On the Relation between Implicitness and Predictability: Combining Annotations of Evoked Questions and Discourse Relations

Nicolaas, Rico

On the Relation between Implicitness and Predictability: Combining Annotations of Evoked Questions and Discourse Relations

Author: Rico Nicolaas

Leiden University

BA Linguistics (Language & Cognition)

Supervisor + Second Reader: Dr. M. Westera, Prof. Dr. L.L. Cheng

Due Date: June 15th, 2025

Abstract

This thesis investigates how readers' evoked questions align with discourse relations and how that alignment governs the omission of explicit connectives. We compared two annotation schemes on the same data, one capturing discourse senses (TED-MDB) and the other capturing evoked questions (TED-Q). This was done by annotating the evoked questions from TED-Q using the same PDTB sense hierarchy as TED-MDB. We show that when a reader's anticipated sense matches the actual relation, authors are more likely to leave the relation implicit. We further demonstrate that causal and resultative relations attract the strongest alignment, indicating that readers' evoked questions predict these relations most reliably. Finally, by comparing sense alignment with the extent to which questions are answered, we find that alignment is a more powerful predictor of implicitness than answered-score. These findings support several linguistic hypotheses surrounding predictability and suggest features for various natural language processing applications.

1. Introduction

With the growing digitization of linguistic data, significantly more corpora have become available to researchers, leading to a marked increase in research on discourse structure over the last two decades (Webber et al., 2012). This wealth of annotated text has proven crucial for various NLP applications: document summarization (III & Marcu, 2009; Ono et al., 1994), information extraction (Maslennikov & Chua, 2007), machine translation (Xiong et al., 2019), and argument mining (Potash et al., 2017). As these tasks fall under NLP, they are part of computational linguistics. In theoretical linguistics, by contrast, one might consider topics such as: pronoun interpretation (Arnold, 2001; Rohde et al., 2007), discourse marking (Yung et al., 2017) and cue effects (Crible et al., 2021).

This study builds on two core discourse annotation frameworks. The Penn Discourse Treebank (PDTB) provides a hierarchy of discourse relations, called “senses” (Prasad et al., 2007; Webber et al., 2019). The Questions Under Discussion (QUD) framework, by contrast, views discourse as a continuation of question–answer pairs, which move the discourse forward (Ginzburg & Sag, 2000; Roberts, 1996; Van Kuppevelt, 1995). Applying both approaches to the same TED Talk transcripts produces two parallel corpora: TED-MDB (Zeyrek et al., 2020) and TED-Q (Westera et al., 2020). TED-MDB contains PDTB-annotated discourse relations, whereas TED-Q contains evoked questions and their corresponding answers, of which the combinations are used to predict QUDs, which are deemed to be crucial for discourse predictability.

TED-MDB and TED-Q have been compared by Westera et al. (2020) to investigate the Uniform Information Density Hypothesis (Frank & Jaeger, 2008) and its two sub-hypotheses: the Causality-by-Default Hypothesis (Sanders, 2005) and the Continuity Hypothesis (Murray, 1997; Segal et al., 1991). The UID Hypothesis states that "predictable linguistic elements are good candidates for reduction". By assuming that evoked questions are good predictors of QUD predictability, they were able to examine how often explicit discourse markers were omitted. Furthermore, van Berkel (2023) investigated the relation between these two frameworks in the same database by zooming in on the alignment between their respective annotations.

Building on his approach, this study directly compares the two frameworks to address the central research question: **"To what extent are discourse relations more implicit when they are predictable?"**. To answer this question, the evoked questions from TED-Q are personally annotated for the PDTB-senses they 'ask for' - that is, the sense readers predict will connect the evoked question to a fitting answer - and they are then compared to the discourse relations from TED-MDB to investigate whether explicit discourse markers are omitted more often when a question's predicted sense aligns with the actual relation. In addition to the main research question, two additional questions were answered to explore factors underlying discourse predictability.

The paper is organized as follows: Chapter 2 reviews the annotation frameworks, explains how discourse annotation operates, and examines how the two frameworks relate. Chapter 3 presents the main research question, describes the dataset and its preprocessing, details the annotation procedure, and outlines the analysis plan. Chapter 4 reports the results, and Chapter 5 discusses their implications for the various hypotheses. Finally, Chapter 6 offers a conclusion.

2. Background

Discourse structure concerns the organization of a text and the relationships between its components. Sentences serve as the fundamental building blocks of discourse, shaping how meaning unfolds. A systematic way to analyze discourse structure is by identifying the relations that connect these building blocks. These relations vary depending on the type of information encoded within them. The interpretation of this information is guided by linguistic features such as semantics, pragmatics, and lexico-syntactic properties. The example below illustrates a discourse relation between two parts of a sentence:

(1) *The boy quarreled with his father* <u>when</u> **he did not get permission to go out.**
(Mak & Sanders, 2013)

The italicized section of the sentence introduces an event, which is further specified in the bolded section by providing a timeframe in which the event took place. These two key components of the discourse relation are connected by a linking word, in this case, *when*. The

resulting relation is classified as a temporal discourse relation, as it progresses the discourse by establishing a chronological sequence. Besides the one stated above, there are many categories of discourse relations, each specifying the nature of the discourse relation.

2.1 Annotation frameworks

2.1.1 Relation-based

Discourse annotation is the process of assigning relations between textual elements to capture how meaning is structured within a text. Various approaches exist, but two are particularly relevant to this study. The first is the relation-based approach, which links adjacent textual units based on a predefined set of discourse relations. Unlike hierarchical approaches, this method focuses on local coherence rather than the overall discourse structure.

The Penn Discourse Treebank (PDTB) is a well-established framework that follows this approach (Prasad et al., 2007). The PDTB provides an extensive inventory of discourse relations and serves as a widely used resource for discourse annotation (Rehbein et al., 2016; Scheffler & Stede, 2016; Song et al., 2024). Now in its third version, PDTB organizes discourse relations into a hierarchical sense structure. Figure 1 below illustrates this hierarchy, where relations, referred to as 'senses', are categorized into three levels, each adding finer distinctions. At the broadest level, four primary categories define discourse relations: Temporal, Contingency, Comparison, and Expansion.

Temporal relations describe the timing of events, distinguishing between those that happen simultaneously and those that follow one another. Contingency relations indicate a dependency between events, often involving cause, condition, or purpose. Comparison relations highlight similarities or differences between two arguments. Expansion relations provide additional information or elaboration. This includes instantiations, where a general statement is followed by an example.

All the level-1 senses are divided into level-2 and level-3 sub-senses. The level-2 senses exist to refine the categorization of discourse relations by providing more detailed information about them. For example, a discourse relation with the label CONTINGENCY.CONDITION denotes

Level-1	Level-2	Level-3
TEMPORAL	SYNCHRONOUS	–
	ASYNCHRONOUS	PRECEDENCE
		SUCCESSION
CONTINGENCY	CAUSE	REASON
		RESULT
		NEGRESULT
	CAUSE+BELIEF	REASON+BELIEF
		RESULT+BELIEF
	CAUSE+SPEECHACT	REASON+SPEECHACT
		RESULT+SPEECHACT
	CONDITION	ARG1-AS-COND
		ARG2-AS-COND
	CONDITION+SPEECHACT	–
	NEGATIVE-CONDITION	ARG1-AS-NEGCOND
		ARG2-AS-NEGCOND
	NEGATIVE-CONDITION+SPEECHACT	–
	PURPOSE	ARG1-AS-GOAL
		ARG2-AS-GOAL
COMPARISON	CONCESSION	ARG1-AS-DENIER
		ARG2-AS-DENIER
	CONCESSION+SPEECHACT	ARG2-AS-DENIER+SPEECHACT
	CONTRAST	–
	SIMILARITY	–
EXPANSION	CONJUNCTION	–
	DISJUNCTION	–
	EQUIVALENCE	–
	EXCEPTION	ARG1-AS-EXCPT
		ARG2-AS-EXCPT
	INSTANTIATION	ARG1-AS-INSTANCE
		ARG2-AS-INSTANCE
	LEVEL-OF-DETAIL	ARG1-AS-DETAIL
		ARG2-AS-DETAIL
	MANNER	ARG1-AS-MANNER
		ARG2-AS-MANNER
	SUBSTITUTION	ARG1-AS-SUBST
		ARG2-AS-SUBST

Figure 1

The PDTB 3.0 Sense hierarchy of discourse relations (Webber et al., 2019).

a relation in which one argument introduces an unrealized situation, which (when realized) leads to the situation described in the other argument.

(2) *Before your next California-bashing editorial, please spend more time out here witnessing the situation* - (Implicit = because) **it just may change your view.**

Level-3 senses constitute a subset of level-2 senses that are asymmetric, meaning the relation between two arguments can be instantiated in either the first argument (Arg1) or the second argument (Arg2). This asymmetry introduces *directionality* to discourse relations, as the placement of the relation's components affects its interpretation. Examples (3) and (4) below illustrate cases of an EXPANSION relation within the LEVEL-OF-DETAIL sub-type, where the more detailed description of a statement appears either in Arg1 or Arg2, demonstrating how the direction of the relation influences discourse structure. Beyond directionality, some of these allow for multiple types of continuation. For instance, a CONTINGENCY.CAUSE sense can extend the discourse in different ways: either by introducing a reason for an event or by specifying its result. In example (5), the discourse progresses by providing the outcome of the event in question.

(3) *Many modern scriptwriters seem to be incapable of writing drama, or anything else, without foul-mouthed cursing. Sex and violence are routinely included even when they are irrelevant to the script, and high-tech special effects are continually substituted for good plot and character development.* In short, **we have a movie and television industry that is either incapable or petrified of making a movie unless it carries a PG-13 or R rating.**
[EXPANSION.LEVEL-OF-DETAIL.ARG1-AS-DETAIL, EXPLICIT]

(4) *An enormous turtle has succeeded where the government has failed:*
(Implicit = specifically) **He has made speaking Filipino respectable.**
[EXPANSION.LEVEL-OF-DETAIL.ARG2-AS-DETAIL, IMPLICIT]

(5) The bill would then declare *that the debt is equity* **and** therefore **isn't deductible.**
[CONTINGENCY.CAUSE.RESULT, EXPLICIT]
(Webber et al., 2019)

The various lexical realizations of discourse relations naturally result in different relation types. First, the presence or absence of a discourse connective determines whether a relation is classified as **Explicit** or **Implicit**. Additionally, explicit relations may include co-occurring connectives, where two or more connectives combine to form a connective pair (e.g., and then, so finally). Second, a discourse relation can be marked as **AltLex** (Alternative Lexicalization) when the relation is expressed through a phrase that is not part of the closed set of discourse connectives. These lexicalizations often appear as multi-word expressions with diverse syntactic constructions. Third, **EntRel** (Entity Relation) refers to a relation between an entity and one of its attributes or characteristics. Unlike semantic relations, EntRel captures a specific type of discourse connection. Since distinguishing it from other relations can sometimes be challenging, it is considered a last-resort strategy. Finally, **NoRel** and **Hypophora** describe distinct cases. NoRel is used when no discourse relation holds between two adjacent sentences, whereas Hypophora marks rhetorical sentence pairs in which a speaker poses a question and immediately answers it. The examples below respectively exhibit an AltLex and EntRel relation:

(6) The moon has moved in front of the sun. *It blocks out most of the light so we can see that dim corona around it.* It would be the same thing **if I put my thumb up and blocked that spotlight that's getting right in my eye,** I can see you in the back row. [EXPANSION:EQUIVALENCE, ALTLEX]

(7) The reason, I would come to find out, was *their prosthetic sockets were painful because they did not fit well.* **The prosthetic socket is the part in which the amputee inserts their residual limb, and which connects to the prosthetic ankle.** [ENTREL]

2.1.2 QUD-based

The second annotation framework used in this study is the Question Under Discussion (QUD) framework, which models discourse structure based on the questions a text raises and the extent to which they are answered (Ginzburg & Sag, 2000; Roberts, 1996; Van Kuppevelt, 1995). Unlike approaches that categorize discourse relations using predefined labels, the QUD

framework views discourse as a structured sequence of implicit or explicit questions that guide interpretation. QUDs, whether explicit or implicit, are crucial to understanding discourse structure. However, implicit QUDs are the hardest to identify, as they must be inferred rather than directly observed.

Typically, QUD-based approaches emphasize the hierarchical nature of discourse, where overarching super-questions structure the text, with sub-questions branching below them (Roberts, 1996). However, this framework has also been applied to the annotation of local discourse structure, which will be reflected in the annotation procedure used in this study. A key motivation for this is that QUD predictability can be monitored more closely: a question is likely to be the QUD of a given passage if it is directly answered by the subsequent text.

Early proponents of the QUD framework aimed to identify structural parallels between discourse relations within sentences and the relationships between the questions these sentences evoke (Van Kuppevelt, 1995; Von Stutterheim & Klein, 1989). This foundational principle continues to underpin the approach today. However, only recently have QUDs been integrated into broader models of discourse coherence (Benz & Jasinskaja, 2017; Ginzburg, 2012; Jasinskaja & Zeevat, 2008; Onea, 2016).

2.2 Annotated datasets

2.2.1 TED-MDB

The TED Multilingual Discourse Bank (TED-MDB) is a corpus of TED Talk transcripts annotated for discourse relations in six languages: English, Polish, German, Russian, European Portuguese, and Turkish (Zeyrek et al., 2020). It follows the Penn Discourse Treebank (PDTB) framework, making it a relation-based approach to discourse annotation. The goal of TED-MDB is to provide a resource for studying discourse structure across languages, improving discourse parsing tools, and enabling cross-linguistic comparisons of discourse relations. TED-MDB consists of six TED Talks, each annotated in all six languages, resulting in a parallel dataset of discourse relations. The overview of all the annotated Ted Talks can be seen in figure 2. These annotations capture various types of discourse relations, which are classified according to the

PDTB framework into Explicit, Implicit, Alternative Lexicalization (AltLex), Entity Relations (EntRel), and No Relation (NoRel).

ID	Author	Title
1927	Chris McKnett	The investment of logic for sustainability
1971	David Sengeh	The sore problem of prosthetic limbs
1976	Jeremy Kasdin	The flower-shaped starshade that might help us detect Earthlike planets
1978	Sarah Lewis	Embrace the near win
2009	Kitra Cahana	A glimpse of life on the road
2150	Dave Troy	Social maps that reveal a city's intersections and separations

Figure 2

TED Talks annotated in TED-MDB (Zeyrek et al., 2020).

TED-MDB contributes to cross-linguistic discourse studies by highlighting how discourse relations vary across languages. By providing a multilingual dataset, it enables researchers to examine how different languages express coherence relations and whether certain relations are more likely to be implicit or explicit depending on linguistic and typological factors. This resource is valuable for discourse parsing, multilingual natural language processing (NLP), and computational models of discourse structure.

2.2.2 TED-Q

TED-Q was developed by applying an evoked question annotation layer to the existing TED-MDB transcripts, thereby introducing an additional layer of discourse annotation (Westera et al., 2020). The annotation process consisted of two distinct phases: the elicitation phase and the comparison phase. During the elicitation phase, the TED-MDB texts were first tokenized using a Python library. These tokenized 'sentences', which could represent smaller segments of longer sentences, were then presented to participants. Their primary task was to generate a question that was naturally evoked by the text up to that point. Additionally, if applicable, participants were instructed to assess the extent to which previously evoked questions had been answered. This was done using a rating scale ranging from 1 to 5, where a higher ANSWERED score indicated that a question had received a more complete answer. The objective of this phase was to ensure

comprehensive coverage of the TED-MDB texts with evoked questions, thereby facilitating a structured analysis of discourse coherence.

Chunk	Highlight	Question	Answered	Related
When we think about mapping cities, we tend to think about roads and streets and buildings, and the settlement narrative the led to their creation, or you might think about the bold vision of an urban designer, but there's other ways of thinking about mapping cities and how they got to be made.	we tend to think about roads and streets	Is there more to map?	3	2.0
	we tend to think about	What is it that we do not think about in regards to mapping cities?	3	1.75
	other ways of thinking about mapping cities	How are cities mapped out?	2	1.75
	mapping cities and how they got to be made	Who maps out cities?	1	2.33

Figure 3

Example of TED-Q annotation process. Taken from TED Talk with id 2150.

Following this, the comparison phase aimed to evaluate the quality and validity of the elicited questions by assessing inter-annotator agreement. This measure is particularly valuable, as it provides insight into whether different annotators similarly interpret discourse relations. To achieve this, annotators were shown snippets of approximately two sentences, accompanied by all the questions that had been elicited for that segment. One of these questions was designated as the 'target,' and a rotation system ensured that every possible pair of questions was compared an equal number of times. Through this process, a RELATED score was assigned to each question, quantifying its degree of similarity to the target question. This score ranged from 1 to 5, where a score of 5 signified that the comparison question was entirely equivalent to the target question. The results of this phase contributed to a more fine-grained understanding of how reliably discourse structure could be annotated using evoked questions.

2.3 Comparison relation-based models and QUD-based models

TED-MDB and TED-Q were both developed to enhance our understanding of discourse structure, yet they differ in key aspects of their annotation frameworks. One major distinction lies in the level of expertise required for annotation. TED-MDB follows the PDTB framework, which provides a predefined, closed inventory of discourse relations. Annotators must choose the most appropriate relation, requiring familiarity with a fixed set of options. In contrast, TED-Q is based on the QUD approach, which offers a more flexible and intuitive process. Rather than selecting from predefined categories, participants generate evoked questions, making the annotation process feel more natural and accessible, particularly for less experienced annotators.

Another key difference is the scope of annotation. QUD-based frameworks assume a hierarchical discourse structure, where an overarching super-question organizes the discourse, with sub-questions branching below it. In contrast, relation-based approaches do not impose such a hierarchy. Instead, they take a more localized approach, annotating discourse relations between adjacent textual units. This localized structure benefits non-expert participants by breaking the process into smaller, more manageable steps.

Despite these differences, TED-Q and TED-MDB share important points of intersection, making comparative research between them particularly valuable. Notably, while they belong to different frameworks, discourse relations and QUDs are closely related. Moreover, TED-Q incorporates a local annotation process as well, matching TED-MDB. However, direct studies comparing the two remain limited. A preliminary analysis by Westera et al. (2020) examined their relationship, finding that 84% of the evoked questions in TED-Q were elicited at points in the discourse where TED-MDB annotated a discourse relation between a question-related element and the following element. More importantly, the study aimed to identify correlations between discourse relation types and evoked question types. Central to this investigation was the Uniform Information Density (UID) Hypothesis, a prominent linguistic theory that proposes information should be distributed evenly throughout an utterance to optimize comprehension (Frank & Jaeger, 2008). More specifically, the hypothesis suggests that more predictable utterances tend to be

expressed more implicitly (Asr & Demberg, 2012). To support this claim, two related hypotheses were tested: the Causality-by-Default Hypothesis (Murray, 1997; Segal et al., 1991), which predicts an implicit bias toward causal inferences, and the Continuity Hypothesis (Sanders, 2005), which posits that discourse is generally structured to maintain coherence and minimize disruptions. The former was tested by investigating whether causal questions get asked and/or answered more, but that was not the case. The latter was also not supported by the findings, since the ANSWERED scores did not prove that continuous questions get answered more.

Despite these insights, certain gaps remain in understanding the correlation between discourse relations and evoked questions. The main limitation of Westera et al. (2020) is that their analysis of predictability relied solely on ANSWERED score. While this metric captures how often evoked questions receive answers, it does not account for whether the type of evoked question aligns with the discourse relation. A high ANSWERED score does not necessarily indicate that the expected discourse relation corresponds to the type of question posed. Moreover, an evoked question could have a low ANSWERED score - meaning that it is not the most likely to be a QUD at that point - but align fully with a discourse relation, meaning that the reader could successfully anticipate what direction the discourse would take at that moment. A more detailed analysis could reveal instances where question types and discourse relations do or do not align, highlighting areas where readers may or may not receive the answers they anticipate. This alignment is essential for text comprehensibility as it reflects whether discourse structure meets reader expectations.

Building on this, van Berkel (2023) examined the alignment between "MDB-senses" and "Q-senses" by applying PDTB-style discourse annotation to TED-Q's evoked questions. His findings indicated that causal relations did not exhibit the highest alignment among level-2 senses. In contrast, continuous relations showed significantly higher alignment than discontinuous ones, suggesting a preference for continuity in discourse structure. However, this preference was not reflected in the ANSWERED scores in TED-Q. Additionally, implicit discourse relations demonstrated greater alignment than explicit or AltLex relations, suggesting a stronger

connection between implicitness and predictability.

He also acknowledged certain limitations in his study. A key concern was the limited sample size, which constrained the robustness of the findings. A larger dataset would allow for a more reliable investigation into the relationship between evoked questions and discourse relations. Additionally, the study lacked a sufficient number of annotators to measure inter-annotator agreement, which has been shown to be a critical factor in annotation research. High inter-annotator agreement ensures that different annotators share a consistent understanding of discourse structure, reinforcing the reliability of the results (Landis & Koch, 1977; Sim & Wright, 2005; Warrens, 2015). Finally, while his finding that implicit relations exhibited higher alignment suggests greater predictability, the study did not explore whether alignment between discourse relations and evoked questions drives implicitness. Given the strong link between implicitness and predictability, further research into these nuances is necessary to refine our understanding of how discourse relations shape comprehension.

3. Methods

3.1 The current study

The PDTB-style annotation framework was used to label the evoked questions in TED-Q with what we term Q-senses, indicating the semantic relationship between each question and its corresponding hypothetical answer (i.e. an answer the reader would expect to be given to the question). Within each Q-sense annotation, the question itself is designated Q-Arg1 and the hypothetical answer Q-Arg2. These annotations were then aligned alongside the PDTB-annotated discourse relations drawn from TED-MDB to construct the dataset. In the TED-MDB annotations, referred to as MDB-senses, each relation spans two sentences: the first sentence is labeled MDB-Arg1 and the second MDB-Arg2. All annotation data were obtained from the TED-QDB repository, which is created by Westera. It combines the data from TED-MDB and TED-Q.

This study investigates the relationship between discourse relations and evoked questions, with a particular focus on determining whether discourse relations are more likely to be implicit

when they are in alignment with evoked questions. The central research question is: **"To what extent are discourse relations more implicit when they are predictable?"** Predictability can be operationalized in various ways, but in this study, it is defined through the alignment between the MDB-senses and Q-senses. Alignment can occur at three hierarchical levels of the PDTB sense hierarchy, but for our main question we focus only on full (Level 1 + 2 + 3) matches. This approach contrasts with that of Westera et al. (2020), who measured predictability using ANSWERED score, indicating how completely an evoked question is resolved. While such scores offer insight into response completeness, they do not capture whether the question anticipates the specific discourse relation intended by the speaker. In contrast, sense alignment provides a more direct lens on how readers interpret and anticipate discourse structure.

To investigate the relationship between discourse relations and evoked questions more generally, a series of cross-variable comparisons were conducted to provide a broad understanding of the relationship between predictability and implicitness. The variables included in the analysis are: ANSWERED score, sense-level alignment, cumulative alignment and the types of discourse relations present. These variables were selected to capture complementary aspects of discourse processing and reader expectation. Alignment was evaluated across all three levels of PDTB annotation. While exact matches between Q-senses and MDB-senses offer the strongest evidence of shared interpretation, partial matches, such as agreement at level-1 or level-2, were also treated as meaningful. For instance, an MDB-sense of CONTINGENCY.CAUSE.REASON and a Q-sense of CONTINGENCY.CAUSE.RESULT represent a closer semantic alignment than a combination like CONTINGENCY.CAUSE.REASON and EXPANSION.CONJUNCTION. Differences in alignment scores across the levels might reflect the change of predictability as discourse relations get more specific.

3.2 Data pre-processing

Before annotation, the TED-QDB dataset, originally stored as a JSON object, was converted to an Excel-compatible format via a Python script, creating a more accessible and structured workspace. All original columns were retained, and a new "Extracted Text" column

was added. Since the source texts and evoked questions were stored separately in TED-QDB, the script extracted the relevant text snippets and placed them adjacent to their corresponding evoked-question entries to prepare for annotation. Once annotation was complete, additional columns were introduced and existing ones adapted to support analysis. Specifically, each Q-sense and MDB-sense was decomposed into three separate columns (one per sense level), and both individual and cumulative alignment scores were computed and recorded in their own columns. Finally, to enable row-wise analysis of multiple Q-sense annotations, the dataset was reshaped into long format, ensuring that each Q-sense label occupied its own record.

3.3 Data annotation

After pre-processing, each entry was annotated by first examining the discourse fragment that evoked the question (MDB-Arg1) alongside the evoked question itself (Q-Arg1), and then formulating a plausible hypothetical answer (Q-Arg2). The semantic relationship between Q-Arg1 and Q-Arg2 was classified according to the PDTB 3.0 sense hierarchy (Webber et al., 2019). Concurrently, each evoked question was evaluated for quality using the "IRR-components" framework: Interrogativity, Relatedness, and Repetition, adapted from van Berkel (2023). Under Interrogativity, questions were marked as open or polar; polar questions (answerable by "yes" or "no") were deemed less likely to reflect substantive discourse expectations. Under Relatedness, questions were flagged if they bore no relevance to the text (e.g. queries about the narrator rather than the narrative). Under Repetition, any question that recurred was marked as repeated. Questions flagged in one or more of these categories were collectively termed "PUR-questions". However, we decided to not omit polar questions, as they could be answered with further elaboration. Applying these criteria led to the omission of 860 evoked questions from the original 2391, yielding a final dataset of 1,531 annotated instances suitable for alignment analysis.

In certain cases, questions could plausibly lead to multiple Q-senses, depending on how the corresponding answer was interpreted. However, to avoid artificially inflating the alignment rate, multiple labels were only assigned when they were clearly justified and strongly supported by the discourse context. In the majority of cases, a single Q-sense was selected to reflect the

most salient discourse relation.

Particular challenges arose when distinguishing between closely related PDTB categories. For example, EXPANSION.INSTANTIATION and EXPANSION.LEVEL-OF-DETAIL were frequently difficult to separate, as both involve elaborative structures or examples. In such cases, the label most directly supported by the immediate discourse context was chosen. Similarly, EXPANSION.CONJUNCTION was often used in situations where the discourse continued without expressing a more specific relation, such as cause or contrast. These trends suggest that EXPANSION-type relations may be overrepresented among the annotated Q-senses, although this assumption requires further empirical validation.

MDB-Arg1	Q-Arg1	Q-Arg2	Q-Sense	IRR-components
(...) I feel so fortunate that my first job was working at the Museum of Modern Art on a retrospective of painter Elizabeth Murray.	Where was the museum?	1) The museum was located in Los Angeles, California.	1) Expansion. Conjunction	Open Related Unrepeated
(...) The reason the near win has a propulsion is because it changes our view of the landscape and puts our goals, which we tend to put at a distance, into more proximate vicinity to where we stand.	Does something that seems close seem more attainable?	1) But because we put them into a more proximate vicinity, they seem to take on a more attainable form. 2) Let's now connect this to the notion of how attainable those goals are.	1) Comparison. Concession. Arg2-As-Denier 2) Expansion. Conjunction	Polar Related Unrepeated
(...) but they sized me up as they made their way to the turf, and spoke to each other not with words but with numbers,	Why did they speak in numbers?	1) Because they were discussing matters about the rings on the board. 2) They spoke in numbers, so that they had a well-informed plan as to how they were going to tackle the training.	1) Contingency. Cause. Reason 2) Contingency. Purpose. Arg2-as-Goal	Open Related Repeated
(...) It's what's called a spirit line, a deliberate flaw in the pattern to give the weaver or maker a way out, but also a reason to continue making work. Masters are not experts	Do you consider yourself a master?	-	NoRel	Polar Unrelated Repeated

Figure 4

Examples of the annotation process.

Figure 4 illustrates examples from the annotation process. It shows several evoked questions alongside their plausible Q-arg2 answers, corresponding annotated Q-senses and IRR-components. Since the MDB-senses are not included in the figure, no single Q-sense is highlighted as correct. Rather, the examples demonstrate the range of potential interpretations that could align with different MDB-senses.

3.4 Analysis

With the annotations done, we evaluated annotation consistency by comparing our labels with those of van Berkel (2023), who applied the same framework to three of the six TED talks in our dataset. Despite the difference in dataset size, this overlap permitted a direct assessment of inter-rater agreement. To quantify this, we computed Cohen's Kappa, a widely used metric for categorical-label agreement (Cohen, 1960). Assessing inter-rater consistency is essential, as it reflects both the reliability of the annotations and the annotators' proficiency in applying them (Landis & Koch, 1977; Sim & Wright, 2005; Warrens, 2015). A high Kappa value would confirm that our independent annotations are reliably aligned, thereby reinforcing confidence that subsequent findings reflect genuine discourse patterns. In contrast, a low Kappa score would signal potential shortcomings in one or both annotation sets, undermining the likelihood that observed effects represent true relations rather than annotation inconsistency.

Then, cumulative alignment between Q-senses and MDB-senses was quantified as follows: when only the sense-category matched, a score of 1 was assigned, when the sense subtype matched, a score of 2 was assigned, and when full alignment, including sense directionality, was achieved, a score of 3 was assigned. To address the primary research question, whether discourse relations become more implicit when they align with evoked questions, a chi-square test of independence was conducted with individual alignment (level-3) and implicitness as factors. Two additional chi-square tests were then performed to examine implicitness rates at alignment level-2 and level-1. **We hypothesize that discourse relations aligned with evoked questions will exhibit higher rates of implicitness.**

To identify which kinds of discourse relations are the most predictable, three one-way ANOVAs were conducted in which cumulative alignment served as the dependent variable and sense-level (sense-category, -subtype or -directionality) as the independent variable, and by comparing mean cumulative alignment scores across sense-levels, the senses that align most strongly on average were revealed. **We hypothesize that the highest mean cumulative alignment scores will occur within the Contingency category, specifically the Cause subtype**,

consistent with the Causality-by-Default hypothesis, which predicts a reader preference for causal relations.

Finally, the study aimed to determine whether cumulative alignment or ANSWERED score provided the best explanation for implicitness in the data, and to this end, a logistic regression model was built using cumulative alignment and ANSWERED score as continuous predictors. **We hypothesize that cumulative alignment will outperform ANSWERED score as a predictor of implicitness**, based on findings by van Berkel (2023), who observed a more pronounced distinction between implicit and explicit relations when predicting alignment than was reported for ANSWERED score by Westera et al. (2020). Moreover, alignment more directly captures a reader's ability to anticipate upcoming discourse, while the ANSWERED score reflects only how completely a question feels satisfied, not how well it was predicted.

All inferential tests were two-tailed with a significance threshold of $\alpha = .05$. For the post-hoc pairwise contrasts following each ANOVA, we applied Tukey's HSD tests, which controls the family-wise error rate across multiple comparisons. No additional correction was applied to the chi-squared tests or the logistic regression.

4. Results

4.1 Descriptive statistics

The core descriptive statistics for our dataset are presented across six tables: Table 1 shows how often each Q-sense slot (1, 2, or 3) was used. Out of 3049 total Q-sense annotations, only 15 occupy the third slot, demonstrating that annotators almost never needed more than two candidate relations. This low rate of level-3 labels points to generally unambiguous question interpretations. However, there are a lot of level-2 labels. There is a high chance that a significant portion of these level-2 labels is taken up by the EXPANSION category, as it is the most annotated category.

Table 1

Annotation counts per Q-sense column (excluding `NoRel`*) and summary of* `NoRel` *occurrences*

	Non-`NoRel` Annotations			**Total**
Column	Q-sense 1	Q-sense 2	Q-sense 3	**Annotations**
Count	1903	644	15	2562
NoRel diagnostics				
Total `NoRel` cells			487	3049

Table 2 compares level-1 sense distributions for MDB-senses versus Q-senses. Two patterns stand out. First, there are far more Q-sense labels than MDB-sense labels, perhaps reflecting differences in annotation expertise. Second, EXPANSION is the top category in both sets, yet it represents an even larger share of Q-senses, suggesting that, when readers ask questions, they disproportionately seek elaborations or examples relative to how often those relations naturally occur. The other sense categories appear in similar absolute counts across TED-MDB and TED-Q, although their relative proportions are slightly higher in TED-MDB, implying that evoked-question annotators may underdetect relations that do not involve explicit elaboration.

Table 2

Distribution of MDB-sense-categories and Q-sense-categories.

Sense-category	**MDB**	**MDB %**	**Q**	**Q %**
Expansion	865	55,3%	1684	65,7%
Contingency	416	25,2%	489	19,1%
Temporal	71	4,5%	67	2,6%
Comparison	237	15,0%	322	12,6%
Total	1574	100%	2562	100%

Table 3 reports cumulative alignment levels between MDB-senses and Q-senses over all 3049 cases. More than half (2048 instances) show no alignment, corroborating van Berkel's finding that readers frequently pose questions that diverge from the textbook PDTB-relation. However, alignments at levels 2 and 3 occur more often here than in van Berkel's work. We attribute this in part to our decision to merge the subtypes CONTINGENCY.CAUSE+BELIEF and CONTINGENCY.CAUSE+SPEECHACT into the general CONTINGENCY.CAUSE label, thereby boosting higher-level match counts under a more flexible scheme. Table 4 presents the distribution of ANSWERED score. The modal rating is 1, indicating that, in most cases, participants judged their question as still unanswered by the subsequent sentence(s).

Table 5 displays the breakdown of relation types in TED-MDB. We excluded EntRel and NoRel from this table because they describe relation types rather than sense categories per se. Among the six relation types, Implicit is most frequent, followed by Explicit; AltLex and EntRel follow, with NoRel the least common. Finally, Table 6 summarizes the IRR-components. Open, Related, and Unrepeated labels predominate over their PUR-question counterparts, confirming that only a small fraction of questions required exclusion from alignment analyses.

Table 3

Cumulative agreement counts per level.

Alignment level	Count	Percent (%)
No alignment	2048	67.2%
1	427	14.0%
2	249	8.2%
3	325	10.6%
Total	3049	100.00%

Table 4

Distribution of ANSWERED-scores

Score	Count	Percent (%)
1	1005	42.03%
2	300	12.55%
3	367	15.35%
4	362	15.14%
5	357	14.93%
Total	2391	100.00%

Table 5

Distribution of relation types in TED-MDB.

Relation type	Count	Percent (%)
Explicit	609	22.8%
Implicit	1091	40.8%
AltLex	135	5.1%
EntRel	502	18.8%
NoRel	335	12.5%
Total	2672	100.00%

Table 6

IRR-components and their number of occurrences.

Category	Annotation	Count	Percent (%)
Interrogativity	Open	1863	77.9%
	Polar	528	22.1%
Relatedness	Related	1904	79.6%
	Unrelated	487	20.4%
Repetition	Unrepeated	2002	83.7%
	Repeated	389	16.3%

4.2 Inter-annotator agreement

Inter-annotator agreement was culculated using Cohen's Kappa. This was done in Python, as all the other parts of the analysis in this study. To contextualize the resulting values, Table 7 summarizes the conventional interpretation ranges for inter-rater reliability.

Table 7

Interpretation of Cohen's Kappa values.

κ range	Interpretation
$0.00 \leq \kappa \leq 0.20$	Slight agreement
$0.21 \leq \kappa \leq 0.40$	Fair agreement
$0.41 \leq \kappa \leq 0.60$	Moderate agreement
$0.61 \leq \kappa \leq 0.80$	Substantial agreement
$0.81 \leq \kappa \leq 1.00$	Perfect perfect agreement

The Cohen's Kappa value was calculated as $\kappa = 0.301$, placing it in the "fair agreement" category. In order to enhance comparability between our annotations and those of Van Berkel, a pre-processing step was applied. Specifically, any sense labels in van Berkel's dataset that lacked a sense-directionality, such as EXPANSION.CONJUNCTION, were programmatically normalized to include a duplicated third level (i.e. EXPANSION.CONJUNCTION.CONJUNCTION). We mirrored this transformation in our own annotations so that every label in both datasets uniformly possessed three sense-levels. Despite this effort, the Kappa value remained at a fair range, implying only a modest degree of overlap between our annotations. Such an outcome underscores the limited stability of these annotations: annotators agreed more often than chance would predict, yet substantial discrepancies persisted. These discrepancies may reflect the inherent complexity of the annotation task or the relative inexperience of us as annotators. Future studies could enhance annotation reliability by increasing the number of annotators or by first providing annotators with a training set. Although inter-annotator agreement was only fair, our reannotation of Van Berkel's data and the expansion to a larger dataset adds significant value. Moreover, showing that key trends persist across both the original and our annotations highlights the consistency of these patterns.

4.3 Results

Chi-squared tests

The first chi–squared test of independence was performed on a 2×2 contingency table crossing Implicitness (Implicit vs. Not Implicit) with Level-3 Alignment (Aligned vs. Not Aligned). The result was significant (χ^2 (1, N = 2269) = 70.56, p < .001) indicating that implicitness and alignment are not independent. To quantify effect size, we computed Cramer's V, which turned out to be approximately 0.18. This is a small-to-medium effect. We also calculated the odds ratio:

$$\text{OR} = \frac{210 \times 1176}{768 \times 115} \approx 2.79,$$

showing that the odds of a relation being implicit is roughly 2.8 times higher when it is Level-3 aligned than when it is not aligned. See figure 5 below.

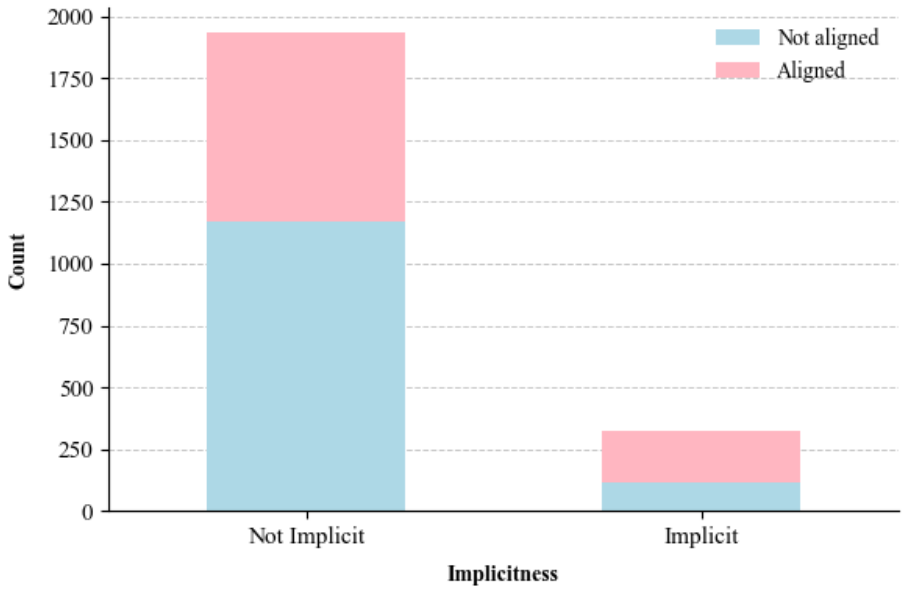

Figure 5

Distribution of implicitness for level-3 alignment.

The second chi–squared test of independence was conducted on a 2×2 table crossing Implicitness (Implicit vs. Not Implicit) with Level-2 Alignment (Aligned vs. Not Aligned). The result was significant (χ^2 (1, N = 2269) = 103.03, p < .001) indicating a non-independent

relationship between implicitness and alignment at Level 2. For effect size, Cramer's V was computed, which turned out to be approximately 0.21, which is a small-to-medium effect. We also calculated the odds ratio:

$$\text{OR} = \frac{352 \times 1069}{626 \times 222} \approx 2.71,$$

showing that the odds of a relation being implicit is about 2.7 times higher if it is Level-2 aligned than when it is not aligned. See figure 6 below.

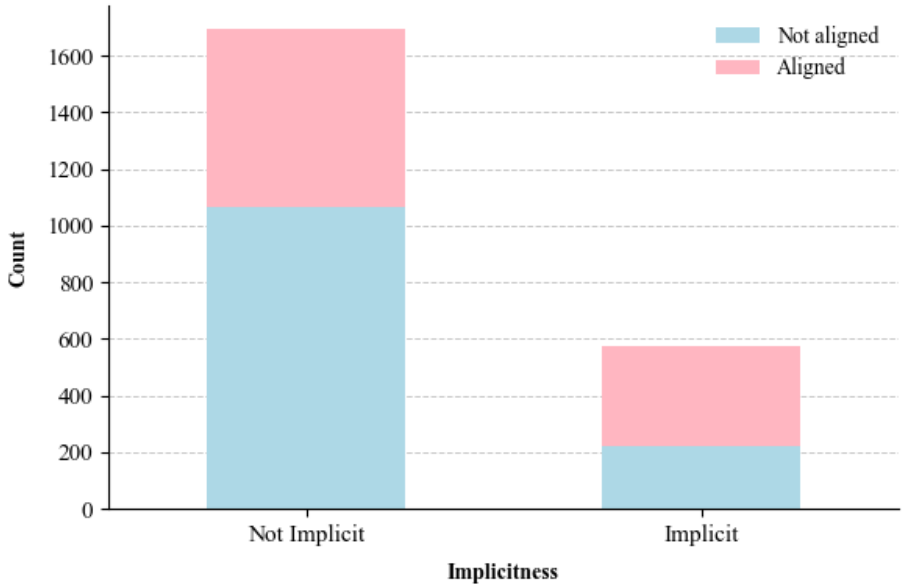

Figure 6

Distribution of implicitness for level-2 alignment.

The third chi–squared test of independence was conducted on a 2×2 table crossing Implicitness (Implicit vs. Not Implicit) with Level-1 Alignment (Aligned vs. Not Aligned). The result was significant (χ^2 (1, N = 2269) = 288.79, p < .001) indicating a non-independent relationship between implicitness and alignment at Level 1. For effect size, Cramer's V was computed, which turned out to be approximately 0.36, which is a medium-to-large effect. We also calculated the odds ratio:

$$\text{OR} = \frac{631 \times 921}{347 \times 370} \approx 4.53,$$

showing that the odds of a relation being implicit is about 4.5 times higher if it is Level-1 aligned than when it is not aligned. See figure 7 below.

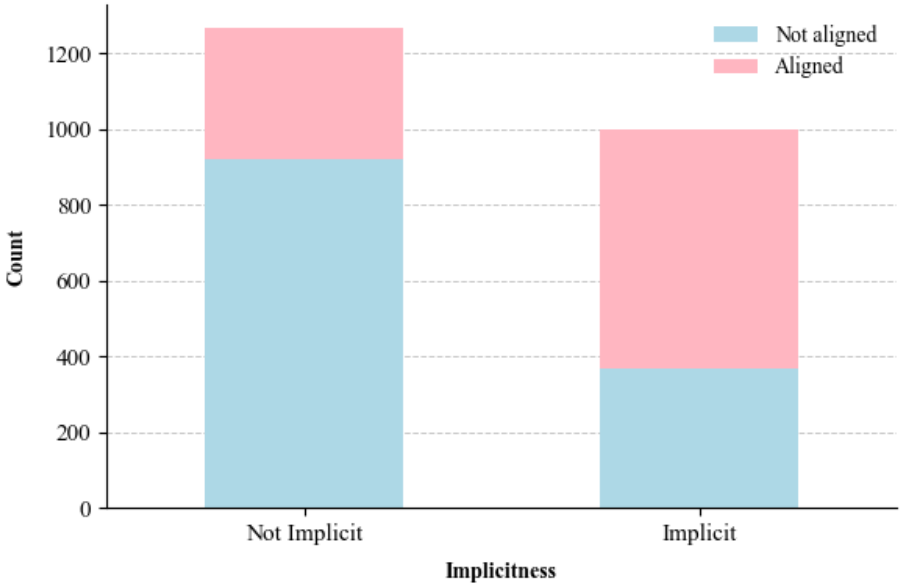

Figure 7

Distribution of implicitness for level-1 alignment.

ANOVAs

The first ANOVA showed that mean cumulative alignment differs significantly across the four sense-categories ($F(3, 2553) = 8.47$, $p < .001$). Post-hoc Tukey HSD tests revealed that CONTINGENCY relations ($M = 0.93$, $SD = 1.35$) align significantly more closely with evoked questions than COMPARISON relations ($M = 0.56$, $SD = 1.05$, $p < .001$) and more than EXPANSION relations ($M = 0.73$, $SD = 0.97$, $p = .001$). EXPANSION relations also align more closely than COMPARISON relations ($p = .046$). No other pairwise differences reached significance; the TEMPORAL category ($M = 0.66$, $SD = 1.21$) did not differ reliably from any other sense. Figure 8 shows the mean cumulative alignment per sense-category.

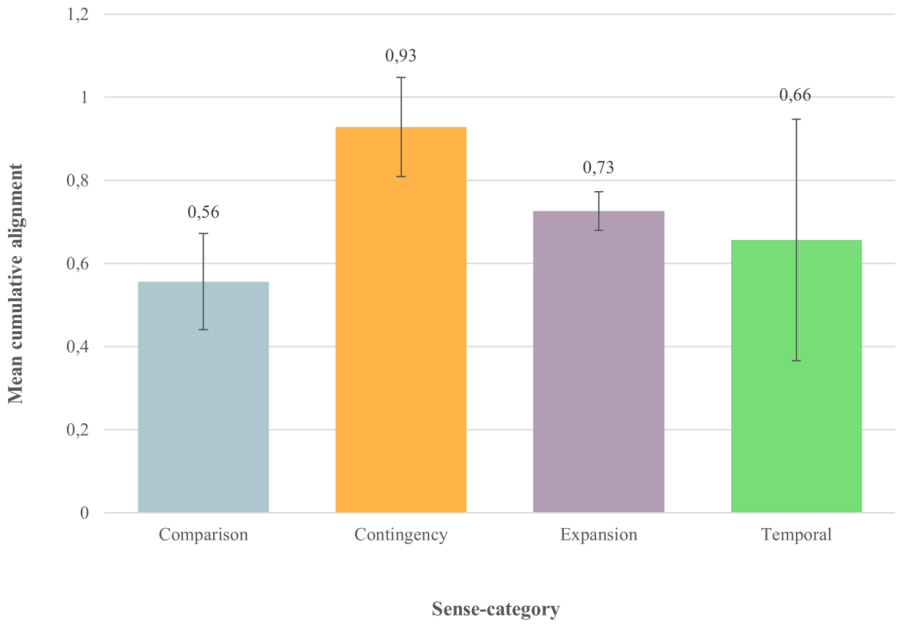

Figure 8

Mean cumulative alignment per sense-category.

The second ANOVA revealed significant differences in mean alignment across the sixteen sense-subtypes ($F(15, 2949) = 5.70$, $p < .001$). Post-hoc Tukey tests showed that CAUSE ($M = 1.00$, $SD = 1.38$) aligns significantly more than CONCESSION ($M = 0.64$, $SD = 1.19$, $p = 0.008$) , CONJUNCTION ($M = 0.74$, $SD = 0.90$, $p = .008$), CONTRAST ($M = 0.43$, $SD = 0.75$, $p < 0.001$), MANNER ($M = 0.40$, $SD = 0.49$, $p < 0.001$), PURPOSE ($M = 0.15$, $SD = 0.36$, $p < 0.001$), INSTANTIATION ($M = 0.71$, $SD = 0.96$, $p = 0.014$) and LEVEL-OF-DETAIL ($M = 0.76$, $SD = 1.10$, $p = 0.048$), and that SUBSTITUTION ($M = 1.55$, $SD = 1.28$) aligns significantly more than CONCESSION ($p = 0.026$), CONTRAST ($p = 0.002$), MANNER ($p = 0.001$), PURPOSE ($p < 0.001$), SIMILARITY ($M = 0.29$, $SD = 0.64$, $p = 0.014$) and EQUIVALENCE ($M = 0.62$, $SD = 0.75$, $p = 0.038$). In contrast, categories such as CONTRAST ($M = 0.43$), MANNER ($M = 0.40$) and PURPOSE ($M = 0.15$) show the lowest alignment rates. These results suggest a clear predictability hierarchy among relations, with causal and substitution-type discourse relations being most reliably anticipated by readers. Figure 9 below shows the mean cumulative alignment per sense-subtype.

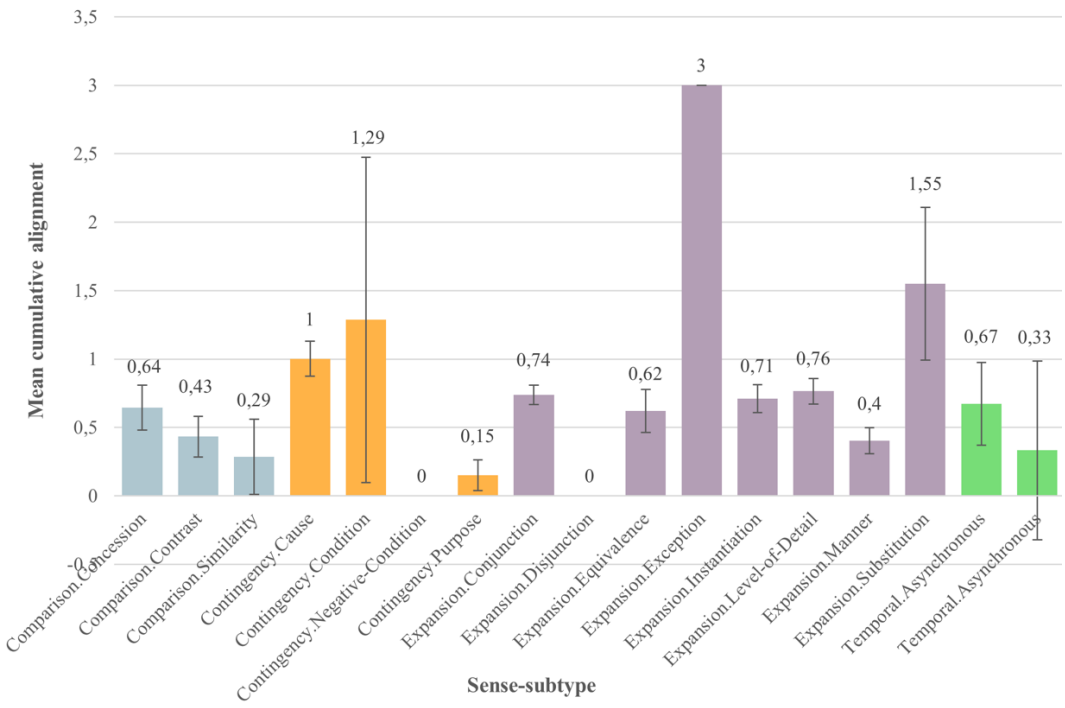

Figure 9

Mean cumulative alignment per sense-subtype.

The third ANOVA revealed significant differences in mean alignment across the third sense-directionalities $F(13,2229) = 6.43$, $p < .001$). The post-hoc Tukey tests showed that the ARG2-AS-SUBST ($M = 1.55$, $SD = 1.28$) directionality aligns significantly more than ARG2-AS-DENIER ($M = 0.64$, $SD = 1.19$, $p = .021$). Also, ARG2-AS-DETAIL ($M = 0.76$, $SD = 1.10$) aligns significantly more than ARG2-AS-GOAL ($M = 0.15$, $SD = 0.37$, $p = .038$). Moreover, RESULT ($M = 1.19$, $SD = 1.44$) aligns significantly more than ARG2-AS-DETAIL ($p = .001$). Lastly, ARG2-AS-EXCPT ($M = 3.00$, $SD = 0.00$) aligns significantly more than ARG2-AS-COND ($M = 1.29$, $SD = 1.60$, $p = .017$) and ARG2-AS-MANNER ($M = 0.40$, $SD = 0.49$, $p = .043$). Because ARG2-AS-EXCPT is based on only two observations and represents an extreme outlier, these particular contrasts should be interpreted with caution and the overall ANOVA regarded as inconclusive for that category. Table 8 shows the mean cumulative alignment per sense-directionality, ranked from highest to lowest.

Table 8

Mean cumulative alignment per sense-directionality.

Sense	Mean cumulative alignment	Number of occurances
Expansion.Exception.Arg2-as-Excpt	3.000	2
Expansion.Substitution.Arg2-as-Subst	1.550	20
Contingency.Condition.Arg2-as-cond	1.286	7
Contingency.Cause.Result	1.189	159
Contingency.Cause.Reason	0.896	279
Temporal.Asynchronous.Precedence	0.765	51
Expansion.Level-of-Detail.Arg2-as-Detail	0.764	522
Expansion.Instantiation.Arg2-as-Instance	0.710	345
Comaprison.Concession.Arg2-as-Denier	0.644	202
Expansion.Manner.Arg2-as-Manner	0.402	102
Temporal.Asynchronous.Succession	0.308	13
Contingency.Purpose.Arg2-As-Goal	0.154	39
Contingency.Negative-Condition.Arg1-as-Negcond	0.000	4
Contingency.Purpose.Arg1-as-Goal	0.000	1

Logistic regression

A binary logistic regression was conducted to predict whether a discourse relation is implicit from cumulative alignment, answered-scores and their interaction. The model was significant, $\chi^2(3) = 250.0$, $p < .001$, and explained approximately 6.9% of the variance in implicitness (Nagelkerke $R^2 = .069$). The main effect of alignment was positive and highly significant (b = 0.59, SE = 0.04, Wald z = 15.17, p < .001), indicating that each one-unit increase in cumulative alignment increased the log-odds of an implicit relation by .59 (OR $\approx$ 1.80; 95% CI [1.66, 1.99]). The main effect of answered-scores was also significant (b = 0.08, SE = 0.03, Wald z = 2.90, p = .004), such that better answered-scores slightly raised the odds of implicitness (OR

≈ 1.08; 95% CI [1.02, 1.14]). Crucially, the interaction term was negative but did not reach significance (b = -0.038, SE = 0.026, Wald z = -1.45, p = .146). Thus, higher cumulative alignment and higher answered-scores each independently increase the odds of an implicit relation. The negative interaction term, although not reaching conventional significance, suggests a small tendency for the effect of predictability to diminish at higher levels of answeredness. Figure 10 shows the effect of the two predictors on implicitness.

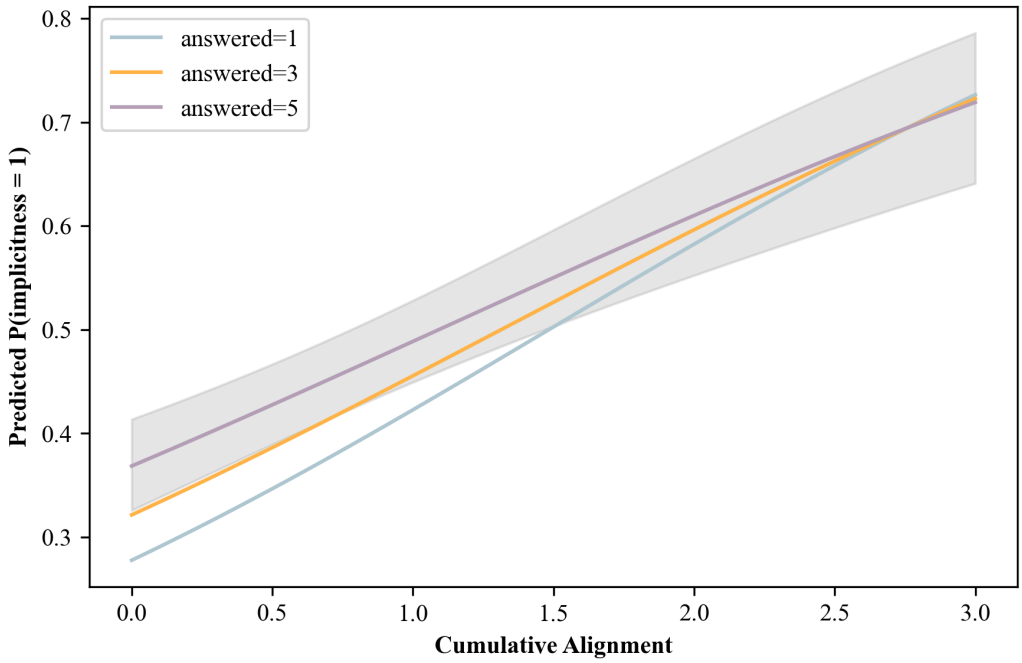

Figure 10

Predicted Probability by alignment and answered-scores.

5. Discussion

Three chi-squared tests were conducted to determine whether discourse relations are rendered more implicit when they align with evoked questions. In the first test, designed to address the primary research question, full alignment, so level-3 alignment between Q-senses and MDB-senses, was used as the criterion. It was found that the odds of a relation being implicit is roughly 2.8 times higher when it is Level-3 aligned than when it is not aligned. Thus, our hypothesis that discourse relations would be more implicit when aligned can be accepted. The

result indicates that, at the population level, the precise anticipation afforded by readers' evoked questions allows the omission of explicit connectives without loss of coherence. A second test examined level-2 alignment, that is, matches on both the overarching class and its subtype (e.g., Cause, Result). Here, it was found that the odds of a relation being implicit is roughly 2.7 times higher when it is Level-2 aligned than when it is not aligned. The similarity of this effect size to that of level-3 alignment suggests that directional detail offers little additional predictive value beyond that provided by class and subtype matching. Finally, coarse level-1 alignment alone was evaluated. In this condition, the odds of a relation being implicit is roughly 4.5 times higher when it is Level-1 aligned than when it is not aligned, producing a significantly larger effect. This finding implies that general expectations regarding discourse progression, once signaled by the evoked question, allow writers to omit explicit markers with greatest confidence. In contrast, finer-grained predictability cues appear to prompt the inclusion of explicit connectives to disambiguate directional subtypes. This might seem counter-intuitive, but it might be because precision is needed more when discourse gets more detailed or specific. Together, these results demonstrate that broad predictability has the strongest influence on implicitness, whereas more detailed anticipatory cues have diminishing effects. These results support the UID hypothesis (Frank & Jaeger, 2008), which predicts that highly predictable discourse relations tend to be left implicit. In our data, implicit relations, by definition omitting explicit connectives, are precisely the instances that readers most reliably anticipate, making them natural QUD candidates (Westera et al., 2020). Because cumulative alignment quantifies how closely a reader's evoked question matches the actual discourse sense, it functions as a direct measure of this predictability. In other words, the higher a relation's alignment score, the more accurately it has been predicted in advance, and therefore the more confidently writers can omit a connective. Implicit relations and high alignment thus go hand in hand: both signal that the upcoming discourse move was already on the reader's radar, satisfying the principle that predictable information can be compressed out of the text.

Also, three ANOVAs were conducted to determine which senses had the highest

cumulative alignment i.e. which senses were 'asked' the most by readers. The first ANOVA compared the mean cumulative alignment scores of the four sense-categories. Results showed that the CONTINGENCY relation aligned the most out of all the categories, followed by EXPANSION, then TEMPORAL and lastly COMPARISON. The results confirm our hypothesis that the highest cumulative alignment scores would occur in the CONTINGENCY category. What this might indicate is that readers are often more interested in the reason or explanation of why something occured, rather than in expanding the discourse, highlighting differences or similarities in the discourse, or temporal sequencing. This contradicts the findings by van Berkel (2023), who found the highest cumulative alignment scores to occur in the EXPANSION category. Viewed through a QUD lens (Ginzburg & Sag, 2000; Roberts, 1996; Van Kuppevelt, 1995), this finding suggests that readers most often drive the discourse forward by asking for causal explanations.

The second ANOVA compared mean cumulative alignment scores across sense-subtypes. The results indicate that the CAUSE subtype, contained within the CONTINGENCY category, has a higher average alignment score than any subtype from the other categories. A few exceptions arise: for instance, the EXCEPTION and SUBSTITUTION subtypes, both contained in the EXPANSION category, also exhibit relatively high mean scores. However, these appear to be outliers and therefore may not reliably reflect their population values. Beyond these outliers, the remaining subtypes, particularly those in EXPANSION category, show fairly uniform mean alignment scores, suggesting that readers may struggle to anticipate more specific senses within a given category. In other words, distinguishing among sense-categories and sense-subtypes is not necessarily intuitive for readers, even though such distinctions are valuable when annotating discourse relations and formulating evoked questions. The fact that the CAUSE subtype yields the highest mean cumulative alignment score is consistent with the Causality-by-Default Hypothesis, which postulates a preference for causal relations (Sanders, 2005). Notably, however, this hypothesis did not find corroborating evidence in TED-Q: causal questions in that dataset did not show a significantly lower answered-score compared to questions beginning with "what" or falling into the "other" category.

The third ANOVA compared mean cumulative alignment scores across sense-directionalities. Results revealed significant differences among directionalities within the EXPANSION category. Additionally, the RESULT directionality exhibited significantly higher alignment scores than many other directionalities. This pattern suggests that readers may be particularly interested in the outcomes of specific actions, rather than in other ways of advancing the discourse. These findings contrast with those of van Berkel (2023), who found only one significant difference within sense-directionality, which occurred between REASON and RESULT, with REASON showing a significantly higher cumulative mean alignment score. His finding suggests that readers are more interested in understanding why an event occurred than in its outcomes. This aligns with Westera et al. (2020), who note that English lacks a dedicated wh-word for resultative relations, whereas "why" is unambiguously tied to explanatory relations. Our findings, however, suggest the reverse: readers appear more curious about an event's outcomes than its causes. This discrepancy could stem from our larger dataset or from differences in how we annotated evoked questions, which is an issue underscored by the moderate Cohen's Kappa score of 0.301.

The logistic regression model was designed to determine which variable, cumulative alignment or ANSWERED score, has a greater influence on whether a discourse relation is implicit or explicit. The results show that while both predictors are significant, cumulative alignment is a substantially stronger predictor of implicitness than ANSWERED score. This finding echoes van Berkel (2023), who reported that alignment differences between implicit and explicit relations are more pronounced than differences in ANSWERED score. What this also means, is that our hypothesis is confirmed. Taken together, these results suggest that alignment is a superior measure of predictability compared to ANSWERED score. In contrast to Westera et al. (2020), who emphasize answered-scores as the primary indicator of predictability, this analysis demonstrates that the mere act of a reader posing a question that accurately forecasts the upcoming discourse carries more weight in driving implicitness than the degree to which that question is ultimately answered.

6. Conclusion

Earlier studies have explored relation-based and QUD-based frameworks separately. Some have even applied both frameworks to the same dataset to uncover patterns linking discourse relations with evoked questions. What remains underexplored, however, is how predictability influences the use of explicit discourse markers. This study addresses that gap by testing whether relations that align with readers' questions are more likely to be left implicit. Our findings confirm that aligned relations are indeed dropped more often, although this effect weakens as alignment becomes more fine-grained. We also find that CONTINGENCY relations, especially resultative causal links, attract the highest mean alignment scores, indicating that readers predict these relations most reliably. Finally, a comparative analysis shows that sense alignment outperforms ANSWERED score in predicting implicitness. Together, these results provide evidence for the UID hypothesis and the Causality-by-Default hypothesis. Also, this new knowledge about discourse structure could provide insight to the NLP applications mentioned earlier: document summarization (III & Marcu, 2009; Ono et al., 1994), information extraction (Maslennikov & Chua, 2007), machine translation (Xiong et al., 2019), and argument mining (Potash et al., 2017). Not only this, but according to Alhaydar (2025), intelligent technology tools such as communicative AI are not yet fully able to understand all the complexities of human language. Therefore, studies like these could prove useful in providing such communicative tools with data.

There were some shortcomings in this study, however. Despite addressing some issues from prior work, such as increasing sample size and calculating inter-annotator agremeent, others persist. Our relative inexperience with discourse annotation likely affected the consistency of labels, which in turn reduces the overall robustness of the findings. In addition, we handled IRR components differently from van Berkel (2023): whereas he excluded all polar questions, we retained them in our dataset. Because these polar questions were evenly distributed across sense categories, their inclusion should not have biased the results, but it did increase the sample size. In his thesis, van Berkel (2023) conducted a rank test to identify which model best fit the data, finding that the version without polar questions ranked fourth out of eight. However, he also noted

that differences in performance among all models were minimal and that all models still significantly explained the data.

Brunetti and Riester (2025) propose that discourse can be modeled as a hierarchical QUD-tree, where each nonterminal node represents an implicit or explicit question driving the text and each terminal node its answer. They highlight, however, that current frameworks lack a clear method for incorporating non-assertive speech acts (such as imperatives and exclamatives) and their associated "meta-questions" into this structure. Future research could address this gap by developing annotation schemes and parsing algorithms that build full QUD-trees over extended discourse, allow any speech act to occupy question or answer nodes, and then align those nodes with established PDTB-style discourse senses. Such an integrated model would broaden our theoretical understanding of how global and local discourse interact and offer a more comprehensive backbone for discourse-aware NLP applications.

References

Alhaydar, A. (2025). NLP and figurative language: A quantitative study exploring the competence of AI-powered bots in understanding metaphors and idioms [Publisher: İzmir Academy Association]. *Journal of AI*, *9*(1), 32–55. Retrieved June 8, 2025, from https://dergipark.org.tr/en/pub/jai/issue/87787/1613027

Arnold, J. E. (2001). The effect of thematic roles on pronoun use and frequency of reference continuation. *Discourse Processes*, *31*(2), 137–162. https://doi.org/10.1207/S15326950DP3102_02

Asr, F. T., & Demberg, V. (2012). Implicitness of discourse relations. *Proceedings of COLING 2012*, 2669–2684. Retrieved January 9, 2025, from https://aclanthology.org/C12-1163.pdf

Benz, A., & Jasinskaja, K. (2017). Questions under discussion: From sentence to discourse. *Discourse Processes*, *54*(3), 177–186. https://doi.org/10.1080/0163853X.2017.1316038

Brunetti, L., & Riester, A. (2025). Non-assertive speech acts and their QUDs [Publisher: Logos Verlag]. *Annotating Text with Questions Under Discussion*. Retrieved June 9, 2025, from http://www.llf.cnrs.fr/sites/llf.cnrs.fr/files/biblio/Brunetti-Riester-informationStructureQuestions.pdf

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Crible, L., Wetzel, M., & Zufferey, S. (2021). Lexical and structural cues to discourse processing in first and second language [Publisher: Frontiers Media SA]. *Frontiers in psychology*, *12*, 685491. Retrieved June 7, 2025, from https://www.frontiersin.org/articles/10.3389/fpsyg.2021.685491/full

Frank, A. F., & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production [Issue: 30]. *Proceedings of the annual meeting of the cognitive science society*, *30*. Retrieved February 25, 2025, from https://escholarship.org/content/qt7d08h6j4/qt7d08h6j4.pdf

Ginzburg, J. (2012). *The interactive stance*. Oxford University Press, USA. Retrieved February 23, 2025, from https: //books.google.nl/books?hl=nl&lr=&id=W4WOWZxKWZAC&oi=fnd&pg=PP1&dq= The+interactive+stance&ots=TIVOLYGbAG&sig=Vr9Dmngic0di6atQdUkEI6x8mJw

Ginzburg, J., & Sag, I. (2000). *Interrogative investigations*. Stanford: CSLI publications. Retrieved March 4, 2025, from https://www.academia.edu/download/101977887/ 64d4657a899b56d5b1271dfb2ee943c8fedb.pdf

III, H. D., & Marcu, D. (2009, July 4). A noisy-channel model for document compression. https://doi.org/10.48550/arXiv.0907.0806

Jasinskaja, K., & Zeevat, H. (2008). Explaining additive, adversative and contrast marking in russian and english [Publisher: Presses universitaires d'Orléans]. *Revue de Sémantique et Pragmatique*, *24*(1), 65–91. Retrieved February 23, 2025, from https://www.researchgate.net/profile/Henk-Zeevat/publication/242074822_Explaining_additive_adversative_and_contrast_marking_in_Russian_and_English/links/0046353033217af9a4000000/Explaining-additive-adversative-and-contrast-marking-in-Russian-and-English.pdf

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data [Publisher: JSTOR]. *biometrics*, 159–174. Retrieved May 26, 2025, from https://www.jstor.org/stable/2529310

Mak, W. M., & Sanders, T. J. M. (2013). The role of causality in discourse processing: Effects of expectation and coherence relations. *Language and Cognitive Processes*, *28*(9), 1414–1437. https://doi.org/10.1080/01690965.2012.708423

Maslennikov, M., & Chua, T.-S. (2007). A multi-resolution framework for information extraction from free text. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 592–599. Retrieved June 7, 2025, from https://aclanthology.org/P07-1075.pdf

Murray, J. D. (1997). Connectives and narrative text: The role of continuity. *Memory & Cognition*, *25*(2), 227–236. https://doi.org/10.3758/BF03201114

Onea, E. (2016). *Potential questions at the semantics-pragmatics interface* (Vol. 33). Brill. Retrieved February 23, 2025, from https://books.google.nl/books?hl=nl&lr=&id=OyYiDAAAQBAJ&oi=fnd&pg=PP5&dq=Potential+questions+at+the+semantics-pragmatics+interface&ots=DeTvzOUUQU&sig=948ozH6Bx7_3Xo0iqo-C_TZZ3Wc

Ono, K., Sumita, K., Research, S. M., & Center, D. (1994). Abstract generation based on rhetorical structure extraction. *arXiv preprint cmp-lg/9411023*. Retrieved June 7, 2025, from https://arxiv.org/abs/cmp-lg/9411023

Potash, P., Romanov, A., & Rumshisky, A. (2017, May 8). Here's my point: Joint pointer architecture for argument mining. https://doi.org/10.48550/arXiv.1612.08994

Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., Robaldo, L., & Webber, B. (2007). The penn discourse treebank 2.0 annotation manual. *December*, *17*, 2007. Retrieved February 19, 2025, from https://catalog.ldc.upenn.edu/docs/LDC2008T05/manual/pdtb-annotation-manual.pdf

Rehbein, I., Scholman, M., & Demberg, V. (2016). Annotating discourse relations in spoken language: A comparison of the PDTB and CCR frameworks. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 1039–1046. Retrieved February 18, 2025, from https://aclanthology.org/L16-1165/

Roberts, C. (1996). Information structure in discourse: Toward a unified theory of formal pragmatics. *Ohio State University Working Papers in Linguistics*, *49*, 91–136. Retrieved March 4, 2025, from https://www-formal.stanford.edu/buvac/95-context-symposium/Papers/croberts.ps

Rohde, H., Kehler, A., & Elman, J. L. (2007). Pronoun interpretation as a side effect of discourse coherence [Issue: 29]. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *29*. Retrieved June 7, 2025, from https://escholarship.org/content/qt9738n5wf/qt9738n5wf.pdf

Sanders, T. (2005). Coherence, causality and cognitive complexity in discourse. *Proceedings/Actes SEM-05, First International Symposium on the exploration and modelling of meaning*, 105–114. Retrieved February 25, 2025, from https://www.researchgate.net/profile/Ted-Sanders-2/publication/46669022_Coherence_Causality_and_Cognitive_complexity_in_discourse/links/53eb50ab0cf2fb1b9b6b0e20/Coherence-Causality-and-Cognitive-complexity-in-discourse.pdf

Scheffler, T., & Stede, M. (2016). Adding semantic relations to a large-coverage connective lexicon of german. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 1008–1013. Retrieved February 18, 2025, from https://aclanthology.org/L16-1160/

Segal, E. M., Duchan, J. F., & Scott, P. J. (1991). The role of interclausal connectives in narrative structuring: Evidence from adults' interpretations of simple stories. *Discourse Processes*, *14*(1), 27–54. https://doi.org/10.1080/01638539109544773

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements [Publisher: Oxford University Press]. *Physical therapy*, *85*(3), 257–268. Retrieved May 26, 2025, from https://academic.oup.com/ptj/article-abstract/85/3/257/2805022

Song, W., Han, H., Han, X., Cheng, M., Gong, J., Wang, S., & Liu, T. (2024). Discriminative explicit instance selection for implicit discourse relation classification. *Frontiers of Computer Science*, *18*(4), 184340. https://doi.org/10.1007/s11704-023-3058-2

Van Kuppevelt, J. (1995). Discourse structure, topicality and questioning [Publisher: Cambridge University Press]. *Journal of linguistics*, *31*(1), 109–147. Retrieved February 23, 2025, from https://www.cambridge.org/core/journals/journal-of-linguistics/article/discourse-structure-topicality-and-questioning/60F3E68601517091AF560CB3CC02C6AC

van Berkel, F. (2023, June 30). *Evoked questions in TED talks and the discourse relations they ask for* [Doctoral dissertation, University of Leiden].

Von Stutterheim, C., & Klein, W. (1989). Referential movement in descriptive and narrative discourse. In *North-holland linguistic series: Linguistic variations* (pp. 39–76, Vol. 54). Elsevier. Retrieved February 23, 2025, from https://www.sciencedirect.com/science/article/pii/B9780444871442500057

Warrens, M. J. (2015). Five ways to look at cohen's kappa [Publisher: OMICS International]. *Journal of Psychology & Psychotherapy*, *5*. Retrieved May 10, 2025, from https://research.rug.nl/en/publications/five-ways-to-look-at-cohens-kappa

Webber, B., Egg, M., & Kordoni, V. (2012). Discourse structure and language technology [Publisher: Cambridge University Press]. *Natural Language Engineering*, *18*(4), 437–490. Retrieved January 30, 2025, from https://www.cambridge.org/core/journals/natural-language-engineering/article/discourse-structure-and-language-technology/E6F91D0B578466AA25F6BC585E8AFDCE

Webber, B., Prasad, R., Lee, A., & Joshi, A. (2019). The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*, *35*, 108. Retrieved May 5, 2024, from https://catalog.ldc.upenn.edu/docs/LDC2019T05/PDTB3-Annotation-Manual.pdf

Westera, M., Mayol, L., & Rohde, H. (2020). TED-q: TED talks and the questions they evoke. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 1118–1127. Retrieved March 12, 2024, from https://aclanthology.org/2020.lrec-1.141/

Xiong, H., He, Z., Wu, H., & Wang, H. (2019). Modeling coherence for discourse neural machine translation [Issue: 01]. *Proceedings of the AAAI conference on artificial intelligence*, *33*, 7338–7345. Retrieved June 9, 2025, from https://ojs.aaai.org/index.php/AAAI/article/view/4721

Yung, F., Duh, K., Komura, T., & Matsumoto, Y. (2017). A psycholinguistic model for the marking of discourse relations [Publisher: The Dialogue & Discourse Board of Editors]. *Dialogue and Discourse*, *8*(1), 106–131. Retrieved June 7, 2025, from https://www.research.ed.ac.uk/en/publications/a-psycholinguistic-model-for-the-marking-of-discourse-relations

Zeyrek, D., Mendes, A., Grishina, Y., Kurfalı, M., Gibbon, S., & Ogrodniczuk, M. (2020). TED multilingual discourse bank (TED-MDB): A parallel corpus annotated in the PDTB style. *Language Resources and Evaluation*, *54*(2), 587–613. https://doi.org/10.1007/s10579-019-09445-9