

# **Harmonizing Divergence in Computational Discourse Analysis**

**Nelson Filipe Costa**

A Thesis in the Department of  
Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements  
for the Degree of

**Doctor of Philosophy in Computer Science**

Concordia University  
Montréal, Québec, Canada

November 2025

© Nelson Filipe Costa, 2025

CONCORDIA UNIVERSITY  
School of Graduate Studies

This is to certify that the thesis prepared

By: **Nelson Filipe Costa**  
Entitled: **Harmonizing Divergence in Computational Discourse Analysis**

and submitted in partial fulfillment of the requirements for the degree of

**Doctor of Philosophy in Computer Science**

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

\_\_\_\_\_ Chair  
*Dr. Suryadipta Majumdar*

\_\_\_\_\_ External Examiner  
*Dr. Giuseppe Carenini*

\_\_\_\_\_ Examiner  
*Dr. Nizar Bouguila*

\_\_\_\_\_ Examiner  
*Dr. Tristan Glatard*

\_\_\_\_\_ Examiner  
*Dr. Mirco Ravanelli*

\_\_\_\_\_ Supervisor  
*Dr. Leila Kosseim*

Approved by \_\_\_\_\_  
Tse-Hsun (Peter) Chen, Graduate Program Director  
Department of Computer Science and Software Engineering

\_\_\_\_\_ 2025

\_\_\_\_\_  
Mourad Debbabi, Dean  
Faculty of Engineering and Computer Science

# Abstract

## Harmonizing Divergence in Computational Discourse Analysis

**Nelson Filipe Costa, PhD**

**Concordia University, 2025**

Understanding discourse is essential for advancing computational models from surface-level text processing to deeper language reasoning, as it captures the logical flow of ideas that shapes meaning into a coherent text. However, progress in computational discourse analysis is hindered by divergent theoretical frameworks, ambiguity in implicit discourse relations and a myopic focus on the English language.

This thesis addresses these challenges through three research objectives. First, it proposes an empirical mapping between the two most widely used discourse frameworks, the Rhetorical Structure Theory and the Penn Discourse Treebank, for explicit and implicit discourse relations. The proposed mapping successfully maps 80.0% of the overlapping annotations between the most prominent corpora following each framework, laying groundwork for cross-framework interoperability. Second, the thesis introduces a novel multi-task classification model, MTask, for Implicit Discourse Relation Recognition (IDRR). The model captures ambiguity in implicit relations by jointly learning multi-label representations of their senses. The model establishes the first benchmark on multi-label IDRR and is also evaluated on the traditional single-label IDRR. Third, the thesis extends the multi-label approach to different languages and presents a hierarchical classification model. The model outperforms MTask in the English language and establishes the first benchmark on multilingual and multi-label IDRR. The thesis further explores prompting strategies using recent large language models and shows that fine-tuning strategies still perform better in this task.

Together, these contributions advance the goal of harmonizing divergence in computational discourse analysis, offering more generalizable and inclusive methods for discourse modeling across frameworks, ambiguity and languages.

# Acknowledgments

Em Novembro de 2018, quando entreguei a minha tese de mestrado, escrevi na dedicatória que se naquele momento tinha chegado onde cheguei, foi porque tinha subido aos ombros de gigantes. Não os mesmos gigantes a que Isaac Newton se referia, pois embora muito grandes que sejam, não chegam aos calcanhares dos gigantes a que me referia - os meus pais. Agora, em Outubro de 2025, momento em que entrego a minha tese de doutoramento, gostava de dedicar não só este feito uma vez mais aos meus gigantes pais, mas todos os feitos que tive a sorte de poder alcançar ao longo destes anos. Se hoje alcancei o que alcancei e se hoje sou o que sou, foi graças a vocês. Obrigado pelo apoio incondicional e incansável que sempre me fizeram sentir. Obrigado por nos momentos mais difíceis nunca me deixarem sentir sozinho. Obrigado por estarem sempre lá para me estenderem a mão e me ajudarem a levantar. Obrigado por serem os pais incríveis que são. E um obrigado muito especial também à minha irmã - uma irmã a quem tenho a sorte de poder chamar amiga e a quem muito em breve também poderei chamar de doutora. Obrigado por todo apoio e toda a cumplicidade que a cada ano cresce mais forte. Aos meus pais, à minha irmã, ao Gonçalo, ao meu tio e às minhas avós, um obrigado de coração cheio a todos vocês. Sem vocês nunca teria chegado tão longe. Sou um privilegiado.

Following my family, I would like to thank my supervisor, Dr. Leila Kosseim, who, in May 2022, took a chance on me and welcomed me into her lab - even though I knew nothing about Natural Language Processing at the time. Now, three years later, under her unwavering support and guidance, I am completing a PhD in this field. From learning what a *word embedding* is to receiving the best conference paper award for our final publication together - thank you for always being there to help me. I could not be more grateful to have had you as my supervisor on this academic journey. And what a journey it has been!

I would now like to thank the friends who have made this adventure in Montreal so special and all the challenges of a PhD so much more bearable. First, I want to thank Fábio - a person I have been lucky to call a friend for as long as I can remember and who joined me on this journey overseas. Thank you for the unconditional friendship and for bringing with you a piece of home and comfort to this side of the ocean. Second, I want to thank

Dylan - who I met here and quickly became such an important person in my life. Thank you for always being there when I needed it the most and for being the very definition of a friend. Je voudrais ensuite remercier les trois mousquetaires québécois, Phil, Max et Caro, que j'ai eu la chance de rencontrer il y a toutes ces années en Belgique et qui ont été la principale raison derrière ma décision de faire un doctorat à Montréal. À vous, ainsi qu'à tous les autres amis québécois que j'ai eu la chance de rencontrer au fil de ces années, Sandy, David, Sara, Pierre Tyson, Ashley, Pierre Goyette, Alex, Dominik, Simon, Myriam, Gab et Catherine, merci de m'avoir accueilli à bras ouverts et de m'avoir toujours fait sentir chez moi dans votre pays. Grâce à vous, je ne me suis jamais senti étranger au Québec - pas une seule fois. To my fellow international friends, Kathryn, Connor Brackley, Masood, Kelsey, Danielle, Guillaume, Camille Jankowski, Camille Lemauff, Aymeric, Connor Ross, Marine, Agnès, Hugo and Christina, I am incredibly grateful destiny brought our paths together in Montreal. Without you, discovering and exploring the wonders of Quebec would never have been such an unforgettable experience.

Finalmente, gostava de deixar um agradecimento especial a todos os meus amigos e pessoas queridas que nunca deixaram de me apoiar desde Portugal. Gostava de agradecer e partilhar esta conquista com o João, Tita e Mila, bem como com os meus padrinhos, Armando e Anabela, e com o Diogo, Alice, Filipa, António e Miguel - que embora não sejam família de sangue, é como se o fossem para mim. Gostava também de agradecer aos meus amigos de longa data, Guilherme, Xandy, Guida, Luísa, Albano, Rodrigo, Pedro, Fontinha, Raquel, Nuno, Francisco Saraiva, Kets e Catarina, bem como aos meus outros amigos de também longa (mas menos longa) data, Tiago, Diogo, Filipe, Francisco Horta, Gonçalo e toda a restante malta de MEFT. A todos vocês, muito obrigado por me aturarem todos estes anos e por todo o apoio que, embora à distância, sempre me fizeram sentir.

# Contents

|                                                            |            |
|------------------------------------------------------------|------------|
| <b>List of Figures</b>                                     | <b>ix</b>  |
| <b>List of Tables</b>                                      | <b>x</b>   |
| <b>List of Abbreviations</b>                               | <b>xii</b> |
| <b>1 Introduction</b>                                      | <b>1</b>   |
| 1.1 Motivation . . . . .                                   | 3          |
| 1.2 Research Objectives . . . . .                          | 5          |
| 1.3 Contributions . . . . .                                | 6          |
| 1.4 Thesis Outline . . . . .                               | 8          |
| <b>2 Discourse Frameworks and Corpora</b>                  | <b>9</b>   |
| 2.1 Rhetorical Structure Theory . . . . .                  | 10         |
| 2.1.1 RST Discourse Treebank . . . . .                     | 12         |
| 2.2 The Penn Discourse Treebank . . . . .                  | 13         |
| 2.2.1 PDTB 3.0 . . . . .                                   | 14         |
| 2.2.2 DiscoGeM . . . . .                                   | 17         |
| 2.2.3 DiscoGeM 2.0 . . . . .                               | 18         |
| 2.3 Other Frameworks . . . . .                             | 19         |
| <b>3 Mapping Explicit and Implicit Discourse Relations</b> | <b>21</b>  |
| 3.1 Previous Work . . . . .                                | 22         |
| 3.2 Methodology . . . . .                                  | 24         |
| 3.2.1 Segment Alignment . . . . .                          | 25         |
| 3.2.2 Relation Mapping . . . . .                           | 28         |
| 3.3 Results . . . . .                                      | 35         |
| 3.3.1 Mapping of Explicit Discourse Relations . . . . .    | 37         |
| 3.3.2 Mapping of Implicit Discourse Relations . . . . .    | 38         |

|          |                                                             |           |
|----------|-------------------------------------------------------------|-----------|
| 3.4      | Discussion . . . . .                                        | 39        |
| <b>4</b> | <b>Multi-Label Implicit Discourse Relation Recognition</b>  | <b>41</b> |
| 4.1      | Previous Work . . . . .                                     | 43        |
| 4.2      | Data Preparation . . . . .                                  | 46        |
| 4.2.1    | DiscoGeM . . . . .                                          | 46        |
| 4.2.2    | PDTB 3.0 . . . . .                                          | 48        |
| 4.2.3    | Corpora Agreement . . . . .                                 | 49        |
| 4.3      | Preliminary Study . . . . .                                 | 50        |
| 4.4      | Methodology . . . . .                                       | 52        |
| 4.4.1    | Multi-Label Classification . . . . .                        | 53        |
| 4.4.2    | Single-Label Classification . . . . .                       | 53        |
| 4.5      | Experiments . . . . .                                       | 54        |
| 4.5.1    | Model and Loss Function Selection . . . . .                 | 54        |
| 4.5.2    | Fine-Tuning of Hyperparameters . . . . .                    | 56        |
| 4.6      | Results . . . . .                                           | 57        |
| 4.6.1    | Results on DiscoGeM . . . . .                               | 58        |
| 4.6.2    | Transfer Learning Results on PDTB 3.0 . . . . .             | 59        |
| 4.7      | Further Analysis . . . . .                                  | 60        |
| 4.7.1    | Individual Sense Results . . . . .                          | 60        |
| 4.7.2    | Sense Coherence across Levels . . . . .                     | 63        |
| 4.8      | Discussion . . . . .                                        | 63        |
| <b>5</b> | <b>Multilingual Implicit Discourse Relation Recognition</b> | <b>66</b> |
| 5.1      | Previous Work . . . . .                                     | 67        |
| 5.2      | Data Preparation . . . . .                                  | 70        |
| 5.3      | Methodology . . . . .                                       | 70        |
| 5.3.1    | Fine-Tuned Classification Model . . . . .                   | 72        |
| 5.3.2    | LLM Prompting Framework . . . . .                           | 74        |
| 5.4      | Experiments . . . . .                                       | 74        |
| 5.5      | Results . . . . .                                           | 77        |
| 5.6      | Further Analysis . . . . .                                  | 80        |
| 5.6.1    | Multi-Task Architecture . . . . .                           | 81        |
| 5.6.2    | Hierarchical Learning . . . . .                             | 82        |
| 5.7      | Discussion . . . . .                                        | 82        |

|                             |           |
|-----------------------------|-----------|
| <b>6 Conclusion</b>         | <b>85</b> |
| 6.1 Contributions . . . . . | 85        |
| 6.2 Future Work . . . . .   | 89        |
| <b>Bibliography</b>         | <b>91</b> |

# List of Figures

|             |                                                                       |    |
|-------------|-----------------------------------------------------------------------|----|
| Figure 2.1  | Examples of the five schemas in the RST framework. . . . .            | 11 |
| Figure 2.2  | RST tree for the text in Example 4. . . . .                           | 11 |
| Figure 2.3  | RST tree for the text in Example 5. . . . .                           | 12 |
| Figure 3.1  | Example of a complex mapping. . . . .                                 | 29 |
| Figure 3.2  | RST tree exemplifying the nucleus path. . . . .                       | 30 |
| Figure 3.3  | Examples of a perfect relation mapping. . . . .                       | 30 |
| Figure 3.4  | Example of a perfect relation mapping. . . . .                        | 31 |
| Figure 3.5  | Example of an imperfect relation mapping. . . . .                     | 31 |
| Figure 3.6  | Example of an imperfect relation mapping. . . . .                     | 32 |
| Figure 3.7  | Example of an impossible mapping due to unclear nucleus path. . .     | 32 |
| Figure 3.8  | Example of an impossible mapping due to multinuclear relation. . .    | 33 |
| Figure 3.9  | Example of an imperfect mapping due to discontinuous segment. . .     | 34 |
| Figure 3.10 | Example of an impossible mapping due to embedded relation. . . .      | 35 |
| Figure 4.1  | Distribution of majority label senses in the DiscoGeM corpus. . . .   | 48 |
| Figure 4.2  | Distribution of different context sizes in the DiscoGeM corpus. . . . | 50 |
| Figure 4.3  | Architecture of the MTask classification model. . . . .               | 53 |
| Figure 4.4  | Confusion matrix for single-label IDRR at level-1 on DiscoGeM. . .    | 62 |
| Figure 4.5  | Confusion matrix for single-label IDRR at level-2 on DiscoGeM. . .    | 62 |
| Figure 5.1  | Architecture of the HArch classification model. . . . .               | 73 |
| Figure 5.2  | Template of the few-shot prompting framework for multi-label IDRR. .  | 75 |

# List of Tables

|            |                                                                                    |    |
|------------|------------------------------------------------------------------------------------|----|
| Table 2.1  | Original set of discourse relations in the RST framework. . . . .                  | 11 |
| Table 2.2  | Inter-annotator agreement in the RST-DT corpus. . . . .                            | 13 |
| Table 2.3  | Hierarchy of sense labels in the PDTB 3.0 framework. . . . .                       | 15 |
| Table 2.4  | Distribution of discourse relations in the PDTB 3.0 corpus. . . . .                | 16 |
| Table 2.5  | Distribution of implicit discourse relations in the DiscoGeM 2.0 corpus. . . . .   | 18 |
| Table 3.1  | Overlapping relations in the RST-DT and the PDTB 3.0 corpora. . . . .              | 25 |
| Table 3.2  | Distribution of alignments between continuous segments. . . . .                    | 27 |
| Table 3.3  | Distribution of alignments between discontinuous segments. . . . .                 | 28 |
| Table 3.4  | Quantitative mapping results of explicit and implicit discourse relations. . . . . | 35 |
| Table 3.5  | Mapping results for the explicit discourse relations. . . . .                      | 38 |
| Table 3.6  | Mapping results for the implicit discourse relations. . . . .                      | 39 |
| Table 4.1  | Distribution of senses in the DiscoGeM corpus. . . . .                             | 47 |
| Table 4.2  | Distribution of level-2 majority label senses on the test sets. . . . .            | 49 |
| Table 4.3  | Annotation agreement between DiscoGeM and PDTB 3.0. . . . .                        | 49 |
| Table 4.4  | Fine-tuning of epochs and batch size for preliminary study. . . . .                | 51 |
| Table 4.5  | Fine-tuning of learning and decay rates for preliminary study. . . . .             | 51 |
| Table 4.6  | Results of single-label IDRR with different contexts. . . . .                      | 52 |
| Table 4.7  | Model and loss selection for multi-label and single-label IDRR. . . . .            | 56 |
| Table 4.8  | Fine-tuning of hyperparameters for multi-label and single-label IDRR. . . . .      | 57 |
| Table 4.9  | Results of multi-label and single-label IDRR on DiscoGeM. . . . .                  | 58 |
| Table 4.10 | Transfer learning results of single-label IDRR on PDTB 3.0. . . . .                | 59 |
| Table 4.11 | Individual results of single-label IDRR on DiscoGeM and PDTB 3.0. . . . .          | 61 |
| Table 4.12 | Coherence results of single-label IDRR on DiscoGeM and PDTB 3.0. . . . .           | 64 |
| Table 5.1  | Distribution of level-1 senses in the DiscoGeM 2.0 corpus. . . . .                 | 71 |
| Table 5.2  | Distribution of level-2 and level-3 senses in the DiscoGeM 2.0 corpus. . . . .     | 71 |
| Table 5.3  | Model selection for multilingual and multi-label IDRR. . . . .                     | 77 |
| Table 5.4  | Results of multi-label IDRR using MTask and HArch on DiscoGeM. . . . .             | 78 |
| Table 5.5  | Results of multilingual and multi-label IDRR. . . . .                              | 79 |

|           |                                                             |    |
|-----------|-------------------------------------------------------------|----|
| Table 5.6 | Evaluation of the multi-task architecture of HArch. . . . . | 81 |
| Table 5.7 | Evaluation of the learnable parameters of HArch. . . . .    | 82 |

# List of Abbreviations

|              |                                              |
|--------------|----------------------------------------------|
| <b>CE</b>    | Cross-Entropy                                |
| <b>EDU</b>   | Elementary Discourse Unit                    |
| <b>eRST</b>  | Enhanced Rhetorical Structure Theory         |
| <b>HArch</b> | Hierarchical Multi-Task Classification Model |
| <b>IDRR</b>  | Implicit Discourse Relation Recognition      |
| <b>JS</b>    | Jensen–Shannon                               |
| <b>LLM</b>   | Large Language Model                         |
| <b>MAE</b>   | Mean Absolute Error                          |
| <b>MSE</b>   | Mean Squared Error                           |
| <b>MTask</b> | Multi-Task Classification Model              |
| <b>NLP</b>   | Natural Language Processing                  |
| <b>PDTB</b>  | Penn Discourse Treebank                      |
| <b>PLM</b>   | Pre-Trained Language Model                   |
| <b>RST</b>   | Rhetorical Structure Theory                  |
| <b>SDRT</b>  | Segmented Discourse Representation Theory    |
| <b>SOTA</b>  | State-of-the-Art                             |

# Chapter 1

## Introduction

Understanding text is a fundamental challenge in the field of Natural Language Processing (NLP). To grasp the full meaning of a text, it is essential to move beyond the interpretation of individual sentences and capture the logical and structured flow of ideas that connects its parts into a coherent and cohesive whole (Hobbs, 1979). Textual coherence and cohesiveness can be achieved through the coordination of various linguistic mechanisms, including discourse relations to signal logical and semantic links between segments of a text and coreference consistency between expressions and entities throughout a text (Sanders et al., 1992; Marcu, 2000; Jurafsky and Martin, 2008). Together, these mechanisms support the construction of a structured discourse and are central to the development of computational models capable of understanding and generating human language.

Among these mechanisms, discourse relations occupy a particularly central role and form the primary focus of this thesis. Discourse relations allow readers to interpret how different ideas of a text relate to one another. Consider the examples below:

**Example 1.** The roads were slippery because it had been raining all night.

**Example 2.** It was raining heavily, but the marathon continued as planned.

In Example 1, there is a causal relation between the roads being slippery and the fact that it had been raining all night introduced by the discourse connective *because*. In Example 2, there is a relation of contrast introduced by the discourse connective *but* between the marathon continuing as planned and the adverse raining conditions. While such discourse relations are often made explicit through the use of connectives, they can also be left implicit, requiring readers to infer the underlying connection based on linguistic context and background knowledge. Understanding and modeling discourse relations is essential for computational models that aim to move beyond shallow text processing toward deeper language understanding, enabling applications such as summarization, machine translation,

question-answering and dialogue systems to produce more coherent and contextually appropriate responses.

The computational study of discourse relations has given rise to multiple theoretical frameworks, each offering a distinct perspective on how coherence is structured and interpreted. One of the earliest and most widely adopted approaches is the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), which models texts as a hierarchical structure of text segments connected by discourse relations. In contrast, the Penn Discourse Treebank (PDTB) (Miltsakaki et al., 2004; Prasad et al., 2008) introduced a lexically grounded approach in which discourse relations are annotated primarily through the identification of discourse connectives. Unlike the hierarchical organization of the RST, the PDTB adopts a shallow annotation strategy, treating relations as flat connections between text segments without imposing a global structure on the discourse. Another influential framework, the Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003; Lascarides and Asher, 2007) represents discourse as a directed acyclic graph, where nodes correspond to text segments and edges represent discourse relations. SDRT combines discourse structure with formal semantics, introducing complex relations and constraints on how segments can be attached within the graph. More recently, the Enhanced Rhetorical Structure Theory (eRST) (Zeldes et al., 2025) has sought to enrich the original RST framework by allowing non-binary tree structures, permitting multiple relations between the same text segments and supporting non-adjacent and discontinuous attachments that break strict hierarchical constraints.

While they share the common goal of capturing the relational structures that ground textual coherence, the existence of multiple and divergent discourse frameworks presents a significant challenge for both theoretical research and the development of computational models of discourse (Zeldes et al., 2019, 2021; Braud et al., 2023). Since each framework adopts distinct approaches towards the annotation of discourse relations, work across frameworks is not easily comparable and progress toward generalized models capable of handling diverse discourse phenomena is limited. As a result, discourse modeling in NLP remains fragmented, with models often tailored to specific theoretical perspectives. Addressing this fragmentation requires investigating the extent to which discourse relations can be mapped across frameworks, identifying points of convergence and developing approaches that move toward greater theoretical and computational harmonization.

A further challenge in computational discourse analysis arises from the fact that many discourse relations are left implicit, with no explicit connective to signal the relationship between text segments. In everyday language, writers frequently omit explicit markers, relying on readers to infer the intended relation based on linguistic cues, world knowledge

and shared assumptions. For computational models, however, the absence of such explicit cues makes the identification and classification of implicit discourse relations significantly more difficult. Consider the example below:

**Example 3.** The lights flickered. The power went out.

In this example, it is not clear whether the two ideas of the relation are related by a temporal or causal relation. One could interpret that the lights flicked and *then* the power went out, or that the lights flickered *because* the power went out. The two interpretations are valid and may even co-exist. The task of disambiguating implicit discourse relations, known as Implicit Discourse Relation Recognition (IDRR), remains one of the most challenging problems in computational discourse analysis. Moreover, research on IDRR has predominantly been conducted on the English language exclusively, leaving open the exploration of computational discourse analysis to multilingual contexts.

This thesis addresses these theoretical and computational challenges by investigating (i) how discourse relations can be mapped across frameworks, (ii) how ambiguous implicit discourse relations can be modeled with multiple senses and (iii) how discourse modeling can be extended to multilingual contexts. By tackling these problems, this thesis aims to contribute to the development of more generalizable approaches to computational discourse analysis with the broader goal of harmonizing divergence within the field.

## 1.1 Motivation

The ability to model and understand discourse is important for the success of a wide range of NLP applications. Many downstream tasks, such as text summarization (Dong et al., 2021; Huang and Kurohashi, 2021; Chen and Yang, 2021; Zhang et al., 2023; Liu et al., 2023a; Liu and Demberg, 2024; Zhong and Litman, 2025; Guan et al., 2025), machine translation (Jiang et al., 2023b; Hu and Wan, 2023; Wang et al., 2023), text classification (Huber and Carenini, 2020; Choubey et al., 2020; Choubey and Huang, 2021; Devatine et al., 2022; Lei et al., 2022; Li and Huang, 2023; Lei and Huang, 2023), question-answering (Huang et al., 2021; Xu et al., 2022; Du et al., 2023; Jahangir et al., 2024) and dialogue systems (Chi and Rudnicky, 2022; Li et al., 2023, 2024a,b; Xu et al., 2024; Mehndiratta and Asawa, 2025), require not only the interpretation of individual sentences but also the ability to capture how sentences and ideas relate to one another. This need has become even more pressing with recent advances in large language models (LLMs), where

increasingly large context windows demand a deeper understanding of discourse to ensure coherent results across longer stretches of text in downstream NLP applications.

For instance, in text summarization, Liu et al. (2023a) and Liu and Demberg (2024) integrate RST discourse information to identify key content and the differences in importance between sentences of a text for long document summarization. Similarly, Zhong and Litman (2025) decompose long texts into discourse-inspired chunks and integrate discourse information to enhance factual consistency in long document summarization with complex sentences. In machine translation, Jiang et al. (2023b) show that translating documents requires a deeper understanding of discourse phenomena, while Hu and Wan (2023) build a discourse parser to infer context boundaries and improve translation coherence over sentence-level models and Wang et al. (2023) investigate how prompting language models with discourse knowledge can improve document-level translation. In text classification, Huber and Carenini (2020) present a method for improving sentiment analysis by using automatically generated discourse structures that reflect how ideas are organized in long documents with complex rhetorical structures and Lei and Huang (2023) leverage both local and global discourse structures through feature concatenation and knowledge distillation techniques to improve fine-grained propaganda identification. In question-answering, Huang et al. (2021) propose a discourse-aware graph network that uses the discourse structure of source texts to support logical reasoning, while Du et al. (2023) build a hierarchical graph network capturing relations between sentences to propagate information and improve answer prediction over long documents. In dialogue systems, Li et al. (2023) extract latent discourse structures from the attention matrices of pre-trained language models (PLMs) and leverage these structures to improve dialogue discourse parsing through unsupervised and semi-supervised methods. Building on this line of work, Li et al. (2024b) reframe dialogue discourse parsing as a generation task, enabling a sequence-to-sequence LLM to explicitly generate the discourse relations between dialogue units and capture the structural organization of the conversation.

The examples above illustrate how discourse plays an important role in improving the coherence and overall quality of NLP models across a variety of applications even in the era of LLMs. As models are increasingly expected to handle longer and more complex texts, the ability to capture discourse structures becomes even more essential. However, in order to move towards more capable discourse models, some key challenges remain: existing approaches to computational discourse are often tied to divergent theoretical frameworks, struggle with the ambiguity of implicit discourse relations and are predominantly limited to the English language. The next section presents the specific research objectives of this thesis, which aim to address these challenges.

## 1.2 Research Objectives

Building on the challenges outlined in the previous section, this thesis sets out to address three sources of divergence that currently limit the development of generalizable models for computational discourse analysis. More precisely, the research focuses on (i) bridging theoretical frameworks, (ii) improving the modeling of implicit discourse relations and (iii) extending discourse analysis beyond monolingual settings.

The first source of divergence arises at the most fundamental level: the lack of consensus on a theoretical framework for modeling discourse. As indicated earlier, multiple frameworks have been proposed to capture discourse coherence, each reflecting different theoretical assumptions and annotation practices. However, no common ground has yet been established, creating obstacles for the comparison, evaluation and interoperability of discourse models. Therefore, the first research objective of this thesis is to establish an empirical mapping between the two most widely used discourse frameworks, the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) and the Penn Discourse Treebank (PDTB) (Miltsakaki et al., 2004; Prasad et al., 2008), for explicit and implicit discourse relations.

**Objective 1.** Map explicit and implicit discourse relations between the RST and the PDTB frameworks.

The second source of divergence concerns the disambiguation of implicit discourse relations within the PDTB framework itself. The task of distinguishing between possible senses of implicit discourse relations, referred to as Implicit Discourse Relation Recognition (IDRR), remains one of the most challenging problems in computational discourse analysis. Traditional approaches to IDRR have followed the general assumption that each discourse relation should be assigned a single sense. This thesis challenges this assumption by arguing that multiple senses may co-exist within the same discourse relation. To this end, the second research objective of this thesis is to develop the first multi-label IDRR classification model capable of capturing the ambiguity of implicit discourse senses.

**Objective 2.** Develop a multi-label IDRR classification model capable of capturing multiple implicit senses.

The third and final source of divergence addressed in this thesis relates to the predominantly monolingual focus of existing IDRR research. Due to the limited availability of annotated discourse corpora in languages other than English, the vast majority of existing work has been conducted in the English language alone, with no research exploring IDRR

across multiple languages simultaneously. The third research objective of this thesis is therefore to develop the first multi-lingual and multi-label IDRR classification model capable of recognizing implicit discourse relations across different languages, while accounting for the ambiguity in their interpretation.

**Objective 3.** Develop a multilingual IDRR classification model capable of capturing multiple implicit senses across different languages.

Together, these three research objectives define the focus of this doctoral thesis. By addressing these challenges, this research aims to contribute to the harmonization of divergence in computational discourse analysis.

## 1.3 Contributions

This doctoral thesis is grounded in four peer-reviewed research papers following the research objectives described in the previous section. Together, they contribute to bridging theoretical divergences between discourse frameworks, improving the modeling of ambiguous implicit discourse relations and extending computational discourse analysis to multi-lingual settings. In addition to the theoretical contributions, publicly available code and resources have been released to support further research in the field. The main contributions of each paper are summarized below.

- (1) The first paper was published at the Recent Advances in Natural Language Processing (RANLP) 2023 (Costa et al., 2023c) and proposes a mapping algorithm between the RST and the PDTB 3.0 discourse frameworks. The proposed algorithm was able to increase the number of mapped discourse relations compared against the previous mapping effort between the RST and the PDTB 2.0 frameworks. The analysis revealed that while implicit discourse relations are easier to align at the segment level, the mapping of their senses is less consistent. These findings provide a deeper understanding of the structural divergences between the two discourse frameworks and offer a robust foundation for cross-framework discourse modeling. GitHub: <https://github.com/CLaC-Lab/Mapping-Discourse-Relations>
- (2) The second paper was published at the Workshop on Computational Approaches to Discourse (CODI) 2024 (Costa and Kosseim, 2024) and explores the potential of training single-label IDRR models on multi-label annotated data. The proposed model was trained on multi-label distributions but evaluated on the traditional single-label setting to ensure comparability with existing IDRR models. The results of this

paper suggested that training on multi-label annotations can help models better generalize and more effectively capture ambiguous implicit discourse relations, providing a first step towards multi-label IDRR. GitHub: <https://github.com/CLaC-Lab/Implicit-Discourse-Relation-Recognition>

- (3) The third paper was published at the Special Interest Group on Discourse and Dialogue (SIGDIAL) 2025 (Costa and Kosseim, 2025a) and introduces the first multi-label approach for IDRR. A multi-task model, MTask, was trained to simultaneously predict multi-label discourse senses across all three sense levels of the PDTB 3.0 framework. The same model was also adapted to the traditional single-label IDRR setting by selecting the sense with the highest probability in the multi-label representation. This work established the first benchmark on multi-label IDRR. It additionally presented the first cross-corpus evaluation of transfer learning between DiscoGeM and PDTB 3.0, revealing both the potential and current limitations of zero-shot generalization between both corpora. GitHub: <https://github.com/nelsonfilipecosta/Implicit-Discourse-Relation-Recognition>
- (4) The fourth paper was also published at the Special Interest Group on Discourse and Dialogue (SIGDIAL) 2025 (Costa and Kosseim, 2025b) and presents the first multilingual and multi-label classification model for IDRR. The classification model, HArch, was fine-tuned to leverage dependencies across the PDTB 3.0 sense hierarchy in different languages in a monolingual and multilingual setting. This model achieves state-of-the-art results and outperforms our previous multi-label IDRR model in the English language and establishes the first benchmark on multilingual IDRR. This paper also evaluated the performance of GPT-4o and LLaMA-4-Maverick in few-shot settings, showing that HArch not only outperformed LLMs in the context of IDRR but also offered significantly lower computational and environmental costs. These results demonstrate the effectiveness and sustainability of fine-tuned encoder-based models for computational discourse analysis. GitHub: <https://github.com/nelsonfilipecosta/Multi-Lingual-Implicit-Discourse-Relation-Recognition>

In addition to these four publications that form the foundation of this thesis, two other peer-reviewed research papers were published concurrently with this work. Although these papers fall outside the primary scope of this thesis, they reflect complementary expertise developed during the same period. The first paper focuses on the automatic identification of paired discourse connectives in Chinese (Costa et al., 2023a), while the second addresses the automatic detection of online persuasion techniques in a multilingual setting (Costa

et al., 2023b). Together, these additional papers further demonstrate the breadth of research conducted throughout the doctoral program.

## **1.4 Thesis Outline**

Chapter 2 presents the necessary theoretical background to this thesis. It introduces the RST and the PDTB discourse frameworks and the discourse annotated corpora relevant to this work. Chapter 3 addresses the first research objective by presenting an empirical mapping of explicit and implicit discourse relations between the RST and the PDTB frameworks. This chapter is based on the work published in Costa et al. (2023c). Chapter 4 addresses the second research objective by presenting a multi-task and multi-label IDRR classification model capable of capturing the ambiguity and multiplicity of implicit discourse senses. This chapter is based on the work published in Costa and Kosseim (2024) and Costa and Kosseim (2025a). Chapter 5 addresses the third research objective by presenting the development of a multilingual and multi-label IDRR classification model capable of recognizing multiple implicit senses across different languages. This chapter is based on the work published in Costa and Kosseim (2025b). Finally, Chapter 6 concludes the thesis and outlines potential directions for future research.

## Chapter 2

# Discourse Frameworks and Corpora

Computational discourse analysis in NLP investigates the structures and relationships within text that enable the coherent communication of ideas. It models how textual units, such as clauses, sentences or larger segments of text, interact to convey meaning and intent beyond the level of isolated sentences. Over the years, different theoretical frameworks have been proposed to describe discourse structure. Among them, the two most influential and widely adopted are the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) and the Penn Discourse Treebank (PDTB) (Miltsakaki et al., 2004; Prasad et al., 2008) frameworks. The RST is a linguistically motivated framework that represents discourse as a hierarchical structure, capturing how all parts of a text contribute to its overall coherence. In contrast, the PDTB adopts a more computationally driven approach, annotating discourse relations as flat pairwise connections between text segments anchored around the idea of discourse connectives. These frameworks reflect distinct theoretical perspectives on how coherence is structured and each has led to the development of discourse annotated corpora widely used in the training of computational discourse models.

This chapter introduces the RST and the PDTB frameworks in detail, as well as discourse annotated corpora that follow them and are which used throughout this thesis. Namely, the RST Discourse Treebank (RST-DT) (Carlson et al., 2001, 2003), which follows the RST framework, and the PDTB 3.0 (Webber et al., 2019), DiscoGeM (Scholman et al., 2022a) and DiscoGeM 2.0 (Yung et al., 2024b), all of which are annotated following the PDTB framework. These corpora provide the foundation for the models and analyses presented in the following chapters. More precisely, the RST-DT and the PDTB 3.0 corpora will be used to address the first research objective on cross-framework mapping in Chapter 3, the PDTB 3.0 and DiscoGeM will be used to address the second research objective on multi-label IDRR in Chapter 4 and DiscoGeM and DiscoGeM 2.0 will be used to address the third research objective on multilingual and multi-label IDRR in Chapter 5.

## 2.1 Rhetorical Structure Theory

The Rhetorical Structure Theory (RST) framework (Mann and Thompson, 1988) considers discourse to be a hierarchical organization of discourse relations between text segments. In the RST, a text is first segmented into minimal syntactic units, referred to as **elementary discourse units (EDUs)**. These units are then recursively connected through an open set of discourse relations, guided by a predefined set of schemas. The application of these relations continues until all units are integrated into a single hierarchical structure. Thus, the RST framework models discourse by proposing a hierarchical structure in which every part of a text has a role to play with respect to other parts of the text (Taboada and Mann, 2006).

There are two main types of EDUs: the **nucleus** and the **satellite**. A nucleus expresses the core information of a discourse relation and can typically stand alone as a complete idea. In contrast, a satellite conveys supporting information and depends on the nucleus to form a complete idea. When establishing a discourse relation between two EDUs, it is therefore crucial to determine which unit serves as the nucleus and which as the satellite. Consider the text in Example 4. According to the RST framework, this text can be segmented into two EDUs, where EDU1 functions as the nucleus and EDU2 as the satellite providing evidence for EDU1. By convention, EDUs are enclosed in square brackets and superscript numbers indicate their sequential order in the text.

**Example 4.** [The striker played really well.]<sup>1</sup> [He was able to score a hattrick.]<sup>2</sup>

In this example, since the satellite EDU2 provides supporting evidence for the nucleus EDU1, the RST framework establishes a **mononuclear** discourse relation between both EDUs. Most relations in the RST framework are mononuclear, where a satellite is connected to a nucleus. However, when both EDUs are considered equally important to the discourse relation, the relation is considered **multinuclear** - with neither unit being considered a satellite to the other. The original open set of discourse relations in the RST framework is presented in Table 2.1. With the exception of the last three relations, SEQUENCE, CONTRAST and JOINT, which are multinuclear, all other relations are mononuclear. Since its creation, additional relations have been added to this set.

All discourse relations in Table 2.1 can be represented by one of the five different schemas illustrated in Figure 2.1. Mononuclear relations in the RST framework are represented by a curved arrow going from the satellite to the nucleus of the relation, which is also defined by a vertical line. All mononuclear relations follow the general schema shown in Figure 2.1a, with the exception of the MOTIVATION and ENABLEMENT relations - which

| Discourse Relations |                       |                |             |
|---------------------|-----------------------|----------------|-------------|
| CIRCUMSTANCE        | EVIDENCE              | PURPOSE        | EVALUATION  |
| SOLUTIONHOOD        | JUSTIFY               | ANTITHESIS     | RESTATEMENT |
| ELABORATION         | VOLITIONAL CAUSE      | CONCESSION     | SUMMARY     |
| BACKGROUND          | NON-VOLITIONAL CAUSE  | CONDITION      | SEQUENCE    |
| ENABLEMENT          | VOLITIONAL RESULT     | OTHERWISE      | CONTRAST    |
| MOTIVATION          | NON-VOLITIONAL RESULT | INTERPRETATION | JOINT       |

Table 2.1: Original open set of discourse relations in the RST framework. Table adapted from (Mann and Thompson, 1988).

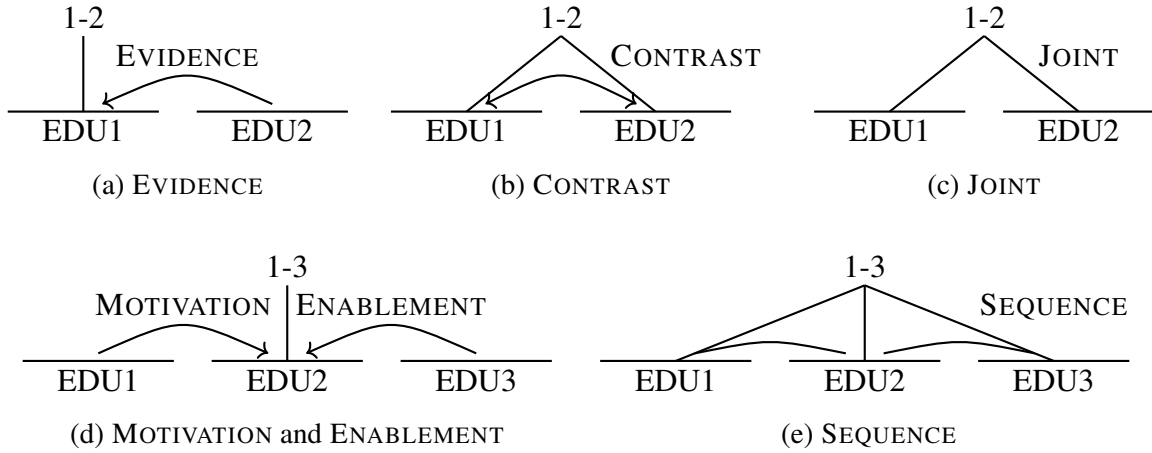


Figure 2.1: Examples of the five schemas in the RST framework. Figure adapted from (Mann and Thompson, 1988).

follow the schema in Figure 2.1d. Conversely, the CONTRAST, JOINT and SEQUENCE multinuclear relations are represented by the schemas in Figures 2.1b, 2.1c and 2.1e, respectively.

Based on the set of discourse relations listed in Table 2.1 and the schema representations in Figure 2.1, the discourse relation in Example 4 corresponds to the EVIDENCE relation in the RST framework and is represented by the schema shown in Figure 2.2.

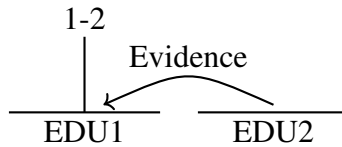


Figure 2.2: RST tree for the text in Example 4.

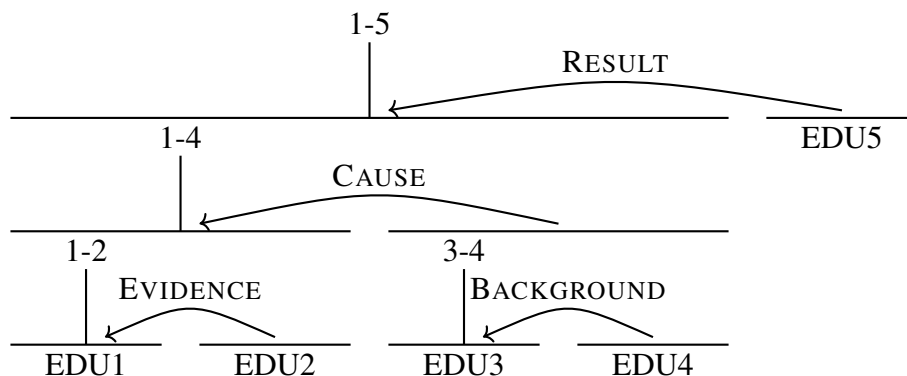


Figure 2.3: RST tree for the text in Example 5.

In the RST framework, these schemas and discourse relations are then applied recursively until the entire text is integrated in a single hierarchical structure. Each relation incrementally builds the discourse tree by connecting smaller units into larger spans. For instance, the discourse structure of the text in Example 5 can be fully represented by the RST tree shown in Figure 2.3.

**Example 5.** [The striker played really well,]<sup>1</sup> [as he was able to score a hattrick.]<sup>2</sup> [The opposing team struggled to defend against him]<sup>3</sup> [because their defense had several key players injured.]<sup>4</sup> [As a result, the match ended with a big win for his team.]<sup>5</sup>

### 2.1.1 RST Discourse Treebank

The RST Discourse Treebank (RST-DT) (Carlson et al., 2001, 2003) is a discourse annotated corpus following the RST framework and comprising 385 Wall Street Journal articles. It contains a total of 21,789 EDUs and 20,017 discourse relations. The average number of words per EDU is 8.1 and the average number of EDUs per article is 56.59. The corpus was developed to provide a practical and scalable resource for discourse analysis, aiming to maintain annotation consistency across a large set of documents. Since previous research (Marcu et al., 1999) had shown high inter-annotator agreement following the tree structure of the RST framework, it was adopted for the annotation of the RST-DT corpus.

The corpus was annotated by multiple professional language analysts with prior experience in data annotation and extensive hands-on training in the RST framework. To ensure consistent annotation on a large scale, grammatical clauses were primarily chosen as EDUs in the annotation of the corpus. Once the EDUs were identified, adjacent spans were connected via discourse relations to form a hierarchical structure. A total of 53 mononuclear and 25 multinuclear relations were used in the annotation of the RST-DT corpus. Table 2.2

| Annotators | Kappa Coefficient |             |            |           | No. of Docs | Avg. No. of EDUs |
|------------|-------------------|-------------|------------|-----------|-------------|------------------|
|            | EDU Segm.         | Hier. Spans | Nuclearity | Relations |             |                  |
| B, E       | 0.96              | 0.89        | 0.85       | 0.78      | 7           | 88.3             |
| A, E       | 0.98              | 0.90        | 0.84       | 0.76      | 6           | 57.7             |
| A, B       | 1.00              | 0.93        | 0.88       | 0.79      | 5           | 58.2             |
| A, C       | 0.95              | 0.84        | 0.78       | 0.68      | 4           | 116.5            |
| A, F       | 0.95              | 0.78        | 0.69       | 0.60      | 4           | 26.5             |
| A, D       | 1.00              | 0.87        | 0.80       | 0.72      | 4           | 23.3             |

Table 2.2: Inter-annotator agreement between six pair of annotators (A, B, C, D, E and F) in the RST-DT who double-annotated four or more documents in terms of the Kappa coefficient. Table adapted from (Carlson et al., 2001).

shows the inter-annotator agreement at four levels: EDU segmentation, hierarchical span identification, nuclearity assignment and relation labeling. The inter-annotator agreement was calculated using the Kappa coefficient for hierarchical structures (Marcu et al., 1999) between six pairs of annotators who double-annotated four or more documents. In this table, a score of 0 indicates no agreement between the annotators, while a score of 1 reflects complete agreement.

## 2.2 The Penn Discourse Treebank

The Penn Discourse Treebank (PDTB) framework (Miltsakaki et al., 2004; Prasad et al., 2008) is one of the most widely used frameworks and considers discourse to be a flat representation of relations anchored around the idea of **discourse connectives**. Discourse connectives are semantic cues, such as the words *but* and *because*, that signal a discourse relation between two text segments, referred to as **arguments**. While arguments in the PDTB framework serve a role analogous to EDUs in the RST framework, they are not constrained to minimal syntactic units. Instead, an argument can be a clause, a sentence or a larger segment of text, as long as it conveys sufficient information to interpret the discourse relation. However, a minimality principle still requires an argument to contain only the minimal amount of information to interpret the relation - avoiding the inclusion of superfluous content.

Discourse relations in the PDTB are broadly categorized into two main types: **explicit** and **implicit**. Explicit discourse relations between arguments are explicitly signaled by the presence of a discourse connective, as shown in Example 6. In this example, the word

*since* acts as the discourse connective linking the two arguments of the relation. Conversely, implicit discourse relations occur when no connective is explicitly present in the text, but a relation can still be inferred. In such cases, annotators insert a suitable connective that could plausibly convey the intended relation, as illustrated in Example 7 - although there is no discourse connective, the word *because* could have been inserted between the two arguments to carry the same meaning. By convention, ARG1 is written in italics, ARG2 is written in bold and the discourse connective is underlined.

**Example 6.** *The striker played really well, since he was able to score a hattrick.*

**Example 7.** *The striker played really well. He was able to score a hattrick.*

While arguments of explicit relations are unconstrained in terms of their location, arguments of implicit relations must be adjacent in the text. In addition to the explicit and implicit discourse relations, there are also AltLex, AltLexC, EntRel, NoRel and hypophora discourse relations in the PDTB framework. However, these are significantly less frequent and out of the scope of this thesis. The explicit discourse relation in Example 6 is classified as intra-sentential (Intra-S), since both of its arguments appear within a single sentence. In contrast, the implicit discourse relation in Example 7 is considered inter-sentential (Inter-S) since its arguments span separate sentences.

In addition to the argument structure of discourse relations, the PDTB framework also provides sense labels for each discourse relation. These senses are organized hierarchically into three levels, as presented in Table 2.3. The top level has four major senses: TEMPORAL, CONTINGENCY, COMPARISON and EXPANSION. Each of these senses is further divided into more specific second-level senses, which in turn may be refined into third-level senses to capture the directionality of asymmetric relations. A discourse relation is considered symmetric if (ARG1, ARG2) and (ARG2, ARG1) are semantically equivalent. Otherwise, the relation is considered asymmetric.

### 2.2.1 PDTB 3.0

The PDTB 3.0 (Webber et al., 2019) is the third and most recent iteration of the discourse annotated PDTB corpus following its namesake framework. It consists of 2,161 Wall Street Journal articles and it contains a total of 53,558 discourse relations. Each successive iteration of the PDTB corpus have led to an increasing number of annotated discourse relations. The first version of the corpus, PDTB 1.0 (Prasad et al., 2006), contained 30,000 discourse relations, while the second version of the corpus, PDTB 2.0 (Prasad et al., 2007), had 40,600 discourse relations. In addition to the larger number of discourse

| 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-EXCEPTION          |
|             |                                | ARG2-AS-EXCEPTION          |
|             | 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-SUBSTITUTION       |
|             |                                | ARG2-AS-SUBSTITUTION       |

Table 2.3: Hierarchy of sense labels in the PDTB 3.0 framework. Table adapted from (Webber et al., 2019).

| Relation Type | Intra-Sentential | Inter-Sentential | Total          |
|---------------|------------------|------------------|----------------|
| Explicit      | 16,905 (69.3%)   | 7,330 (25.1%)    | 24,235 (45.3%) |
| Implicit      | 6,188 (25.4%)    | 15,544 (53.3%)   | 21,732 (40.6%) |
| AltLex        | 785 (3.2%)       | 712 (2.4%)       | 1,497 (2.8%)   |
| AltLexC       | 130 (0.5%)       | 5 (0.0%)         | 135 (0.3%)     |
| EntRel        | 358 (1.5%)       | 5,172 (17.7%)    | 5,530 (10.3%)  |
| Hypophora     | 4 (0.0%)         | 142 (0.5%)       | 146 (0.3%)     |
| NoRel         | -                | 283 (1.0%)       | 283 (0.5%)     |
| Total         | 24,370 (100%)    | 29,188 (100%)    | 53,558 (100%)  |

Table 2.4: Distribution of discourse relations in the PDTB 3.0 corpus. Table adapted from (Webber et al., 2019).

relations, each iteration of the corpus had the annotation guidelines revised to increase the quality of the discourse annotations. Among the various changes in the guidelines, the most significant updates in the PDTB 3.0 corpus include a revised sense hierarchy, the addition of new types of discourse relations and the removal of the attribution scheme.

Of the 53,558 annotated discourse relations in the PDTB 3.0, there is a balanced distribution between 24,370 Intra-S and 29,188 Inter-S relations. Table 2.4 presents the distribution of discourse relations in the corpus. While the vast majority of Intra-S relations are of the explicit type (69.3%), more than half of the Inter-S relations are of the implicit type (53.3%) - highlighting a significant contrast in how discourse relations manifest across sentence boundaries. There is also a contrasting difference on the total number of explicit and implicit discourse relations when compared against AltLex, AltLexC, EntRel, NoRel and hypophora relations. At the sense annotation level, it is also noteworthy that 1,075 (4.9%) of the implicit discourse relations were annotated with two senses as the annotators believed both interpretations could hold simultaneously.

Unfortunately, no information is available regarding inter-annotator agreement statistics for the PDTB 3.0 corpus. Therefore, the analysis presented here focuses on the statistics reported for the PDTB 2.0 corpus (Prasad et al., 2007). In the PDTB 2.0 corpus, discourse relations and their arguments were initially annotated by two annotators per relation and subsequently adjudicated by a team of four experts. Inter-annotator agreement was measured based on the extent of the argument spans using an exact match criterion. According to this metric, agreement between argument spans of explicit discourse relations reached 90.2%, while agreement for implicit discourse relations reached 85.1%.

Similar to the annotation of discourse relations, inter-annotator agreement statistics at

the sense level are only available for the PDTB 2.0 corpus (Prasad et al., 2007). The sense annotation was performed by two annotators and agreement was evaluated across all three levels of the PDTB sense hierarchy. Disagreement at level-1 was recorded only when annotators selected different level-1 sense labels - differences at lower levels within the same level-1 were not counted as disagreement at this level. Similarly, disagreement at level-2 was recorded only when annotators chose different level-2 labels without considering differences at level-3. However, any disagreement at a higher level of the hierarchy was propagated downward to the lower levels. Based on these metrics, the reported inter-annotator agreement was 94% at level-1, 84% at level-2, and 80% at level-3

### 2.2.2 DiscoGeM

The DiscoGeM corpus (Scholman et al., 2022a) contains 6,807 implicit discourse relations annotated following the PDTB framework and distributed across four textual genres: 2,800 relations from political texts, 3,060 from literary texts, 645 from Wikipedia texts and 302 relations extracted from the PDTB 3.0 corpus itself. The set of political discourse relations in DiscoGeM were extracted from the Europarl corpus (Koehn, 2005; Cartoni and Meyer, 2012) and the set of literary discourse relations were selected from 20 novels. These novels were originally written in English, German, French and Czech, with all non-English novels translated into English. For each of these languages, 5 novels were selected.

Contrary to previous annotation efforts of discourse corpora, such as with the RST-DT and the PDTB 3.0, the DiscoGeM corpus captures a distribution of sense labels per discourse relation instead of a single label. To achieve this, the authors collected annotations from at least 10 different crowdsourced annotators per relation using a proxy task. Annotators were presented with two arguments of a discourse relation and asked to select the discourse connective that best conveyed the meaning between them. These connectives were selected from a curated list, where each connective represents a sense in the PDTB framework (see Table 2.3) - the BELIEF and SPEECH-ACT level-2 senses, however, were excluded from the annotation process. Based on the selected connective, the authors inferred the sense of the relation. Through this crowdsourced method, instead of relying on a few expert annotators to find a single label for each relation, the authors of the corpus were able to collect multiple sense labels to capture the full range of possible interpretations of each relation.

In total, the authors collected 65,863 annotations from 199 participants for the total of 6,807 implicit discourse relations. To ensure the quality and consistency of the annotations, participants were required to be native English speakers residing in the UK or Ireland.

| Language<br>Original Lang. | Political |       |       |     | Literature |       |       |       |
|----------------------------|-----------|-------|-------|-----|------------|-------|-------|-------|
|                            | Eng       | Ger   | Fre   | Cze | Eng        | Ger   | Fre   | Cze   |
| English (Eng)              | 418       | 417   | 414   | -   | 800        | 787   | 758   | 777   |
| German (Ger)               | 701       | 701   | -     | -   | 800        | 683   | -     | -     |
| French (Fre)               | 739       | -     | 727   | -   | 780        | -     | 729   | -     |
| Czech (Cze)                | 700       | -     | -     | 697 | 680        | -     | -     | 526   |
| Total                      | 2,558     | 1,118 | 1,141 | 697 | 3,060      | 1,470 | 1,487 | 1,303 |

Table 2.5: Distribution of implicit discourse relations in the DiscoGeM 2.0 corpus according to textual genre and language. Table adapted from (Yung et al., 2024b).

In addition, they had to successfully complete a selection task before being admitted to the annotation process. Throughout the task, their responses were also subject to quality control checks to maintain annotation reliability.

### 2.2.3 DiscoGeM 2.0

Following the annotation procedure the first version of the corpus, DiscoGeM 2.0 (Yung et al., 2024b) extends the multi-label annotation of implicit discourse relations to the German, French and Czech languages. The second version of the corpus thus contains a total of 13,063 parallel implicit discourse relations across four languages: 5,847 in English, 2,588 in German, 2,628 in French and 2,000 in Czech. Unlike the first version, DiscoGeM 2.0 only contains relations from political texts and literary texts. The literary relations were selected from the same set of novels as in the original corpus. Relations that had already been annotated in English remained unchanged, while those originally translated into English were also annotated in their original language. Additionally, all relations from novels originally written in English were translated and annotated in the remaining languages. A similar procedure was applied to the political relations from the Europarl corpus, except that English relations were not translated into Czech. Table 2.5 shows the distribution of implicit discourse relations in DiscoGeM 2.0 according to textual genre and language.

Each relation in the corpus was annotated in its respective language by at least 10 different crowdsourced annotators. For each language, a curated list of discourse connectives was provided, with each connective representing a specific sense in the PDTB framework - except for the BELIEF and SPEECH-ACT level-2 senses, which were also not included in this version of the corpus. Annotators were asked to select the connective that best conveyed the meaning between the two arguments of the discourse relation, from which the

authors inferred the sense of the relation. Only native speakers of the target language were allowed to participate in the annotation process. As in the previous version of the corpus, annotators were required to pass a selection task before participating and their contributions were subject to quality control throughout the process.

## **2.3 Other Frameworks**

In addition to the RST and the PDTB, other frameworks have been proposed to model discourse structure, such as the Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003; Lascarides and Asher, 2007) and the more recent Enhanced Rhetorical Structure Theory (eRST) (Zeldes et al., 2025). While these frameworks are not used in this thesis, they offer valuable perspectives on discourse modeling and are briefly reviewed in this section.

### **Segmented Discourse Representation Theory**

The Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003; Lascarides and Asher, 2007) is a framework for modeling discourse that accounts for the complex ways in which larger text segments relate to one another. It represents discourse as a directed acyclic graph, where each node corresponds to a text segment and edges represent discourse relations. A central idea in the SDRT is that discourse structure is not strictly linear or hierarchical but graph-based, allowing segments to be connected in non-sequential and non-adjacent ways. This allows the SDRT to model discourse structure more flexibly than strictly hierarchical frameworks like the RST. However, the SDRT still imposes the right frontier constraint, which restricts the attachment of new segments to either the most recently introduced node or one of the nodes that dominate it in the discourse structure. This contrasts with the PDTB framework, where arguments of explicit discourse relations can be connected without constraints on their position in the text. Unlike many other frameworks, the SDRT also allows multiple discourse relations to hold simultaneously between the same text segments, reflecting the often ambiguous nature of discourse.

### **Enhanced Rhetorical Structure Theory**

The Enhanced Rhetorical Structure Theory (eRST) (Zeldes et al., 2025) is a recent discourse framework that builds upon the RST while integrating key insights from the PDTB and the SDRT. It preserves the main strengths of the RST, such as hierarchical nuclearity and full-text tree coverage, while overcoming some of its limitations. More precisely, the

eRST allows multiple relations between the same text segments as well as non-sequential discourse structures like the SDRT and it also distinguishes between implicit and explicit discourse relations like the PDTB. Like the RST, the eRST represents discourse using a primary tree structure, where nodes correspond to EDUs and edges represent discourse relations. Each non-terminal node includes at least one nucleus, maintaining the hierarchical structure central to the RST. However, eRST goes further by introducing secondary edges, which provide additional connections between nodes to represent overlapping or concurrent discourse relations. By combining the hierarchical structure of the RST with the representational flexibility of the SDRT and the distinguishable discourse relations of the PDTB, the eRST aims to provide a more unified framework for discourse.

This chapter provided an overview of the main discourse frameworks relevant to this thesis, the RST and the PDTB, and a concise introduction to the SDRT and the eRST. It also reviewed the discourse annotated corpora used in this work: the RST-DT, PDTB 3.0, DiscoGeM, and DiscoGeM 2.0 corpora. Together, these frameworks and corpora illustrate the diversity of theoretical and empirical approaches to computational discourse analysis and lay the foundation for the methodology presented in the following chapters.

## Chapter 3

# Mapping Explicit and Implicit Discourse Relations

As outlined in Chapter 2, different theoretical frameworks have been proposed to describe discourse structure. The two most widely used frameworks are the RST (Mann and Thompson, 1988) and the PDTB (Miltsakaki et al., 2004; Prasad et al., 2008). Despite their widespread use, these frameworks differ substantially in how they model discourse, which leads to challenges in the comparison, evaluation and interoperability of discourse models developed under each paradigm. To address these challenges, this chapter presents an empirical mapping between the RST and the PDTB frameworks, in line with the first research objective of this thesis introduced in Chapter 1: **to map explicit and implicit discourse relations between the RST and the PDTB frameworks**. The proposed mapping provides valuable insights into points of convergence and divergence between the two frameworks, laying the groundwork for more robust and generalizable discourse models. This chapter is based on the paper published at the Recent Advances in Natural Language Processing (RANLP) 2023 (Costa et al., 2023c).

The proposed mapping is established using the two most widely adopted corpora for each framework: the RST-DT (Carlson et al., 2001, 2003) and the PDTB 3.0 (Webber et al., 2019). Despite the differences in their annotation methodologies, both corpora have been partly annotated on overlapping Wall Street Journal articles, which enables a direct comparison of their annotations and the establishment of a mapping between their respective discourse frameworks. Given the importance of this task, previous research had already explored both theoretical (Bunt et al., 2012; Chiarcos, 2012, 2014; Chiarcos and Sukhareva, 2015; Bunt and Prasad, 2016; Bunt et al., 2020; Sanders et al., 2021) and empirical mappings (Rehbein et al., 2016; Scheffler and Stede, 2016; Polakova et al., 2017; Demberg

et al., 2019) between discourse frameworks. Notably, Demberg et al. (2019) proposed an empirical mapping between the RST-DT with the PDTB 2.0 corpus (Prasad et al., 2007). However, no previous work has attempted to establish a mapping involving the most recent PDTB 3.0 corpus, which introduced significant changes to the sense hierarchy and annotation guidelines.

### 3.1 Previous Work

Several studies have addressed the challenge of aligning different discourse frameworks, either by proposing theoretical correspondences between their underlying structures or by developing empirical methods for mapping annotated discourse relations across corpora. This section reviews previous work on this topic, distinguishing between theoretical and empirical approaches. Theoretical approaches aim to align the conceptual definitions and taxonomies of discourse relations across frameworks, while empirical approaches seek to align annotated instances of discourse relations in corpora that follow different frameworks. Previous work following both approaches highlights the conceptual motivations and practical methodologies that have shaped current efforts toward discourse framework interoperability.

A common method for establishing a theoretical mapping between discourse frameworks consists of introducing an intermediary representation to which all frameworks can be linked. This method allows for scalability, as new frameworks can be integrated by mapping them to the shared intermediary representation rather than to each individual framework. The following works exemplify this strategy by proposing intermediary representations to facilitate cross-framework alignment:

- Chiarcos (2012, 2014) and Chiarcos and Sukhareva (2015) present the Ontologies of Linguistic Annotation (OLiA), a modular framework designed to improve interoperability between different discourse frameworks by linking them through a shared intermediate representation known as the OLiA Reference Model. Unlike direct mapping approaches, OLiA allows for nuanced linking that can represent ambiguity, conceptual mismatches and alternative interpretations explicitly. This flexibility makes it particularly suited for aligning annotations in discourse corpora.
- Bunt et al. (2012, 2020) and Bunt and Prasad (2016) propose an open and extendable set of core relations, referred to as DR-Core, which are frequently found across existing discourse frameworks. Instead of defining a fixed inventory, they allow users to select or extend relation sets based on the needs of specific applications. These

relations are designed to serve as an interoperable intermediary layer that facilitates mapping between annotations grounded in different discourse frameworks.

- Sanders et al. (2021) introduce Unifying Dimensions (UniDim), a method to map discourse relations across frameworks by decomposing discourse relation labels into a set of linguistic motivated dimensions. Rather than relying on rigid one-to-one mappings or reducing discourse relations to a minimal common denominator, the authors introduce different dimensions composed of descriptive features that can flexibly represent the nuances of each discourse framework.

Despite the appeal of intermediary representations, theoretical mapping approaches face notable challenges. Differences in the granularity of discourse relation labels often lead to one-to-many or many-to-one mappings, limiting the precision of cross-framework alignment. Additionally, distinct motivations behind discourse frameworks result in divergent interpretations of discourse relations, further complicating the mapping process. These limitations motivate the need for empirical approaches that examine how discourse relations align in practice rather than in theory alone.

Empirical approaches attempt to establish mappings based on actual annotations discourse corpora. However, comparing these annotations is not straightforward. For instance, the RST and the PDTB frameworks differ not only in the sense labels used to define discourse relations but also in their segmentation and structural assumptions. For example, the RST framework segments text into EDUs and builds discourse trees spanning whole documents, while the PDTB framework segments text into intra-sentential or inter-sentential relations between arguments in adjacent sentences - which results in fewer annotated discourse relations. These differences pose challenges for direct comparison and motivate empirical studies that aim to reconcile discourse frameworks through aligned annotations. Below, we outline previous work following empirical approaches to map discourse frameworks:

- Rehbein et al. (2016) created an English corpus of spoken discourse relations annotated following the PDTB 3.0 and the Cognitive approach to Coherence Relations (CCR) (Sanders et al., 1992) frameworks. To ensure consistency across frameworks, they applied a shared segmentation procedure before annotating each relation according to both schemes. Their comparative analysis revealed substantial overlap between the PDTB 3.0 and CCR annotations, though differences in annotation guidelines and relation granularity led to some mismatches.

- Scheffler and Stede (2016) present a method for mapping annotations based on the RST and the PDTB 3.0 frameworks for explicitly discourse relations using the German Potsdam Commentary Corpus (PCC) (Stede and Neumann, 2014), which is independently annotated in both frameworks. The mapping highlights areas of compatibility and divergence, particularly in cases where PDTB annotations specify relations that are absent or simplified in the RST representation. However, their mapping does not consider implicit relations as they are not annotated in the PCC.
- Polakova et al. (2017) examine 472 implicit relations from the PDTB 2.0 corpus that aligned with matching relation labels and argument spans in the RST-DT annotation. Using the RST Signalling Corpus (Das and Taboada, 2018), they analyze what kinds of lexical or structural indicators might help signal these implicit relations. They conclude that implicit discourse relations cannot be easily mapped based on the signals annotated in the RST Signalling corpus.
- Demberg et al. (2019) introduce a method for automatically aligning discourse segments and map the corresponding discourse relation labels between the RST-DT and the PDTB 2.0 corpora. Their results show that while the two frameworks agree reasonably well on explicit relations, alignment is substantially weaker for implicit relations due to differences in segmentation strategies and annotation granularity. The work also provides an aligned corpus where both annotation layers can be queried simultaneously and highlights key sources of systematic discrepancies.

Empirical mapping approaches offer a data-driven alternative to theoretical alignment but face several challenges. Key difficulties arise from differences in segmentation guidelines and annotation scope across frameworks. For example, the RST annotates discourse relations for their nuclearity, whereas the PDTB adopts a flat structure without nuclearity distinctions. These structural mismatches complicate alignment and can lead to inconsistencies or ambiguous mappings. Despite previous efforts, no empirical approach has yet attempted to establish a mapping between the RST-DT and the PDTB 3.0 corpora. The following section introduces a mapping approach between these two corpora, focusing on both explicit and implicit discourse relations.

## 3.2 Methodology

The RST-DT and the PDTB 3.0 corpora were both annotated on top of Wall Street Journal articles (see Sections 2.1.1 and 2.2.1), with 385 articles co-annotated by both corpora.

|                     | RST-DT | PDTB 3.0 |
|---------------------|--------|----------|
| Segments            | 21,789 | 18,738   |
| Discourse Relations | 20,017 | 9,369    |

Table 3.1: Number of segments and discourse relations in the RST-DT and the PDTB 3.0 corpora in the set of 385 overlapping Wall Street Journal articles.

The mapping proposed in this section is based on these overlapping annotations. For consistency, the term **segments** will be used throughout this chapter to refer to both EDUs in the RST framework and arguments in the PDTB framework. Table 3.1 shows the number of segments and discourse relations annotated by each corpus in the set of 385 overlapping articles. Due to its flat annotation scheme, the PDTB 3.0 corpus contains significantly fewer discourse relations than the RST-DT, despite covering a comparable number of segments. Out of the 9,369 PDTB 3.0 relations, 4,169 (44.5%) are explicit and 3,965 (42.3%) are implicit discourse relations - which corresponds to a combined total of 8,134 (86.8%) discourse relations. The remaining 1,235 (13.2%) PDTB 3.0 relations consists of AltLex, AltLexC, EntRel, NoRel and hypophora discourse relations, which are not considered in the mapping approach described in this chapter.

Given the smaller number of discourse relations in the PDTB 3.0 compared to the RST-DT in Table 3.1, the PDTB 3.0 annotations are used as the starting point for the alignment and mapping of discourse relations. This strategy is consistent with the approach adopted by Demberg et al. (2019). The first step in establishing the mapping is to align the text segments annotated under the two frameworks. The next section describes the proposed alignment procedure of text segment, while the subsequent section details the describes the mapping of discourse relations for the aligned segments.

### 3.2.1 Segment Alignment

The goal of the alignment is to match PDTB 3.0 segments to their closest RST-DT segment. RST-DT segments can either consist of a single EDU or multiple contiguous EDUs, whereas PDTB 3.0 segments can either be continuous or discontinuous, as shown in Example 8. In this example, the PDTB 3.0 segment ARG2 is discontinuous since it is split into ARG2a and ARG2b. For easier interpretation, segments extracted from the RST-DT and the PDTB 3.0 will both be delimited by square brackets in the examples presented throughout this chapter. The superscript of RST-DT segments will indicate its EDU number in the corpus and the superscript of PDTB 3.0 segments will indicate the argument number within the relation. The examples provided in this section are drawn from the overlapping

Wall Street Journal articles annotated by both corpora.

**Example 8.** Mr. Carpenter notes [that these types of investors]<sup>ARG2a</sup> also [are sophisticated enough not to complain about Kidder’s aggressive use of program trading]<sup>ARG2b</sup>.

Depending on whether a PDTB 3.0 segment is continuous or discontinuous, there are two possible alignment methods.

### Continuous PDTB 3.0 Segments

For each continuous PDTB 3.0 segment, the alignment methodology begins by identifying all RST-DT segments that overlap with it. The identified segments are then ranked using an overlap metric, which maximizes the number of shared characters between the segments and penalizes extra characters present in the RST-DT segment. Based on this ranking, the highest scoring RST-DT segment is selected as the initial alignment match for the PDTB 3.0 segment. Following this overlap metric, successful alignments are classified as either perfect or imperfect.

The alignment is considered **perfect** when all of the characters of the PDTB 3.0 segment match exactly with those of the corresponding RST-DT segment. However, if there are extra characters in the the PDTB 3.0 segment and these consist of punctuation and explicit connectives, both segments are still considered to be perfectly aligned since PDTB 3.0 segments systematically exclude terminal punctuation and explicit connectives contrary to RST-DT segments. In Example 9, ARG1 of the PDTB 3.0 relation is perfectly aligned with EDU67, since only punctuation differs, and ARG2 is perfectly aligned with the contiguous RST-DT segment EDU68-69.

**Example 9. PDTB:** [We’ve had a good relationship with GE]<sup>ARG1</sup>, [which is the first time you could say that]<sup>ARG2</sup>.

**RST-DT:** [We’ve had a good relationship with GE,]<sup>67</sup> [which is the first time]<sup>68</sup> [you could say that.]<sup>69</sup>

Conversely, the alignment is considered **imperfect** if the RST-DT segment with the highest overlap with the PDTB 3.0 has either additional or missing characters. In Example 10, ARG1 is imperfectly aligned with EDU92 because the PDTB 3.0 segment contains the additional text *of the opportunity*, which is not present in the RST-DT segment. Nevertheless, among all available RST-DT segments, EDU92 remains the one with the highest overlap and is therefore selected as the alignment match.

| Type      | ARG : EDU | Count         | Total          |
|-----------|-----------|---------------|----------------|
| Perfect   | 1 : 1     | 7,621 (49.8%) | 12,959 (84.8%) |
|           | 1 : n     | 5,338 (34.9%) |                |
| Imperfect | 1 : 1     | 1,705 (11.2%) | 2,329 (15.2%)  |
|           | 1 : n     | 624 (4.1%)    |                |
| Total     |           | 15,288 (100%) | 15,288 (100%)  |

Table 3.2: Distribution of perfect and imperfect alignments between continuous PDTB 3.0 segments and RST-DT segments.

**Example 10. PDTB:** ... [of the opportunity to rebuild a franchise at Kidder]<sup>ARG1</sup>.  
**RST-DT:** ... of the opportunity [to rebuild a franchise at Kidder.]<sup>92</sup>

Table 3.2 summarizes the distribution of perfect and imperfect alignments after aligning continuous PDTB 3.0 segments onto RST-DT segments. As the table shows, 12,959 (84.8%) of the alignments found for all of the continuous PDTB 3.0 segments were considered to be perfect. Among these, 7,621 (49.8%) PDTB 3.0 segments were aligned to an RST-DT segment consisting of a single EDU, while 5,338 (34.9%) were aligned to an RST-DT segment consisting of multiple contiguous EDUs. The same pattern is observed for imperfect alignments, where the majority of the PDTB 3.0 segments were also aligned with an RST-DT segment consisting of a single EDU.

### Discontinuous PDTB 3.0 Segments

For discontinuous PDTB 3.0 segments, each of the split segments is aligned to an RST-DT segment using the same method as for continuous segments. For instance, in Example 11, the two constituents of the PDTB 3.0 segments, ARG2a and ARG2b, are respectively aligned with the RST-DT segments EDU110 and EDU110-111. Both alignments in this example are considered imperfect as the RST-DT segments do not match exactly the characters of either constituent of the PDTB 3.0 segment.

**Example 11. PDTB:** Mr. Carpenter notes [that these types of investors]<sup>ARG2a</sup> also [are sophisticated enough not to complain about Kidder’s aggressive use of program trading]<sup>ARG2b</sup>.  
**RST-DT:** Mr. Carpenter notes [that these types of investors also are sophisticated enough]<sup>110</sup> [not to complain about Kidder’s aggressive use of program trading.]<sup>111</sup>

| Type      | ARG : EDU | Count        | Total         |
|-----------|-----------|--------------|---------------|
| Perfect   | 1 : 1     | 762 (38.3%)  | 936 (47.1%)   |
|           | 1 : n     | 174 (8.7%)   |               |
| Imperfect | 1 : 1     | 818 (41.1%)  | 1,053 (52.9%) |
|           | 1 : n     | 235 (11.8%)  |               |
| Total     |           | 1,989 (100%) | 1,989 (100%)  |

Table 3.3: Distribution of perfect and imperfect alignments between discontinuous PDTB 3.0 segments and RST-DT segments.

Table 3.3 presents the distribution of perfect and imperfect alignments after aligning discontinuous PDTB 3.0 segments onto RST-DT segments. In contrast to the continuous segments, where 84.8% were perfectly aligned (see Table 3.2), only 47.1% of the discontinuous segments were perfectly aligned. Thus, slightly more than half of the discontinuous PDTB 3.0 segments were imperfectly aligned with RST-DT segments. Nevertheless, in both continuous and discontinuous cases, the majority of the PDTB 3.0 segments were aligned with an RST-DT segment consisting of a single EDU.

The 15,288 aligned continuous PDTB 3.0 segments from Table 3.2 combined with the 1,989 aligned discontinuous PDTB 3.0 segments from Table 3.3 result in a total of 17,277 aligned PDTB 3.0 segments out of the 18,738 segments present in the 385 overlapping Wall Street Journal articles (see Table 3.1). This corresponds to an overall alignment of 92.2% of the PDTB 3.0 segments onto an RST-DT segment, reflecting the effectiveness of the proposed alignment methods. The resulting set of aligned segments serves as the basis for mapping discourse relations across the two corpora.

### 3.2.2 Relation Mapping

Given the hierarchical nature of the RST framework, some PDTB 3.0 discourse relations can only be mapped to complex RST relations embedded within nested structures involving contiguous EDUs. For example, in Figure 3.1, the PDTB 3.0 discourse relation COMPARISON.CONCESSION between segments ARG1 and ARG2 can only be mapped onto the RST-DT discourse relation CONCESSION nested at the third level of the discourse tree between segments EDU1 and EDU2–4. In this example, the segments of both discourse relations are perfectly aligned, which makes the mapping straightforward. However, such alignment is not always guaranteed, which introduces additional challenges in the mapping of discourse relations. All nucleus segments of RST discourse relations are

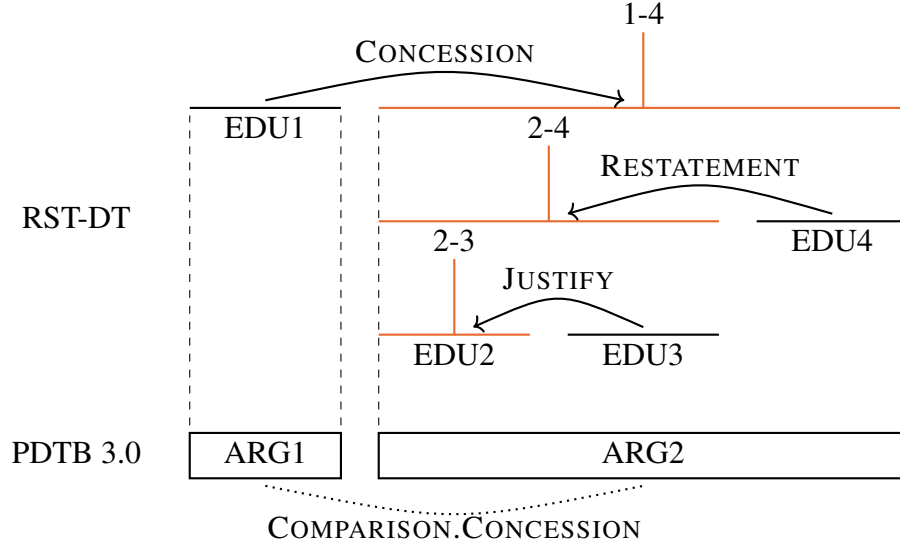


Figure 3.1: Example of a mapping between a PDTB 3.0 discourse relation and an RST-DT discourse relation nested at the third level of the RST tree.

highlighted in orange throughout this chapter to support visual identification and to help following the nucleus path.

This section describes the mapping approach of discourse relations considering different scenarios. In some of these scenarios, the notion of the strong nuclearity principle (Marcu, 2000) and of the nucleus path (Demberg et al., 2019) in the RST framework will be important to guarantee the most appropriate mapping of discourse relations. The **strong nuclearity principle** states that if a discourse relation holds between RST-DT segments composed of multiple contiguous EDUs, the same relation must also hold between the nuclei of these segments. The **nucleus path** refers to the path between the nucleus of a discourse relation composed of multiple contiguous EDUs and the single nuclear EDU at the bottom of its RST discourse tree by following only the segments annotated as nucleus. As an example, consider the RST tree in Figure 3.2.

In this example, the nucleus of the RST-DT discourse relation PREPARATION is the contiguous segment EDU2-5, which in turn was established from the BACKGROUND relation between the contiguous satellite segment EDU2-3 and the contiguous nuclear segment EDU4-5. The same principle is then applied to the discourse relation that led to EDU4-5, which leads to the single nuclear EDU5. Thus, the nucleus path of the PREPARATION discourse relation leads to EDU5, which is considered the single nuclear EDU at the bottom of this discourse tree. Applying the same principle to the previous example in Figure 3.1, the nucleus path of the CONCESSION discourse relation leads to EDU2.

Building on the concepts of strong nuclearity and nucleus path, six different scenarios

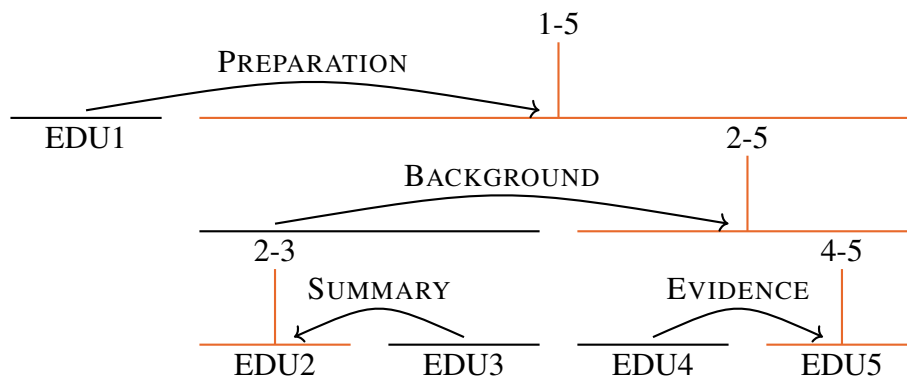


Figure 3.2: RST tree exemplifying the nucleus path.

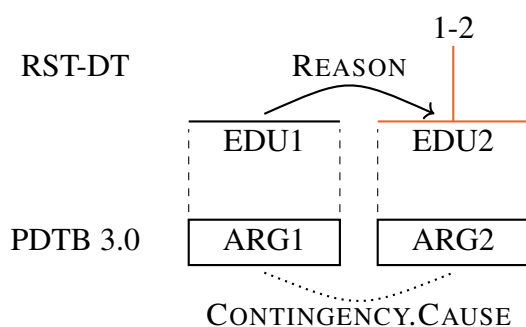


Figure 3.3: Examples of a perfect relation mapping.

were identified in the process of mapping discourse relations across the RST-DT and the PDTB 3.0 corpora. The six different scenarios are described below.

### Perfect Mapping

If both segments of a PDTB 3.0 discourse relation are continuous and perfectly aligned with two RST-DT segments of a discourse relation, the PDTB 3.0 sense can be directly mapped onto the corresponding RST-DT discourse relation. Figures 3.3 and 3.4 show two examples of a perfect mapping. In Figure 3.3, ARG1 and ARG2 are perfectly aligned with EDU1 and EDU2, which allows a direct mapping of the PDTB 3.0 sense CONTINGENCY.CAUSE onto the RST-DT discourse relation REASON. In Figure 3.4, the PDTB 3.0 segments ARG1 and ARG2 are perfectly aligned with the contiguous RST-DT segments EDU1-2 and EDU3-4. Even though the RST-DT segments are composed of contiguous EDUs, they are perfectly aligned with the PDTB 3.0 segments, which allows the PDTB 3.0 sense TEMPORAL.SYNCHRONOUS to be directly mapped onto the RST discourse relation TEMPORAL-SAME-TIME. In this case, the internal discourse relations between the individual EDUs that originated EDU1-2 and EDU3-4 are not considered in the mapping.

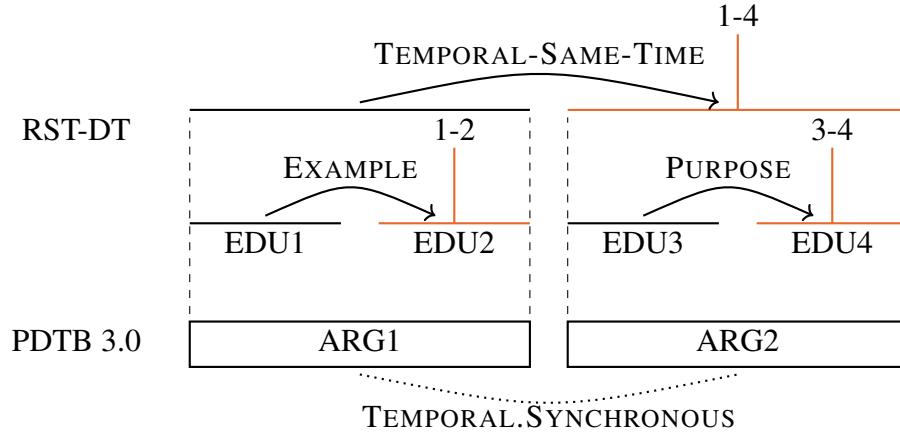


Figure 3.4: Example of a perfect relation mapping.

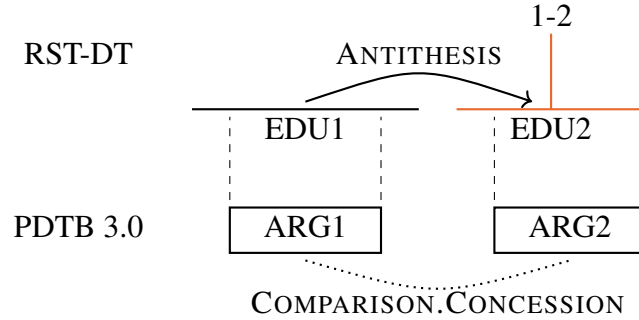


Figure 3.5: Example of an imperfect relation mapping.

### Imperfect Mapping

If one or both PDTB 3.0 segments of a discourse relation are not perfectly aligned with the corresponding RST-DT segments, a mapping can still be established under certain conditions. Figures 3.5 and 3.6 illustrate two such imperfect mapping scenarios. In Figure 3.5, the PDTB 3.0 segments are imperfectly aligned with their RST-DT counterparts, but both of the RST-DT segments consist of a single EDU. In this situation, even though the mapping is not considered perfect, it is still possible to map the PDTB 3.0 sense `COMPARISON.CONCESSION` onto the RST-DT discourse relation `ANTITHESIS`.

In Figure 3.6, however, the alignment of segments is slightly more complex. While the PDTB 3.0 segment `ARG2` is perfectly aligned with the RST-DT segment `EDU3`, the PDTB 3.0 segment `ARG1` is imperfectly aligned with the contiguous RST-DT segments `EDU1-2`. To determine whether the PDTB 3.0 sense `TEMPORAL.ASYNCHRONOUS` can be mapped onto the second level RST-DT discourse relation `TEMPORAL-AFTER`, it is necessary to trace the nucleus paths of both `EDU1-2` and `EDU3-4`. Since `ARG1` is not perfectly aligned with the full `EDU1-2` segment, the mapping is valid only if the nucleus of `EDU1-2`

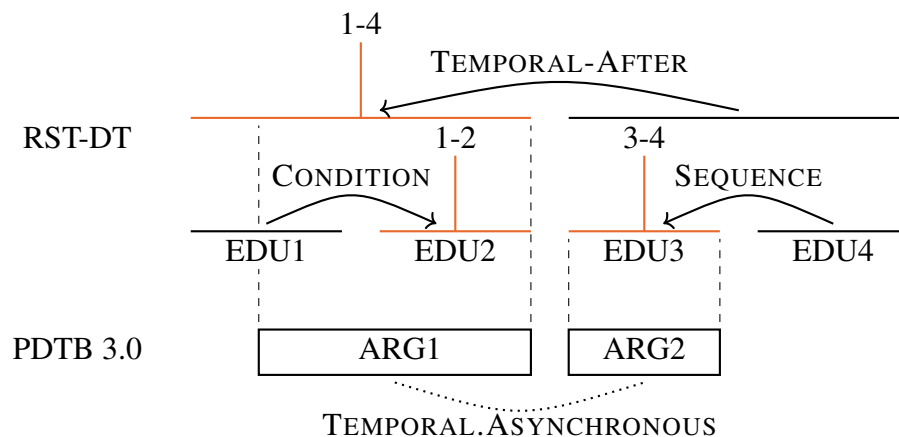


Figure 3.6: Example of an imperfect relation mapping.

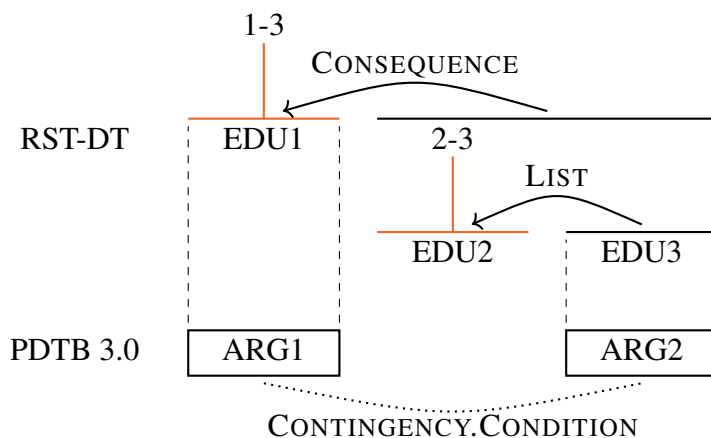


Figure 3.7: Example of an impossible mapping due to unclear nucleus path.

overlaps with ARG1. Similarly, although ARG2 is perfectly aligned with EDU3, it is not perfectly aligned with the full EDU3-4 segment which is the satellite in the RST relation. Thus, for the mapping to hold, EDU3 must be the nucleus of EDU3-4. In this example, the nucleus path of EDU1-2 leads to EDU2, which overlaps with ARG1, and the nucleus of EDU3-4 is EDU3, which is perfectly aligned with ARG2. Since the nuclei of EDU1-2 and EDU3-4 overlap with ARG1 and ARG2, the strong nuclearity principle dictates that the RST-DT discourse relation **TEMPORAL-AFTER** between EDU1-2 and EDU3-4 can be mapped onto the PDTB 3.0 sense **TEMPORAL.ASYNCHRONOUS**.

### Unclear Nucleus Path

If one or both PDTB 3.0 segments of a discourse relation are aligned to RST-DT segments that are only connected through a nested discourse relation, but the nuclei of those

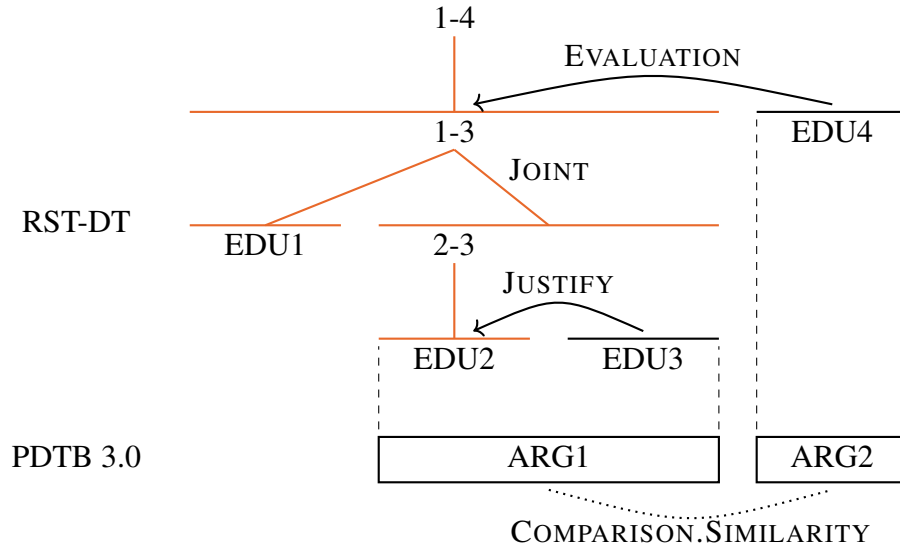


Figure 3.8: Example of an impossible mapping due to multinuclear relation.

RST-DT segments does not overlap with the corresponding PDTB 3.0 segments, a mapping becomes impossible since the strong nuclearity principle does not hold. Consider the example in Figure 3.7. Here, ARG1 is aligned with EDU1 and ARG2 is aligned with EDU3. However, the first RST-DT discourse relation between EDU1 and EDU3 occurs only at second level between EDU1 and EDU2-3. In this example, since the nucleus of EDU2-3 is EDU2, which does not overlap with ARG2, the RST-DT relation CONSEQUENCE cannot be reliably mapped onto the PDTB 3.0 sense CONTINGENCY.CONDITION.

### Multinuclear Relation

If one or both PDTB 3.0 segments are aligned to contiguous RST-DT segments that are only connected through a nested relation whose nucleus path leads to a multinuclear discourse relation, a mapping also becomes impossible. In such cases, the strong nuclearity principle cannot be applied because no unique nucleus can be identified for the contiguous segments involved. For example, in Figure 3.8, the RST-DT discourse relation JOINT connects EDU1 and EDU2-3 in a multinuclear configuration, making it unclear which of the two segments serves as the nucleus of EDU1-4. If EDU2-3 were identifiable as the nucleus, a mapping could be established between the RST-DT discourse relation EVALUATION and the PDTB 3.0 sense COMPARISON.SIMILARITY, since EDU2-3 is aligned with ARG1. However, due to the multinuclear nature of the relation, such a mapping cannot be reliably determined.

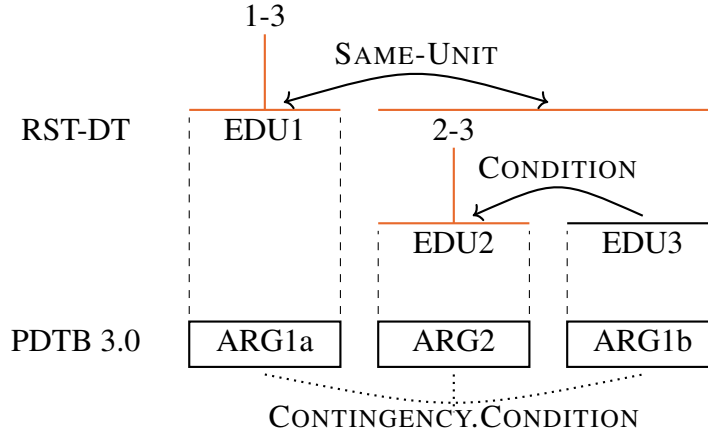


Figure 3.9: Example of an imperfect mapping due to discontinuous PDTB 3.0 segment.

### Discontinuous Relation

In cases where one of the PDTB 3.0 discourse segments is discontinuous and the other segment appears between the discontinuous segments, a mapping is possible if the RST-DT segments aligned with the discontinuous PDTB 3.0 segment are connected by a SAME-UNIT discourse relation. If this condition is met, and one of these segments is also involved in a discourse relation with the segment aligned to the non-discontinuous PDTB 3.0 argument, then the RST-DT relation can be mapped onto the PDTB 3.0 sense. Such mappings are considered imperfect due to the discontinuity of the PDTB 3.0 segment. An example of a possible imperfect mapping with a discontinuous PDTB 3.0 segment is shown in Figure 3.9. In this example, the RST discourse relation CONDITION can be mapped onto the PDTB 3.0 sense CONTINGENCY.CONDITION, since EDU2-3 is connected to EDU1 by a SAME-UNIT relation and the nucleus of EDU2-3 overlaps with ARG2. If this condition was not met, a mapping would be impossible.

### Embedded Relation

Finally, if both segments of a PDTB 3.0 discourse relation are embedded with an RST-DT segment consisting of a single EDU, a mapping is impossible. Consider the example in Figure 3.10. In this example, the PDTB 3.0 EXPANSION.EQUIVALENCE cannot be mapped onto an RST-DT discourse relation since both PDTB 3.0 segments are contained in an RST-DT segment consisting of a single EDU. This scenario occurs due to a difference in segmentation granularity between both discourse frameworks.

The six scenarios described above illustrate how the proposed mapping approach systematically accounts for various alignment configurations and maps discourse relations

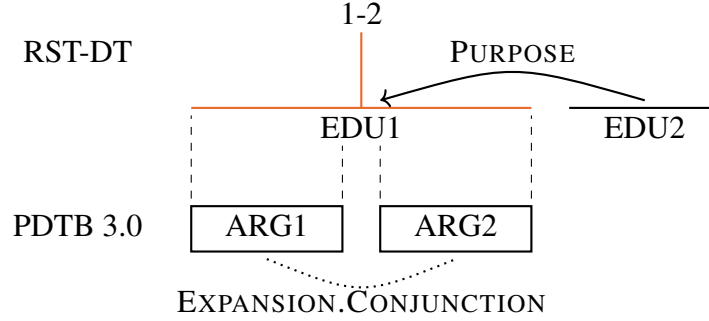


Figure 3.10: Example of an impossible mapping due to embedded relation.

| Relation Mapping | Discourse Relation | Scenario               | Count         | Sub-Total     | Total         |
|------------------|--------------------|------------------------|---------------|---------------|---------------|
| Possible         | Explicit           | Perfect Mapping        | 2,237 (27.5%) | 3,073 (37.8%) | 6,510 (80.0%) |
|                  |                    | Imperfect Mapping      | 836 (10.3%)   |               |               |
|                  | Implicit           | Perfect Mapping        | 3,062 (37.6%) | 3,437 (42.3%) |               |
|                  |                    | Imperfect Mapping      | 375 (4.6%)    |               |               |
| Impossible       | Explicit           | Embedded Relation      | 106 (1.3%)    | 1,096 (13.5%) | 1,624 (20.0%) |
|                  |                    | Unclear Nucleus Path   | 64 (0.8%)     |               |               |
|                  |                    | Multinuclear Relation  | 197 (2.4%)    |               |               |
|                  |                    | Discontinuous Relation | 729 (9.0%)    |               |               |
|                  | Implicit           | Embedded Relation      | 50 (0.6%)     | 528 (6.5%)    |               |
|                  |                    | Unclear Nucleus Path   | 81 (1.0%)     |               |               |
|                  |                    | Multinuclear Relation  | 183 (2.2%)    |               |               |
|                  |                    | Discontinuous Relation | 214 (2.6%)    |               |               |
| Total            |                    |                        | 8,134 (100%)  | 8,134 (100%)  | 8,134 (100%)  |

Table 3.4: Quantitative mapping results of explicit and implicit discourse relations.

across the RST-DT and the PDTB 3.0 corpora. The mapping approach proposed in this chapter are formally expressed in Algorithm 1.

### 3.3 Results

Following the proposed mapping algorithm, Algorithm 1, it was possible to map the explicit and implicit discourse relations between the RST-DT and the PDTB 3.0 corpora. Table 3.4 presents the quantitative results of the mapping algorithm for explicit and implicit discourse relations based on the six mapping scenarios described in the Section 3.2.2.

The proposed mapping algorithm was able to map 6,510 (80.0%) of the 8,134 explicit and implicit discourse relations in the 385 overlapping Wall Street Journal articles annotated in the RST-DT and the PDTB 3.0 corpora. More precisely, the mapping algorithm

---

**Algorithm 1** Mapping algorithm for explicit and implicit discourse relations

---

```
for all continuous PDTB 3.0 segments do
  Identify overlapping RST-DT segments
  Compute overlap score: overlapping characters - extra RST-DT characters
  Select RST-DT segment with the highest overlap score
  if segments match exactly (ignoring punctuation/connectives) then
    Mark alignment as Perfect
  else
    Mark alignment as Imperfect
  end if
end for
for all discontinuous PDTB 3.0 segments do
  Align each constituent individually using the same procedure
end for

for all continuous PDTB 3.0 relations with aligned segments do
  if both segments are perfectly aligned with segments of an RST-DT relation then
    Perfect Mapping
    Map PDTB 3.0 sense to RST-DT relation
  else if imperfectly aligned but overlap RST-DT nuclei via nucleus path then
    Imperfect Mapping
    Map PDTB 3.0 sense to RST-DT relation
  else if do not overlap RST-DT nuclei via nucleus path then
    Mapping Impossible: Unclear Nucleus Path
  else if nucleus path reaches a multinuclear relation then
    Mapping Impossible: Multinuclear
  else if both PDTB 3.0 segments are embedded in a single RST-DT EDU then
    Mapping Impossible: Embedded
  end if
end for

for all discontinuous PDTB 3.0 relations with aligned segments do
  if discontinuous segments are aligned with segments of an RST-DT SAME-UNIT
  relation then
    Imperfect Mapping
    Map PDTB 3.0 sense to RST-DT relation
  else
    Mapping Impossible: Discontinuous Relation
  end if
end for
```

---

was able to map 3,073 (73.7%) of the 4,169 explicit relations and 3,437 (86.7%) of the 3,965 implicit relations. The higher percentage of successfully mapped implicit discourse relations in comparison to explicit discourse relations can be attributed to the different segmentation guidelines in the PDTB 3.0 framework for the arguments of these relations. In the PDTB 3.0 framework, explicit discourse relations are annotated around discourse connectives, which allows for greater flexibility in the location and extent of their arguments. In contrast, implicit discourse relations lack an explicit connective and therefore require their arguments to be adjacent to form a relation. This segmentation constraint closely reflects the RST segmentation guidelines where texts are segmented into adjacent EDUs, thus making it easier to align arguments of implicit relations.

The same reasoning explains the contrast between perfect and imperfect mappings of explicit and implicit discourse relations. Out of the 3,073 successfully mapped explicit discourse relations, 2,237 (72.8%) were perfectly mapped while 836 (27.2%) imperfectly mapped. Meanwhile, out of the 3,437 successfully mapped implicit discourse relation, 3,062 (89.1%) were perfectly mapped while only 375 (10.9%) were imperfectly mapped. Finally, it also worth to note that 729 (66.5%) of the 1,096 explicit discourse relations that were impossible to map was due to discontinuous relations - which is also related to the more permissive segmentation of explicit arguments in the PDTB 3.0 framework.

Having successfully mapped 6,510 explicit and implicit discourse relations across both frameworks, the next two sections explore how PDTB 3.0 sense labels of explicit and implicit discourse relations are mapped onto RST discourse relations.

### 3.3.1 Mapping of Explicit Discourse Relations

Table 3.5 presents the mapping of PDTB 3.0 sense labels of the explicit discourse relations onto RST discourse relations. To keep the table readable, only discourse relations for which at least one of their mappings had at least 30 instances are presented. Percentages and color gradients are calculated row-wise. From the perspective of RST discourse relations, the majority tend to map predominantly onto a single PDTB 3.0 sense. For instance, the RST discourse relation LIST is mapped 95% of the time onto the PDTB 3.0 sense EXPANSION.CONJUNCTION. Similarly, the RST discourse relations TEMPORAL-AFTER and TEMPORAL-SAME-TIME are mapped onto the PDTB 3.0 sense TEMPORAL.ASYNCHRONOUS and TEMPORAL.SYNCHRONOUS, respectively, 94% and 98% of the time. Conversely, the mapping onto RST discourse relations from the PDTB 3.0 perspective is less clear. With the exception of the CONTINGENCY.CONDITION sense which is mapped 127

| PDTB<br>RST-DT     | COMPARISON  |            | CONTINGENCY |             | EXPANSION   | TEMPORAL   |             | Total      |
|--------------------|-------------|------------|-------------|-------------|-------------|------------|-------------|------------|
|                    | CONCESSION  | CONTRAST   | CAUSE       | CONDITION   | CONJUNCTION | ASYNCH     | SYNCH       |            |
| CONTRAST           | 60.5% (138) | 25.9% (59) | 0% (0)      | 0.4% (1)    | 9.2% (21)   | 0% (0)     | 3.9% (9)    | 100% (228) |
| LIST               | 2.4% (17)   | 0.3% (2)   | 0.1% (1)    | 0% (0)      | 95.0% (664) | 0.3% (2)   | 1.9% (13)   | 100% (699) |
| SEQUENCE           | 2.3% (2)    | 0% (0)     | 0% (0)      | 0% (0)      | 72.1% (62)  | 23.3% (20) | 2.3% (2)    | 100% (86)  |
| ANTITHESIS         | 83.8% (207) | 7.3% (18)  | 0% (0)      | 0.4% (1)    | 2.8% (7)    | 1.2% (3)   | 4.5% (11)   | 100% (247) |
| CIRCUMSTANCE       | 7.3% (20)   | 0.4% (1)   | 8.0% (22)   | 6.6% (18)   | 5.5% (15)   | 31.4% (86) | 40.9% (112) | 100% (274) |
| CONCESSION         | 87.6% (170) | 5.7% (11)  | 0% (0)      | 0% (0)      | 2.1% (4)    | 1.5% (3)   | 3.1% (6)    | 100% (194) |
| CONDITION          | 2.6% (4)    | 1.3% (2)   | 0% (0)      | 84.1% (127) | 0% (0)      | 8.6% (13)  | 3.3% (5)    | 100% (151) |
| ELAB-ADDITIONAL    | 29.8% (54)  | 5.0% (9)   | 2.2% (4)    | 0.6% (1)    | 55.8% (101) | 3.9% (7)   | 2.8% (5)    | 100% (181) |
| EXPLAN-ARGUMENT    | 19.0% (11)  | 0% (0)     | 65.5% (38)  | 0% (0)      | 0% (0)      | 1.7% (1)   | 13.8% (8)   | 100% (58)  |
| REASON             | 0% (0)      | 1.3% (1)   | 71.1% (54)  | 0% (0)      | 7.9% (6)    | 6.6% (5)   | 0% (0)      | 100% (76)  |
| TEMPORAL-AFTER     | 1.9% (1)    | 0% (0)     | 0% (0)      | 0% (0)      | 3.8% (2)    | 94.3% (50) | 0% (0)      | 100% (53)  |
| TEMPORAL-SAME-TIME | 0% (0)      | 0% (0)     | 2.2% (1)    | 0% (0)      | 0% (0)      | 0% (0)     | 97.8% (44)  | 100% (45)  |
| Total              | (624)       | (103)      | (130)       | (148)       | (882)       | (190)      | (215)       | (2,292)    |

Table 3.5: Mapping results for the explicit discourse relations. The table shows only discourse relations for which at least one of their mappings had at least 30 instances, which results in a total of 2,292 relations presented instead of 3,073. Percentages and color gradients are calculated row-wise.

(85.8%) out of 148 times onto the RST discourse relation `CONDITION` and the `EXPANSION.CONJUNCTION` sense which is mapped 664 (75.3%) out of 882 times onto the RST discourse relation `LIST`, most of the PDTB 3.0 senses are not predominantly mapped onto a single RST discourse relation. This asymmetry suggests that the definitions of RST discourse relation are more fine-grained than the definitions of PDTB 3.0 sense.

### 3.3.2 Mapping of Implicit Discourse Relations

Table 3.6 presents the mapping of PDTB 3.0 sense labels of the implicit discourse relations onto RST discourse relations. Similar to Table 3.5, only discourse relations for which at least one of their mappings had at least 30 instances are presented to keep the table readable. Percentages and color gradients are calculated row-wise. Contrary to the mapping results of the explicit discourse relations, the mapping of implicit discourse relations from both the RST and the PDTB 3.0 perspective is less clear. With the exception of the RST discourse relation `PURPOSE` which is mapped 96% of the time onto the PDTB 3.0 sense `CONTINGENCY.PURPOSE` and vice-versa, there are no clear mappings between both frameworks. The less clear mapping can be attributed to the inherently higher subjectivity in the interpretation of implicit discourse relations in the absence of explicit connectives, which in consequence leads to more ambiguous annotations. Similar results were also observed by Demberg et al. (2019) when mapping implicit discourse relations from the PDTB 2.0 onto the RST-DT corpora.

| RST-DT \ PDTB     | COMPARISON | CONTINGENCY |             | EXPANSION   |            |             | TEMPORAL   | Total      |
|-------------------|------------|-------------|-------------|-------------|------------|-------------|------------|------------|
|                   | CONCESSION | CAUSE       | PURPOSE     | CONJUNCTION | INSTANT    | LEV-DETAIL  | ASYNCH     |            |
| LIST              | 3.8% (18)  | 19.2% (92)  | 0.2% (1)    | 63.0% (302) | 1.9% (9)   | 0.2% (1)    | 9.4% (45)  | 100% (479) |
| SEQUENCE          | 8.2% (6)   | 6.8% (5)    | 0% (0)      | 12.3% (9)   | 0% (0)     | 5.5% (4)    | 67.1% (49) | 100% (73)  |
| CONSEQUENCE       | 7.4% (6)   | 50.6% (41)  | 4.9% (4)    | 18.5% (15)  | 3.7% (3)   | 4.9% (4)    | 9.9% (8)   | 100% (81)  |
| ELAB-ADDITIONAL   | 8.8% (77)  | 26.8% (236) | 0.5% (4)    | 35.4% (311) | 4.6% (40)  | 19.2% (169) | 4.8% (42)  | 100% (879) |
| ELAB-GEN-SPECIFIC | 1.0% (1)   | 15.5% (15)  | 0% (0)      | 13.4% (13)  | 17.5% (17) | 51.5% (50)  | 1.0% (1)   | 100% (97)  |
| EVIDENCE          | 2.3% (2)   | 13.6% (12)  | 0% (0)      | 12.5% (11)  | 39.8% (35) | 30.7% (27)  | 1.1% (1)   | 100% (88)  |
| EXAMPLE           | 0% (0)     | 12.0% (13)  | 0% (0)      | 8.3% (9)    | 63.0% (68) | 15.7% (17)  | 0.9% (1)   | 100% (108) |
| EXPLAN-ARGUMENT   | 5.7% (14)  | 53.4% (132) | 0% (0)      | 7.3% (18)   | 12.6% (31) | 20.2% (50)  | 0.8% (2)   | 100% (247) |
| PURPOSE           | 0% (0)     | 3.5% (8)    | 96.1% (222) | 0.4% (1)    | 0% (0)     | 0% (0)      | 0% (0)     | 100% (231) |
| REASON            | 0% (0)     | 72.9% (35)  | 12.5% (6)   | 6.3% (3)    | 0% (0)     | 6.3% (3)    | 2.1% (1)   | 100% (48)  |
| Total             | (124)      | (589)       | (237)       | (692)       | (203)      | (336)       | (150)      | (2,331)    |

Table 3.6: Mapping results for the implicit discourse relations. The table shows only discourse relations for which at least one of their mappings had at least 30 instances, which results in a total of 2,331 relations presented instead of 3,437. Percentages and color gradients are calculated row-wise.

### 3.4 Discussion

The results presented Section 3.3 reflect the annotations guidelines of the RST and the PDTB 3.0 frameworks. As shown in Table 3.4, implicit discourse relations in the PDTB 3.0 were more easily aligned with discourse relations in the RST-DT than explicit relations. This outcome can be attributed to fundamental differences in how each framework handles segmentation. In the RST framework, texts are fully segmented into adjacent EDUs. In contrast, in the PDTB 3.0 framework, segmentation depends on the presence of discourse connectives. When an explicit connective is present, the corresponding arguments are minimally specified and are more flexibly annotated within the text. However, in the absence of a connective, the PDTB 3.0 framework requires the two argument of an implicit discourse relation to be adjacent, which results in segmentation boundaries that more closely resemble the EDU segmentation in the RST framework. This segmentation similarity helps explain the higher proportion of successfully aligned implicit discourse relations in Table 3.4 compared to explicit discourse relations.

However, once the arguments were successfully aligned, the results in Table 3.5 show that a majority of explicit discourse relations exhibited clear one-to-one mappings between the RST-DT and the PDTB 3.0 corpora. These mappings were more consistent when approached from the perspective of the RST framework, suggesting that its taxonomy of discourse relations is more fine-grained than the PDTB 3.0 sense definitions. In contrast, Table 3.6 shows that, with the exception of the strong mapping between the RST discourse

relation PURPOSE and the PDTB 3.0 sense CONTINGENCY.PURPOSE, most implicit discourse relations did not find such direct mappings. This can be attributed to the inherently more subjective nature of implicit discourse relations. In the absence of an explicit connective, annotators rely more heavily on their particular interpretation of the relation, which increases variability and reduces mapping consistency across frameworks. In comparison, explicit discourse relations benefit from the presence of explicit connectives, which anchor their intended meaning and allow more objective annotations.

Although no previous work has attempted an empirical mapping between the RST and the PDTB 3.0 frameworks, it is possible to contextualize the results of the mapping approach presented in this chapter by comparing them to the empirical mapping between the RST and the PDTB 2.0 frameworks proposed by Demberg et al. (2019). Both studies focus on mapping discourse relations within the same set of Wall Street Journal articles. While Demberg et al. (2019) successfully mapped 76% of discourse relations between the RST-DT and the PDTB 2.0 corpora, the approach presented in this chapter achieves a slightly higher mapping coverage of 80% between the RST-DT and the PDTB 3.0 corpora. However, since the present work builds on a revised version of the PDTB framework, this 4% increase in coverage cannot be directly attributed to the mapping methodology alone. It may also result from updates in the PDTB 3.0 annotation guidelines which might facilitate the alignment with the RST framework. Further analysis is needed to isolate the effects of the different proposed mapping approaches and the PDTB framework revision differences before drawing definitive conclusions on the relative effectiveness of either approach.

Future work should consider extending the proposed mapping approach to include the AltLex and AltLexC discourse relations in the PDTB 3.0. Although these relation types represent only 3.1% of the total annotations in the corpus (see Table 2.4), they are also annotated with discourse senses, similar to explicit and implicit discourse relations, and thus offer additional opportunities for cross-framework mapping. Moreover, the current approach was evaluated solely on manually annotated discourse relations. As more reliable RST and PDTB 3.0 discourse parsers become available, future studies should evaluate the proposed mapping algorithm on automatically parsed texts. Such an evaluation would provide further validation of the proposed mapping approach and extend its utility beyond gold-standard corpora. Finally, future work should extended the proposed mapping to other discourse frameworks such as the Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003; Lascarides and Asher, 2007) and the Enhanced Rhetorical Structure Theory (eRST) (Zeldes et al., 2025). Incorporating these frameworks would broaden the scope of discourse interoperability and provide deeper insights into how different frameworks converge and diverge in practice.

## Chapter 4

# Multi-Label Implicit Discourse Relation Recognition

Building on the insights of the cross-framework mapping between the RST and the PDTB 3.0 frameworks presented in Chapter 3, this chapter turns to the second research objective of this thesis introduced in Chapter 1: **to develop a multi-label IDRR classification model capable of capturing multiple implicit senses**. While the mapping results presented in Chapter 3 showed some points of convergence across discourse frameworks, they also revealed that implicit discourse relations often lead to more ambiguous interpretations. This observation motivates the need for models that can move beyond single-label classification and instead capture the multiple possible senses of implicit relations. This chapter presents a multi-label IDRR classification model designed to predict multiple senses simultaneously following the PDTB 3.0 framework. The majority of this chapter is based on the paper published at the Special Interest Group on Discourse and Dialogue (SIGDIAL) 2025 (Costa and Kosseim, 2025a).

Implicit discourse relation recognition (IDRR) is one of the most challenging tasks in computational discourse analysis. The goal of IDRR is to identify the sense of discourse relations connecting text arguments in the absence of explicit discourse connectives, such as *but* and *because*. The PDTB 3.0 framework organizes discourse senses across three hierarchical levels with increasing degrees of detail (see Table 2.3). However, understanding the sense of a discourse relation can be a complex and subjective task without the guidance of discourse connectives. Consider the two arguments of the implicit discourse relation used in Example 3 of Chapter 1:

**Arg1:** The lights flickered.

**Arg2:** The power went out.

In this example, it is not clear whether the two arguments of the relation are related by a temporal or causality relation. One could interpret that the lights flicked and *then* the power went out, or that the lights flickered *because* the power went out. This ambiguity illustrates the main challenge of IDRR. Despite its difficulty, IDRR plays a crucial role in downstream applications that rely on text coherence. In text summarization, for example, it helps preserve the logical flow and rhetorical intent of the source text (Liu and Demberg, 2024; Zhong and Litman, 2025). In text classification, it enriches predictions by capturing the implicit relations and contextual meaning of sentences within the broader text (Huber and Carenini, 2020; Lei and Huang, 2023). In dialogue systems, it supports coherence and intent modeling by inferring how utterances relate to each other (Li et al., 2023, 2024b).

To date, most research on IDRR has relied on the different iterations of the PDTB corpus (Prasad et al., 2006, 2007; Webber et al., 2019). However, despite significant efforts, state-of-the-art (SOTA) performance on IDRR has plateaued in recent years at F1-Scores of 72.07 at level-1 and 57.62 at level-2 (Long and Webber, 2022; Wu et al., 2023; Zhao et al., 2023; Liu and Strube, 2023; Chan et al., 2023; Jiang et al., 2023a; Zeng et al., 2024; Long and Webber, 2024). One possible reason can be attributed to the inherently subjective nature of implicit discourse interpretation, which can be difficult to capture using predominantly single-label annotated corpora (Stede, 2008; Rohde et al., 2016; Scholman and Demberg, 2017; Hoek et al., 2021; Ru et al., 2023; López Cortez and Jacobs, 2023; Liu et al., 2024; Aktas and Roth, 2025). In light of this challenge, recent studies have advocated for the multi-label annotation of implicit relations to better capture their nuanced and complex nature in discourse corpora (Yung et al., 2019; Pyatkin et al., 2020; Scholman et al., 2022a,b; Marchal et al., 2022; Pyatkin et al., 2023; Yung et al., 2024b; Yung and Demberg, 2025). This paradigm shift has led to the creation of DiscoGeM (Scholman et al., 2022a) - the first multi-label annotated corpus of implicit discourse relations (see Section 2.2.2).

Following recent work in NLP (Pavlick and Kwiatkowski, 2019; Chen et al., 2020a; Nie et al., 2020; Fornaciari et al., 2021; Meissner et al., 2021; Basile et al., 2021; Uma et al., 2021; Plank, 2022; Jiang and de Marneffe, 2022; Jayaweera and Dorr, 2025), the multi-label approach to IDRR presented in this chapter intends to see disagreement between humans annotators not as noise, but as an opportunity to enhance the relevance of IDRR models by capturing the complexity of human reasoning. Therefore, this chapter addresses the subjectivity in IDRR by proposing a multi-task classification model, MTask, capable of learning both multi-label and single-label representations of implicit discourse relations using the DiscoGeM corpus. The model is then evaluated on the traditional single-label classification task using the PDTB 3.0 corpus in a transfer learning setting.

## 4.1 Previous Work

The advent of transformer-based models (Vaswani et al., 2017) has led to significant advances in natural language understanding. Despite this progress, IDRR remains one of the most challenging tasks in computational discourse analysis. This section reviews previous work on IDRR, with a focus on how SOTA approaches have evolved in recent years. Traditionally framed as a single-label classification task, recent approaches have addressed IDRR by fine-tuning (Long and Webber, 2022; Liu and Strube, 2023; Jiang et al., 2023a) or prompt-tuning (Wu et al., 2023; Zhao et al., 2023; Chan et al., 2023; Zeng et al., 2024; Long and Webber, 2024) pre-trained language models (PLMs). More precisely:

- Long and Webber (2022) apply contrastive learning (Chen et al., 2020b) by leveraging the full three-level sense hierarchy of the PDTB 3.0 to guide the selection of negative examples. Rather than treating the hierarchy merely as a label set, they use it to structure the embedding space such that closely related senses are drawn together while unrelated ones are pushed apart. Their experiments show how different strategies for negative sampling and a simple data augmentation technique, consisting of inserting explicit connectives from PDTB 3.0 metadata to duplicate implicit discourse relations, can further enhance performance on IDRR.
- Wu et al. (2023) introduce a connective prediction approach, CP-KD, based on knowledge distillation to improve IDRR. By training a student model with soft labels generated by a teacher model that captures latent correlations between connectives and discourse senses, the method overcomes limitations of hard one-hot targets and reduces dependency on handcrafted templates. Their approach also reduces reliance on rigid template designs and demonstrates flexibility by generalizing across different templates and extending to explicit discourse relation recognition.
- Zhao et al. (2023) present a parameter-efficient framework, PEMI, that leverages prompt-tuning to mitigate data scarcity in IDRR and preserve the structure of PLMs. The proposed approach uses soft prompts to align input arguments with the pre-trained feature space and introduces a hierarchical label refining method to infuse discourse sense hierarchy into the verbalizer component. By dynamically aggregating lower-level label representations into higher-level categories, the model enables joint learning across all three levels in the PDTB 3.0.

- Liu and Strube (2023) propose a joint model for IDRR that mimics the annotation process of the PDTB 3.0 corpus by explicitly generating discourse connectives between argument pairs and using them to help relation prediction. To address the challenge of low-quality generated connectives in early training stages, the model gradually shifts from gold to generated connectives during training. This end-to-end approach enables the model to better balance the use of arguments and connectives for inference and demonstrates that including connectives as input, rather than as auxiliary objectives, enhances performance.
- Chan et al. (2023) introduce a prompt-based model, DiscoPrompt, that models IDRR as a path prediction task over the hierarchical structure of discourse sense labels. Instead of predicting flat labels or relying solely on connective cues, the proposed method constructs hierarchical prompts that explicitly encode the relation path and adds it to beginning of the input. By integrating both hierarchical label information and connective examples into prompt-tuning, the model captures richer semantic dependencies between discourse senses and improves relational prediction.
- Jiang et al. (2023a) present a contrastive learning framework, GOLF, that leverages both global and local hierarchical structures of discourse sense labels. The global hierarchy captures the full taxonomy of discourse relations, while the local hierarchy reflects the specific label sequence associated with each instance. By incorporating both perspectives through multi-task learning and two complementary contrastive objectives, one aligning text and label embeddings and the other regulating representation distances based on label sequence similarity, the model learns more structured and semantically informed discourse relation representations.
- Zeng et al. (2024) propose a prompt-tuning framework, GLHPT, for IDRR that incorporates both global and local hierarchical knowledge. The proposed method integrates global hierarchy information directly into the output distribution of a multi-level verbalizer using a bottom-up propagation strategy, while also modeling local instance-specific label sequences through contrastive learning. By jointly optimizing both aspects, the framework adapts PLMs more effectively to the hierarchical structure of discourse relations without relying on additional data or structural modules.
- Long and Webber (2024) suggest an alternative to manually designed verbalizers for IDRR by introducing a prototype-based approach that leverages the hierarchical structure of sense labels. Instead of relying on ambiguous or overlapping connectives, the method learns semantic prototype representations for each sense and uses

contrastive learning to structure the distances between instances and sense prototypes in accordance with the discourse hierarchy.

While most of these approaches (Long and Webber, 2022; Wu et al., 2023; Zhao et al., 2023; Liu and Strube, 2023; Jiang et al., 2023a; Zeng et al., 2024; Long and Webber, 2024) use RoBERTa<sub>base</sub> (Liu et al., 2019) as their PLM backbone, Chan et al. (2023) uses T5<sub>base</sub> (Raffel et al., 2020). In a different line of work, the recent shift toward multi-label annotation in discourse data has led to an initial wave of research on multi-label IDRR (Yung et al., 2022; Long et al., 2024). For instance:

- Yung et al. (2022) explore the benefits of modeling implicit discourse relations as multi-label distributions using the DiscoGeM corpus. While they aim to capture the ambiguity inherent in implicit discourse relations, their approach converts their label distributions into multiple single-label instances by duplicating each instance in proportion to its label frequency. This allows the model to capture some of the variability present in multi-label annotations. However, their model is ultimately trained following a single-label methodology.
- Long et al. (2024) treat IDRR as a multi-label classification problem. However, while they claim to adopt a multi-label perspective, they train their model on the PDTB 3.0 corpus where only 4.9% of the implicit relations are annotated with two senses. The authors use these few instances to build a model that can at most predict up to two senses per relation. Yet, since the vast majority of the training data (95.1%) is annotated with a single-label, the model largely behaves as a single-label classifier. Despite this, the study highlights the potential benefits of considering multi-label annotations to capture the ambiguity of implicit discourse relations.

These efforts mark an important step toward acknowledging the ambiguity and subjectivity inherent to IDRR. However, they fall short of embracing a fully multi-label representation of implicit discourse relations. The next sections address this gap by introducing a novel approach to IDRR focusing on multi-label representations of implicit discourse relations. To enable meaningful comparisons with previous work, which almost exclusively frames IDRR as a single-label task, the proposed multi-label approach is also adapted to generate single-label predictions. This dual evaluation demonstrates the versatility of the proposed approach in both multi-label and single-label settings and facilitates direct comparison with existing IDRR models.

## 4.2 Data Preparation

To allow the evaluation of the proposed approach in a multi-label and single-label setting, the work presented in this chapter is based on the DiscoGeM and the PDTB 3.0 discourse annotated corpora (see Sections 2.2.1 and 2.2.2). The model is trained exclusively on DiscoGeM and evaluated both on DiscoGeM (for multi-label and single-label classification) and PDTB 3.0 (for single-label classification). Although the two corpora differ in their annotation methodologies, both are grounded in the PDTB 3.0 discourse framework (see Section 2.2.2). This shared theoretical foundation allows the evaluation of the proposed model across both corpora in a transfer learning setting.

### 4.2.1 DiscoGeM

Most work in IDRR adopts the standard set of 14 level-2 senses proposed by Kim et al. (2020), which corresponds to the senses that appear at least 100 times in the PDTB 3.0 corpus. While all level-1 senses are typically included in IDRR, level-3 senses are rarely considered due to the difficulty in predicting them. The model proposed in this chapter, however, is trained to predict discourse senses across all three sense levels simultaneously. In order to effectively evaluate the potential of transfer learning from the DiscoGeM to the PDTB 3.0 corpus, it is first necessary to align the set of labels used for training and evaluating the model across both corpora. Following the common approach to IDRR, a slightly modified version of standard set of 14 level-2 senses proposed by Kim et al. (2020) is used in this chapter.

Unlike the PDTB 3.0 corpus, the DiscoGeM corpus provides multi-label sense annotations of discourse relations. After filtering out senses not included in the standard set of 14 level-2 senses, the remaining sense distributions were renormalized to ensure they summed to 1 for each instance. However, DiscoGeM does not include the CAUSE+BELIEF sense from the standard set of 14 level-2 senses (see Section 2.2.2). To maintain a consistent set of 14 labels, this missing sense was replaced by the SIMILARITY sense, which was the most frequent among the remaining level-2 senses. Table 4.1 shows the distribution of senses across all three levels in the original DiscoGeM corpus and in the adapted label set used in this chapter. Finally, the level-2 senses SYNCHRONOUS, CONTRAST, SIMILARITY, CONJUNCTION and EQUIVALENCE in the 14 label set do not have corresponding level-3 senses. In order to conduct IDRR at level-3, their level-3 sense label was defaulted to their corresponding level-2 label.

To adapt the DiscoGeM corpus for single-label IDRR, the multi-label sense distribution

| Level-1     | Sum      |         | Level-2                | Sum      |         | Level-3              | Sum      |         |
|-------------|----------|---------|------------------------|----------|---------|----------------------|----------|---------|
|             | Original | Adapted |                        | Original | Adapted |                      | Original | Adapted |
| TEMPORAL    | 584.3    | 619.5   | <b>SYNCHRONOUS</b>     | 95.8     | 102.7   | SYNCHRONOUS*         | 95.8     | 102.7   |
|             |          |         | <b>ASYNCHRONOUS</b>    | 488.5    | 516.8   | PRECEDENCE           | 448.1    | 474.2   |
|             |          |         |                        |          |         | SUCCESSION           | 40.5     | 42.7    |
| CONTINGENCY | 1,745.1  | 1,822.9 | <b>CAUSE</b>           | 1,740.2  | 1,819.0 | REASON               | 382.5    | 400.1   |
|             |          |         |                        |          |         | RESULT               | 1,357.7  | 1,418.9 |
|             |          |         |                        |          |         | NEGRESULT            | 0.0      | 0.0     |
|             |          |         | <b>CONDITION</b>       | 1.2      | 1.2     | ARG1-AS-COND         | 0.1      | 0.1     |
|             |          |         |                        |          |         | ARG2-AS-COND         | 1.1      | 1.1     |
|             |          |         | NEG-CONDITION          | 1.1      | 0.0     | ARG1-AS-NEGCOND      | 1.0      | 0.0     |
|             |          |         |                        |          |         | ARG2-AS-NEGCOND      | 0.1      | 0.0     |
|             |          |         | <b>PURPOSE</b>         | 2.6      | 2.7     | ARG1-AS-GOAL         | 1.7      | 1.7     |
| COMPARISON  | 831.5    | 878.0   | <b>CONCESSION</b>      | 517.1    | 548.7   | ARG1-AS-DENIER       | 169.3    | 179.9   |
|             |          |         |                        |          |         | ARG2-AS-DENIER       | 347.7    | 368.8   |
|             |          |         | <b>CONTRAST</b>        | 202.4    | 213.4   | CONTRAST*            | 202.4    | 213.4   |
|             |          |         | <b>SIMILARITY</b>      | 112.0    | 116.0   | SIMILARITY*          | 112.0    | 116.0   |
| EXPANSION   | 3,034.6  | 3,184.6 | <b>CONJUNCTION</b>     | 1,441.1  | 1,518.0 | CONJUNCTION*         | 1,441.1  | 1,518.0 |
|             |          |         | DISJUNCTION            | 2.9      | 0.0     | -                    | -        | -       |
|             |          |         | <b>EQUIVALENCE</b>     | 19.0     | 19.8    | EQUIVALENCE*         | 19.0     | 19.8    |
|             |          |         | EXCEPTION              | 1.4      | 0.0     | ARG1-AS-EXCEPTION    | 0.3      | 0.0     |
|             |          |         |                        |          |         | ARG2-AS-EXCEPTION    | 1.3      | 0.0     |
|             |          |         | <b>INSTANTIATION</b>   | 388.8    | 403.8   | ARG1-AS-INSTANCE     | 16.9     | 17.8    |
|             |          |         |                        |          |         | ARG2-AS-INSTANCE     | 371.9    | 386.0   |
|             |          |         | <b>LEVEL-OF-DETAIL</b> | 1,137.0  | 1,196.1 | ARG1-AS-DETAIL       | 160.8    | 170.2   |
|             |          |         |                        |          |         | ARG2-AS-DETAIL       | 976.2    | 1,025.9 |
|             |          |         | <b>MANNER</b>          | 4.6      | 4.8     | ARG1-AS-MANNER       | 1.3      | 1.3     |
|             |          |         |                        |          |         | ARG2-AS-MANNER       | 3.3      | 3.4     |
|             |          |         | <b>SUBSTITUTION</b>    | 39.7     | 42.1    | ARG1-AS-SUBSTITUTION | 0.0      | 0.0     |
|             |          |         |                        |          |         | ARG2-AS-SUBSTITUTION | 39.7     | 42.1    |

Table 4.1: Distribution of senses across all three levels in the original DiscoGeM corpus and in the adapted set of 14 level-2 senses (in bold) used in this chapter. Each value represents the sum of the corresponding sense across all instances. Level-3 senses that did not have a sense label (marked with \*) were defaulted to their level-2 sense as replacement.

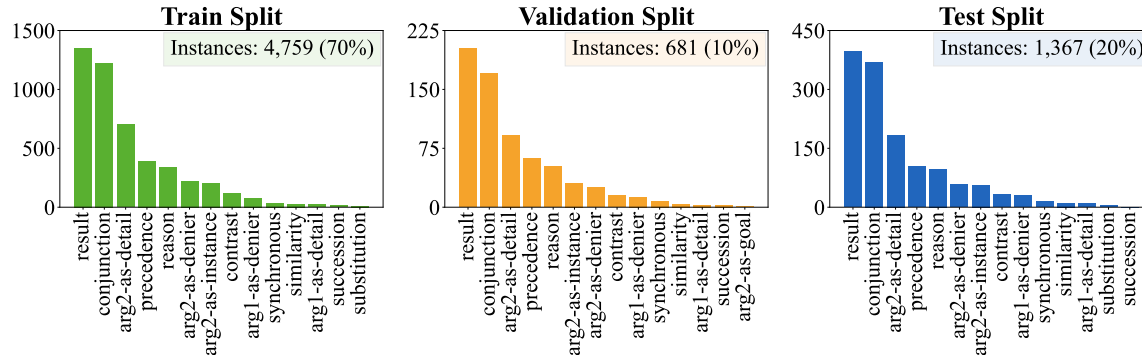


Figure 4.1: Distribution of majority label senses across all splits of the DiscoGeM corpus.

of each discourse relation was replaced by the sense label with the highest score - referred to as the majority label. To reduce computational complexity and maintain consistency across model evaluations, the same data split was used for both multi-label and single-label experiments. The corpus was divided into 70% for training, 10% for validation and 20% for testing. To ensure a balanced distribution, all splits follow the same distribution of majority labels, as shown in Figure 4.1. Due to the extensive number of experiments conducted in this chapter, a fixed data split was chosen over cross-validation to streamline the experimentation process.

## 4.2.2 PDTB 3.0

Since there are currently no benchmarks for multi-label IDRR using DiscoGeM and since most previous work has framed IDRR as a single-label classification task using the PDTB 3.0 corpus, the model proposed in this chapter is also evaluated on PDTB 3.0 corpus via transfer learning. To enable a fair comparison, the same adapted set of 14 level-2 senses used for DiscoGeM was applied to PDTB 3.0. However, following the standard practice in IDRR, only level-1 and level-2 sense labels were used for this evaluation. To compare the performance of the model against previous work, the model was evaluated on the two commonly used test splits of the PDTB 3.0 corpus: the Lin (Lin et al., 2009) and the Ji (Ji and Eisenstein, 2015) test splits. The former uses section 23 in the PDTB 3.0 corpus as the test set, while the latter uses sections 21 and 22. However, to address recent concerns regarding the reliability of these small test sets for meaningful evaluation (Shi and Demberg, 2017; Kim et al., 2020), the model was also evaluated following the cross-validation scheme proposed by Kim et al. (2020). Table 4.2 summarizes the number of level-2 sense instances across the different PDTB 3.0 test splits.

| Level-1     | Level-2         | DiscoGeM | PDTB 3.0 |       |                       |
|-------------|-----------------|----------|----------|-------|-----------------------|
|             |                 | Test     | Lin      | Ji    | Cross                 |
| TEMPORAL    | SYNCHRONOUS     | 13       | 19       | 43    | $42.85 \pm 11.94$     |
|             | ASYNCHRONOUS    | 103      | 33       | 105   | $99.54 \pm 24.95$     |
| CONTINGENCY | CAUSE           | 515      | 271      | 406   | $465.38 \pm 77.28$    |
|             | CONDITION       | 0        | 10       | 15    | $16.31 \pm 3.61$      |
|             | PURPOSE         | 0        | 46       | 89    | $110.08 \pm 22.60$    |
| COMPARISON  | CONCESSION      | 104      | 83       | 98    | $121.23 \pm 22.87$    |
|             | CONTRAST        | 27       | 52       | 54    | $69.92 \pm 18.99$     |
|             | SIMILARITY      | 10       | 1        | 2     | $2.23 \pm 1.48$       |
| EXPANSION   | CONJUNCTION     | 336      | 161      | 236   | $349.38 \pm 86.25$    |
|             | EQUIVALENCE     | 0        | 24       | 30    | $27.54 \pm 10.27$     |
|             | INSTANTIATION   | 53       | 70       | 124   | $118.62 \pm 24.97$    |
|             | LEVEL-OF-DETAIL | 202      | 194      | 208   | $263.69 \pm 26.94$    |
|             | MANNER          | 0        | 4        | 17    | $17.92 \pm 6.49$      |
|             | SUBSTITUTION    | 4        | 27       | 26    | $30.23 \pm 8.28$      |
| Total       |                 | 1,367    | 995      | 1,453 | $1,734.92 \pm 271.89$ |

Table 4.2: Distribution of level-2 majority label senses across the test set of DiscoGeM and the different test splits of the PDTB 3.0. The values in PDTB 3.0 Cross correspond to the averaged distribution of majority label senses across all 12 test folds.

| Top-1      | Top-3       | Top-5       | Top-10      |
|------------|-------------|-------------|-------------|
| 92 (30.5%) | 164 (54.3%) | 197 (65.2%) | 215 (71.2%) |

Table 4.3: Number of instances where the reference label in the PDTB 3.0 was found within the top-k labels in DiscoGeM for the set of 302 co-annotated relations.

### 4.2.3 Corpora Agreement

Given that a total of 302 implicit discourse relations taken from the PDTB 3.0 were also independently annotated in DiscoGeM, it was possible to calculate the annotation agreement between the two corpora over this overlapping set of relations. Table 4.3 reports the number of relations where the PDTB 3.0 reference label matches the majority label (top-1) in DiscoGeM, or appears within its top-3, top-5 or top-10 labels. The fact that for 28.8% of the jointly annotated relations, the reference label from the PDTB 3.0 corpus was not selected by at least one of the annotators of DiscoGeM illustrates the variability in human interpretation of implicit discourse relations - even when annotations are based on the same underlying framework.

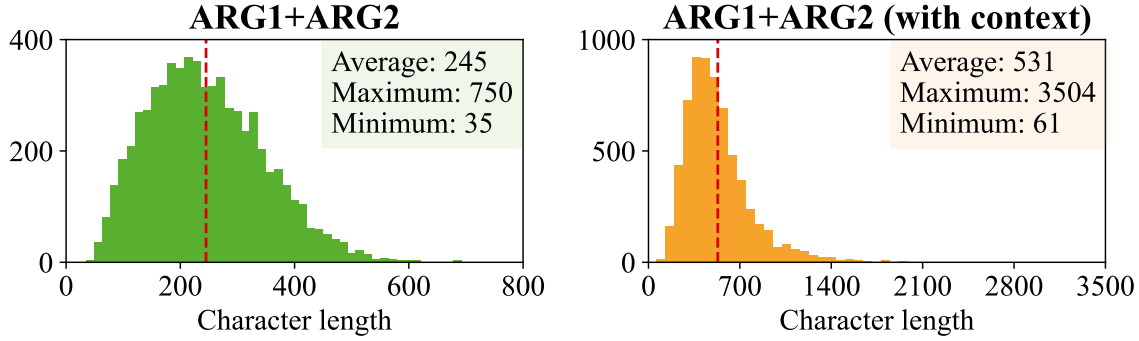


Figure 4.2: Distribution of the character length size for the argument pairs of the discourse relations in the DiscoGeM corpus with and without additional textual context.

### 4.3 Preliminary Study

Before introducing the proposed multi-label IDRR model, this section presents a preliminary study based on the paper published at the Workshop on Computational Approaches to Discourse (CODI) 2024 (Costa and Kosseim, 2024). Its objective is to better understand the characteristics of implicit discourse relations in the DiscoGeM corpus and identify factors that may enhance model performance. More precisely, this section investigates whether providing additional contextual information to the arguments of implicit discourse relations can improve the performance of a simple IDRR classification model. To this end, two datasets were created from the DiscoGeM corpus. The first dataset includes only the argument pairs associated with each multi-label sense annotation, while the second provides additional context to the argument pairs by including the surrounding sentences immediately before and after each argument. Figure 4.2 shows the distribution of the character length size for the argument pairs of the discourse relations in these two datasets. The instances in the dataset containing only the arguments of the discourse relations have an average length of 245 characters and a maximum length of 750 characters. In contrast, the instances in the dataset with added context have an average length of 531 characters and the longest instance has 3,504 characters.

The model is trained using the multi-label sense annotations provided in the DiscoGeM corpus but evaluated only on its ability to predict a single sense at the third-level of the PDTB 3.0 sense hierarchy. This evaluation setup was chosen to facilitate this preliminary study with a simple IDRR classification model. Following most approaches described in Section 4.1, the model architecture consists of a RoBERTa<sub>base</sub> (Liu et al., 2019) encoder

followed by a single feedforward classification layer. The model is trained to predict multi-label senses from which the predicted sense with the highest score in the multi-label distribution is selected and compared against the sense with the highest score at the input. The model uses the Adam optimizer (Kingma and Ba, 2015) to minimize a soft cross-entropy loss with a mean reduction over batches to allow training with multi-label distributions.

To ensure a balanced distribution of senses in the training and evaluation of the model, the data was split based on the third-level PDTB 3.0 sense with the highest score for each relation. This strategy helps ensure a similar distribution of highest scoring senses across the training, validation and testing splits. To fine-tune the model, a series of simple experiments were conducted using the validation split of the dataset without context to determine the optimal hyperparameter configuration. Table 4.4 shows the results of varying the number of training epochs and batch sizes, while keeping a constant learning rate of  $1 \times 10^{-5}$  and using no decay function. Conversely, Table 4.5 shows the results of varying the learning rate and the impact of introducing a linear decay function with a decay rate of 0.1, for 10 epochs and a batch size of 64.

| EP / BS        | F1           | Precision    | Recall       |
|----------------|--------------|--------------|--------------|
| 10 / 16        | 49.98        | 49.55        | 51.35        |
| 10 / 32        | 50.91        | 50.62        | 51.58        |
| <b>10 / 64</b> | <b>51.38</b> | <b>51.54</b> | <b>52.19</b> |
| 20 / 64        | 50.59        | 50.67        | 51.04        |

Table 4.4: Evaluation of the model with different epochs (EP) and batch sizes (BS), with a constant learning rate of  $1 \times 10^{-5}$  and no decay.

| LR / DR                           | F1           | Precision    | Recall       |
|-----------------------------------|--------------|--------------|--------------|
| $5e^{-5}$ / 0.0                   | 48.77        | 49.91        | 49.42        |
| $2e^{-5}$ / 0.0                   | 49.43        | 49.64        | 50.88        |
| <b><math>1e^{-5}</math> / 0.0</b> | <b>51.38</b> | <b>51.54</b> | <b>52.19</b> |
| $1e^{-5}$ / 0.1                   | 49.67        | 50.78        | 50.73        |

Table 4.5: Evaluation of the model with different learning rates (LR) and decay rates (DR), for 10 epochs and a batch size of 64.

The values highlighted in bold in both Table 4.4 and Table 4.5 indicate the optimal hyperparameter configuration of the model in this preliminary study: 10 epochs, batch size of 64, learning rate of  $1 \times 10^{-5}$  and no decay function. Following this configuration, the model was trained and evaluated separately on the datasets with and without additional context. Table 4.6 presents the results of the model on both datasets.

The results in Table 4.6 suggest that providing additional context to the arguments of implicit discourse relations does not improve model performance. In fact, it leads to worse results. One possible explanation could be due to the nature of the additional context used. In this study, the additional context was selected from the sentences immediately preceding

| Dataset                      | F1           | Precision    | Recall       |
|------------------------------|--------------|--------------|--------------|
| <b>No Additional Context</b> | <b>51.38</b> | <b>51.54</b> | <b>52.19</b> |
| With Additional Context      | 43.67        | 43.22        | 45.43        |

Table 4.6: Results of single-label IDRR at the third-level of the PDTB 3.0 sense framework using arguments of implicit discourse relations in the DiscoGeM corpus with and without additional context.

and following each argument, but these may not always contain the most relevant information to interpret the discourse relation. Manually selecting more targeted contextual information for each discourse relation might have led to better results. Another potential factor is input truncation due to the inclusion of longer sequences. However, since RoBERTa<sub>base</sub> accepts inputs up to 512 tokens and the average length of the inputs with additional context is of 531 characters (as shown in Figure 4.2), not much information should be lost to truncation.

A more comprehensive study would have to be conducted before drawing a definite conclusion, but this preliminary study suggests that for the 6,807 implicit discourse relations in the DiscoGeM corpus (see Section 2.2.2) the potential informational gain from additional context may not outweigh the increased input complexity. As a result, all subsequent IDRR models presented in this thesis will be trained using only the arguments of the discourse relations without any additional context.

## 4.4 Methodology

This section presents a multi-task IDRR classification model, MTask, capable of learning both multi-label and single-label representations of implicit discourse relations across all three sense levels defined in the PDTB 3.0 framework. Figure 4.3 illustrates the architecture of the model. The model takes as input a concatenated pair of discourse arguments, which is encoded using RoBERTa<sub>base</sub> (Liu et al., 2019). The resulting contextualized embedding is passed through a shared linear and dropout layer and subsequently fed to three distinct classification heads - each corresponding to a sense level in the hierarchy. In the multi-label setting, each head outputs a probability distribution over the senses at its corresponding level. For single-label predictions, a pooling layer selects the sense with the highest probability from each distribution. The model is trained using the Adam optimizer (Kingma and Ba, 2015) and the loss function is calculated as the sum of the individual classification losses of each head.

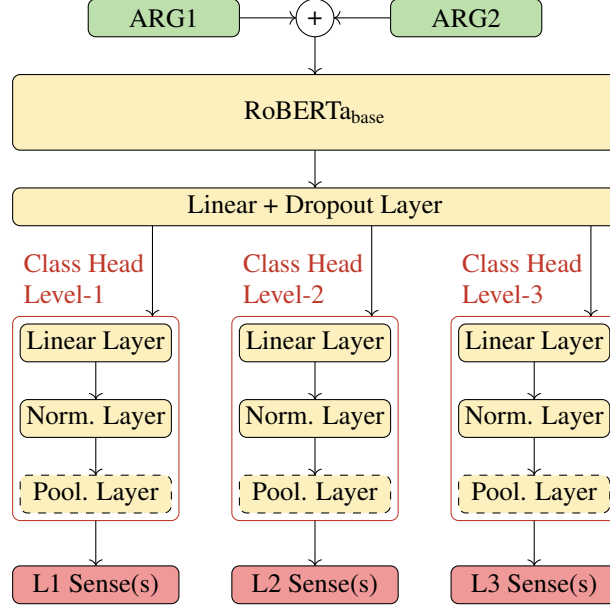


Figure 4.3: Architecture of the multi-task model, MTask, for IDRR. Given a pair of discourse arguments as input, each classification head of the model generates an output for the corresponding sense level in the PDTB 3.0 framework. In the multi-label setting, each output is a probability distribution over the senses at its corresponding level. In the single-label setting, the sense with the highest probability is selected via the pooling layers (indicated by dashed lines).

#### 4.4.1 Multi-Label Classification

In the multi-label setting, the loss of each classification head is computed using the mean absolute error (MAE) loss function, which was selected based on its superior performance in the experiments that will be discussed in Section 4.5.1. The performance of the model in the multi-label setting is evaluated using the Jensen-Shannon (JS) distance (Lin, 1991) to measure the similarity between the predicted and reference sense probability distributions. This choice follows previous work on multi-label classification for NLP tasks (Meissner et al., 2021; Uma et al., 2021; Pyatkin et al., 2023; Yung and Demberg, 2025), where probabilistic outputs reflect annotator disagreement and uncertainty.

#### 4.4.2 Single-Label Classification

In the single-label setting, the loss of each classification head is computed using the cross-entropy (CE) loss function which emphasizes the correct classification of the majority label sense. This choice is supported by the experiments discussed later in Section 4.5.1. The performance of the model in the single-label setting is computed using the weighted

F1-score for the majority labels in the predicted and reference sense distributions.

## 4.5 Experiments

To optimize the performance of the proposed MTask model, a series of experiments were conducted to evaluate different model configurations. These experiments were performed using the validation split of the DiscoGeM corpus to ensure that hyperparameter fine-tuning and model selection were guided by held-out data. All reported results in the subsequent sections represent the average performance across three independent runs, each initialized with a different random seed. Models were trained for 10 epochs with a batch size of 16.

### 4.5.1 Model and Loss Function Selection

Motivated by prior work suggesting that PLMs trained with a next sentence prediction (NSP) objective perform better on IDRR (Shi and Demberg, 2019), this experiment compares using BERT<sub>base</sub> (Devlin et al., 2019), which includes the NSP objective, and RoBERTa<sub>base</sub> (Liu et al., 2019), which omits it, as the PLM encoder of the MTask model. RoBERTa<sub>base</sub> has also been widely adopted in most previous work in IDRR (Long and Webber, 2022; Wu et al., 2023; Zhao et al., 2023; Liu and Strube, 2023; Jiang et al., 2023a; Zeng et al., 2024; Long and Webber, 2024), making it a strong baseline. To further explore the trade-off between performance and computational efficiency, distilled versions of both models were also evaluated (Sanh et al., 2019). Each model configuration was tested using different loss functions to evaluate their impact on performance under both multi-label and single-label classification settings. Letting  $\hat{y}^h$  denote the predicted distribution of the classification head  $h$  (where  $h \in 1, 2, 3$  - see Figure 4.3),  $y^h$  the corresponding reference distribution,  $C^h$  the number of sense classes at level- $h$  and  $N$  the number of instances per batch, the following loss functions were evaluated:

- Cross-Entropy (CE)

$$loss^h(\hat{y}^h, y^h) = -\frac{1}{N} \sum_{n=1}^N \sum_{c=1}^{C^h} \log \left( \frac{e^{\hat{y}_{n,c}^h}}{\sum_{i=1}^{C^h} e^{\hat{y}_{n,i}^h}} \right) y_{n,c}^h \quad (1)$$

- Mean Absolute Error (MAE)

$$loss^h(\hat{y}^h, y^h) = \frac{1}{NC^h} \sum_{n=1}^N \sum_{i=1}^{C^h} |\hat{y}_{n,i}^h - y_{n,i}^h| \quad (2)$$

- Mean Squared Error (MSE)

$$loss^h(\hat{y}^h, y^h) = \frac{1}{NC^h} \sum_{n=1}^N \sum_{i=1}^{C^h} (\hat{y}_{n,i}^h - y_{n,i}^h)^2 \quad (3)$$

- Huber Loss (Huber)

$$loss^k(\hat{y}^h, y^h) = \sum_{n=1}^N \sum_{i=1}^{C^h} \begin{cases} \frac{(\hat{y}_{n,i}^h - y_{n,i}^h)^2}{2NC^h} & , \text{ if } |\hat{y}_{n,i}^h - y_{n,i}^h| < 0.5 \\ \frac{2|\hat{y}_{n,i}^h - y_{n,i}^h| - 1}{2NC^h} & , \text{ otherwise} \end{cases} \quad (4)$$

The results of these experiments are presented in Table 4.7. As the results in the table show, RoBERTa<sub>base</sub> consistently outperforms BERT<sub>base</sub>, yielding lower Jensen-Shannon (JS) distances in the multi-label setting and higher F1-scores in the single-label setting across all sense levels and loss functions. These findings suggest that the representational benefits of using larger models, such as RoBERTa<sub>base</sub>, outweigh the potential advantages previously associated with models trained with the NSP objective in IDRR (Shi and Demberg, 2019). Additionally, both BERT<sub>base</sub> and RoBERTa<sub>base</sub> outperform their respective distilled variants, reaffirming the importance of model size for this task. The results in Table 4.7 also show that the choice of loss function plays an important role. Models trained with MAE, MSE and the Huber loss function consistently achieve lower JS distances in the multi-label setting, while model trained with CE loss function achieve higher F1-scores in the single-label setting.

The mathematical formulation of the loss functions evaluated in these experiments significantly influences the behavior of the model during training. The MAE and MSE losses in Equations 2 and 3, respectively, aim to minimize the overall difference between the predicted and reference distributions by penalizing errors across all possible labels. MAE minimizes the average absolute differences, leading to predictions that are closer to the reference distribution in an averaged sense, while MSE places a larger penalty on larger errors due to its quadratic nature, which can result in a stronger emphasis on outliers. The Huber loss in Equation 4 combines the properties of MAE and MSE in function of the  $|\hat{y}_{n,i}^h - y_{n,i}^h| < 0.5$  condition. In other words, the Huber loss behaves like MAE for smaller

| PLM                           | Loss  | JS Distance $\searrow$ (Multi-Label) |                                     |                                     | F1-Score $\nearrow$ (Single-Label) |                                    |                                    |
|-------------------------------|-------|--------------------------------------|-------------------------------------|-------------------------------------|------------------------------------|------------------------------------|------------------------------------|
|                               |       | Level-1                              | Level-2                             | Level-3                             | Level-1                            | Level-2                            | Level-3                            |
| BERT <sub>base</sub>          | MAE   | 0.314 $\pm$ 0.004                    | 0.461 $\pm$ 0.007                   | 0.535 $\pm$ 0.004                   | 63.34 $\pm$ 1.41                   | 46.44 $\pm$ 0.82                   | 39.20 $\pm$ 1.82                   |
|                               | MSE   | 0.319 $\pm$ 0.005                    | 0.462 $\pm$ 0.002                   | 0.544 $\pm$ 0.005                   | 63.82 $\pm$ 0.83                   | 48.68 $\pm$ 0.16                   | 41.36 $\pm$ 1.55                   |
|                               | Huber | 0.318 $\pm$ 0.007                    | 0.463 $\pm$ 0.001                   | 0.542 $\pm$ 0.002                   | 64.36 $\pm$ 0.48                   | 49.45 $\pm$ 0.96                   | 41.58 $\pm$ 1.08                   |
|                               | CE    | 0.328 $\pm$ 0.004                    | 0.563 $\pm$ 0.002                   | 0.634 $\pm$ 0.004                   | 63.49 $\pm$ 1.38                   | 51.42 $\pm$ 0.58                   | 47.65 $\pm$ 0.92                   |
| DistilBERT <sub>base</sub>    | MAE   | 0.330 $\pm$ 0.006                    | 0.472 $\pm$ 0.003                   | 0.543 $\pm$ 0.003                   | 61.29 $\pm$ 0.96                   | 39.29 $\pm$ 0.66                   | 31.75 $\pm$ 0.55                   |
|                               | MSE   | 0.336 $\pm$ 0.007                    | 0.481 $\pm$ 0.003                   | 0.558 $\pm$ 0.003                   | 60.10 $\pm$ 0.96                   | 43.77 $\pm$ 0.86                   | 37.66 $\pm$ 1.14                   |
|                               | Huber | 0.337 $\pm$ 0.002                    | 0.479 $\pm$ 0.003                   | 0.557 $\pm$ 0.006                   | 60.41 $\pm$ 0.86                   | 44.06 $\pm$ 0.64                   | 37.29 $\pm$ 0.28                   |
|                               | CE    | 0.341 $\pm$ 0.005                    | 0.571 $\pm$ 0.006                   | 0.639 $\pm$ 0.003                   | 60.28 $\pm$ 0.58                   | 45.55 $\pm$ 1.52                   | 41.19 $\pm$ 1.50                   |
| RoBERTa <sub>base</sub>       | MAE   | <b>0.304 <math>\pm</math> 0.010</b>  | <b>0.448 <math>\pm</math> 0.006</b> | <b>0.530 <math>\pm</math> 0.004</b> | 65.12 $\pm$ 1.20                   | 51.81 $\pm$ 1.54                   | 42.99 $\pm$ 1.09                   |
|                               | MSE   | 0.308 $\pm$ 0.012                    | 0.455 $\pm$ 0.001                   | 0.534 $\pm$ 0.005                   | 65.28 $\pm$ 0.36                   | <b>53.94 <math>\pm</math> 1.80</b> | 44.91 $\pm$ 1.31                   |
|                               | Huber | 0.305 $\pm$ 0.007                    | 0.450 $\pm$ 0.004                   | 0.539 $\pm$ 0.003                   | 65.13 $\pm$ 0.60                   | 53.39 $\pm$ 0.19                   | 44.33 $\pm$ 0.64                   |
|                               | CE    | 0.317 $\pm$ 0.003                    | 0.561 $\pm$ 0.003                   | 0.630 $\pm$ 0.005                   | <b>65.48 <math>\pm</math> 0.82</b> | 53.51 $\pm$ 0.39                   | <b>49.55 <math>\pm</math> 0.11</b> |
| DistilRoBERTa <sub>base</sub> | MAE   | 0.318 $\pm$ 0.001                    | 0.463 $\pm$ 0.003                   | 0.539 $\pm$ 0.004                   | 61.30 $\pm$ 0.38                   | 42.75 $\pm$ 1.63                   | 34.96 $\pm$ 1.29                   |
|                               | MSE   | 0.313 $\pm$ 0.004                    | 0.470 $\pm$ 0.005                   | 0.555 $\pm$ 0.003                   | 64.32 $\pm$ 1.99                   | 50.48 $\pm$ 0.29                   | 42.43 $\pm$ 0.22                   |
|                               | Huber | 0.318 $\pm$ 0.004                    | 0.472 $\pm$ 0.001                   | 0.561 $\pm$ 0.004                   | 64.11 $\pm$ 0.72                   | 50.46 $\pm$ 0.73                   | 42.41 $\pm$ 0.30                   |
|                               | CE    | 0.329 $\pm$ 0.002                    | 0.564 $\pm$ 0.005                   | 0.636 $\pm$ 0.003                   | 65.02 $\pm$ 0.69                   | 52.83 $\pm$ 0.71                   | 47.94 $\pm$ 0.88                   |

Table 4.7: Results of experimenting with different pre-trained language models (PLMs) and different loss functions on multi-label classification (JS distance) and on single-label classification (weighted F1-score). The results are the average of three different runs with random starts. Values in bold show the best score for each metric.

errors and like MSE for larger errors. The results of experimenting with the MAE, MSE and Huber losses in Table 4.7 were therefore relatively similar. However, the MAE loss was slightly better and thus better suited for multi-label IDRR. Conversely, the CE loss in Equation 1 focuses on maximizing the probability of the majority label in the reference distribution which makes it better suited for single-label IDRR. Based on these findings, RoBERTa<sub>base</sub> was selected as the encoder backbone of MTask (see Figure 4.3), using MAE loss for multi-label classification and CE loss for single-label classification.

## 4.5.2 Fine-Tuning of Hyperparameters

After selecting RoBERTa<sub>base</sub> as the encoder backbone of MTask and identifying the most effective loss functions for each classification setting, the next experiments involved fine-tuning the model to optimize its performance. Fine-tuning was conducted separately for multi-label and single-label IDRR using different learning rates. Table 4.8 summarizes the results of these experiments. Initial tests showed that learning rates of  $1 \times 10^{-5}$  and  $5 \times 10^{-6}$  consistently led to strong performance in both settings. To further improve model performance, two decay functions were evaluated for these two learning rates: linear decay and cosine annealing with warm restarts (Loshchilov and Hutter, 2017).

| LR        | Decay  | JS Distance $\searrow$ (Multi-Label with MAE) |                                     |                                     | F1-Score $\nearrow$ (Single-Label with CE) |                                    |                                    |
|-----------|--------|-----------------------------------------------|-------------------------------------|-------------------------------------|--------------------------------------------|------------------------------------|------------------------------------|
|           |        | Level-1                                       | Level-2                             | Level-3                             | Level-1                                    | Level-2                            | Level-3                            |
| $1e^{-4}$ | —      | $0.386 \pm 0.004$                             | $0.509 \pm 0.003$                   | $0.569 \pm 0.005$                   | $36.07 \pm 1.67$                           | $21.61 \pm 0.58$                   | $10.05 \pm 0.99$                   |
| $5e^{-5}$ | —      | $0.339 \pm 0.041$                             | $0.478 \pm 0.039$                   | $0.554 \pm 0.016$                   | $64.71 \pm 2.39$                           | $49.76 \pm 1.65$                   | $45.52 \pm 1.59$                   |
| $1e^{-5}$ | —      | $0.304 \pm 0.010$                             | $0.448 \pm 0.006$                   | $0.530 \pm 0.004$                   | <b><math>65.48 \pm 0.82</math></b>         | $53.51 \pm 0.39$                   | $49.55 \pm 0.11$                   |
|           | Linear | $0.305 \pm 0.009$                             | $0.459 \pm 0.007$                   | $0.547 \pm 0.004$                   | $64.70 \pm 0.35$                           | $54.32 \pm 2.15$                   | $48.96 \pm 0.81$                   |
|           | CosAn  | <b><math>0.299 \pm 0.004</math></b>           | <b><math>0.447 \pm 0.008</math></b> | <b><math>0.529 \pm 0.008</math></b> | $65.15 \pm 0.61$                           | $53.99 \pm 1.17$                   | $49.56 \pm 1.25$                   |
| $5e^{-6}$ | —      | $0.303 \pm 0.006$                             | $0.461 \pm 0.005$                   | $0.548 \pm 0.005$                   | $65.45 \pm 0.36$                           | $53.67 \pm 0.45$                   | $49.77 \pm 1.15$                   |
|           | Linear | $0.315 \pm 0.006$                             | $0.480 \pm 0.007$                   | $0.566 \pm 0.007$                   | $64.84 \pm 1.14$                           | <b><math>55.86 \pm 1.10</math></b> | <b><math>50.34 \pm 1.55</math></b> |
|           | CosAn  | $0.306 \pm 0.003$                             | $0.469 \pm 0.002$                   | $0.558 \pm 0.004$                   | $65.40 \pm 1.04$                           | $54.97 \pm 1.85$                   | $49.62 \pm 1.26$                   |
| $1e^{-6}$ | —      | $0.348 \pm 0.003$                             | $0.522 \pm 0.005$                   | $0.613 \pm 0.003$                   | $60.66 \pm 1.38$                           | $47.66 \pm 0.64$                   | $40.96 \pm 0.48$                   |

Table 4.8: Results of experimenting with different learning rates and decay functions on multi-label classification (JS distance) using the MAE loss and on single-label classification (weighted F1-score) using the CE loss. The results are the average of three different runs with random starts. Values in bold show the best score for each metric.

For the linear decay, the learning rate was gradually reduced over the first 5 epochs to half its initial value. For the cosine annealing, the learning rate oscillated between its original value and half that value over two full cycles across 10 epochs. As shown in Table 4.8, the cosine annealing function with a learning rate of  $1 \times 10^{-5}$  achieved the best results on multi-label classification, while the linear decay function with a learning rate of  $5 \times 10^{-6}$  led to the best results on single-label classification. Therefore, these two learning rate configurations were selected for training MTask in the respective classification settings throughout the remainder of this chapter.

## 4.6 Results

Having selected the optimal model configuration for MTask in both the multi-label and single-label classification settings, this section presents the results of the proposed multi-task IDRR model. Section 4.6.1 presents the results of the model on the DiscoGeM corpus in both multi-label and single-label settings, while Section 4.6.2 presents the transfer learning results of the model on different test splits of the PDTB 3.0 corpus in the single-label setting after being trained exclusively on DiscoGeM.

### 4.6.1 Results on DiscoGeM

Table 4.9 presents the performance of the MTask model on the DiscoGeM test set for both multi-label and single-label IDRR. Results in the multi-label setting are reported using Jensen-Shannon (JS) distance, while results in the single-label setting are reported using weighted F1-score. As there is no previous work on full multi-label IDRR, a custom baseline was created for comparison. This baseline emulates the DiscoGeM annotation process by generating ten single-label predictions per relation (similarly to the crowdsourced annotators - see Section 2.2.2) and then averaging them to form a probability distribution. The single-label predictions at each sense level were sampled according to the distribution of majority label senses at the respective level in the DiscoGeM corpus (see Figure 4.1 for level-3 distribution). This simulated distribution is then used as the multi-label baseline in Table 4.9, while in the single-label setting only the majority label sense of the distribution is retained for evaluation.

| Model              | JS Distance $\searrow$ (Multi-Label) |                                     |                                     | F1-Score $\nearrow$ (Single-Label) |                                    |                                    |
|--------------------|--------------------------------------|-------------------------------------|-------------------------------------|------------------------------------|------------------------------------|------------------------------------|
|                    | Level-1                              | Level-2                             | Level-3                             | Level-1                            | Level-2                            | Level-3                            |
| Baseline           | $0.519 \pm 0.002$                    | $0.636 \pm 0.001$                   | $0.714 \pm 0.002$                   | $34.72 \pm 1.63$                   | $22.42 \pm 1.34$                   | $18.81 \pm 0.80$                   |
| Yung et al. (2022) | —                                    | —                                   | —                                   | —                                  | $23.66 \pm 1.19$                   | —                                  |
| Preliminary Study  | —                                    | —                                   | —                                   | —                                  | —                                  | <b><math>51.38 \pm 1.93</math></b> |
| MTask w/ MAE       | <b><math>0.299 \pm 0.002</math></b>  | <b><math>0.446 \pm 0.003</math></b> | <b><math>0.523 \pm 0.002</math></b> | $65.39 \pm 0.54$                   | $50.13 \pm 1.54$                   | $41.60 \pm 0.59$                   |
| MTask w/ CE        | $0.323 \pm 0.003$                    | $0.564 \pm 0.003$                   | $0.634 \pm 0.002$                   | <b><math>65.89 \pm 1.35</math></b> | <b><math>55.99 \pm 1.73</math></b> | <b><math>50.82 \pm 1.26</math></b> |

Table 4.9: Final results on the test set of DiscoGeM on multi-label classification (JS distance) and on single-label classification (weighted F1-score). The results are the average of three different runs with random starts. Values in bold show the best score for each metric.

The results in Table 4.9 show that the MTask model consistently outperforms the baseline across all sense levels in both the multi-label and the single-label settings. These results also reinforce the earlier findings that training the model with the MAE loss function leads to lower JS distances in the multi-label setting, while training with the CE loss function leads to higher F1-scores in the single-label setting. In the single-label setting, the model achieves a substantial improvement over the only previously published IDRR model evaluated on DiscoGeM (Yung et al., 2022), which was only trained and evaluated at level-2. It is also interesting to note that the baseline performance is comparable to the model proposed by Yung et al. (2022) when accounting for standard deviation, which supports the reliability of the baseline as a point of reference. At level-3, MTask achieves similar results to the model used in the preliminary study in Section 4.3. However, it is important to emphasize that the model in the preliminary study was trained to predict only level-3 senses,

whereas MTask is trained to predict all three sense levels simultaneously.

#### 4.6.2 Transfer Learning Results on PDTB 3.0

To explore the potential of transfer learning between discourse annotated corpora, this section evaluates the MTask model on different test splits of the PDTB 3.0 corpus after training it exclusively on DiscoGeM. While the two corpora differ in their annotation methodologies, they are grounded in the same underlying discourse framework, making it possible to investigate whether a model trained on the multi-label annotations of DiscoGeM can generalize effectively on the single-label annotations of PDTB 3.0. This evaluation serves as the first analysis on the potential of transfer learning between the DiscoGeM and the PDTB 3.0 corpora on single-label IDRR. Table 4.10 presents the results of MTask in this transfer learning setting, alongside current SOTA models trained and evaluated on single-label IDRR using the PDTB 3.0 corpus (Long and Webber, 2022; Wu et al., 2023; Zhao et al., 2023; Liu and Strube, 2023; Chan et al., 2023; Jiang et al., 2023a; Zeng et al., 2024; Long and Webber, 2024) as well as the model proposed by Yung et al. (2022) which was trained on both corpora.

| Model                  | F1-Score $\nearrow$ (PDTB 3.0)     |                                    |                                          |                                          |                                    |                                    |
|------------------------|------------------------------------|------------------------------------|------------------------------------------|------------------------------------------|------------------------------------|------------------------------------|
|                        | Lin                                |                                    | Ji                                       |                                          | Cross                              |                                    |
|                        | Level-1                            | Level-2                            | Level-1                                  | Level-2                                  | Level-1                            | Level-2                            |
| Yung et al. (2022)     | —                                  | —                                  | —                                        | 38.07 $\pm$ 2.25                         | —                                  | —                                  |
| Long and Webber (2022) | —                                  | —                                  | 70.05 $\pm$ <i>n/a</i>                   | <b>57.62 <math>\pm</math> <i>n/a</i></b> | —                                  | —                                  |
| Wu et al. (2023)       | —                                  | —                                  | <b>72.07 <math>\pm</math> <i>n/a</i></b> | 50.12 $\pm$ <i>n/a</i>                   | —                                  | —                                  |
| Zhao et al. (2023)     | —                                  | —                                  | 69.06 $\pm$ <i>n/a</i>                   | 52.73 $\pm$ <i>n/a</i>                   | —                                  | —                                  |
| Liu and Strube (2023)  | —                                  | —                                  | 71.15 $\pm$ 0.47                         | 54.92 $\pm$ 0.81                         | <b>70.06 <math>\pm</math> 1.72</b> | <b>55.26 <math>\pm</math> 1.32</b> |
| Chan et al. (2023)     | —                                  | —                                  | 70.84 $\pm$ <i>n/a</i>                   | 49.03 $\pm$ <i>n/a</i>                   | —                                  | —                                  |
| Jiang et al. (2023a)   | —                                  | —                                  | 70.88 $\pm$ <i>n/a</i>                   | 55.30 $\pm$ <i>n/a</i>                   | —                                  | —                                  |
| Zeng et al. (2024)     | —                                  | —                                  | 71.59 $\pm$ <i>n/a</i>                   | 56.50 $\pm$ <i>n/a</i>                   | —                                  | —                                  |
| Long and Webber (2024) | —                                  | —                                  | 71.19 $\pm$ <i>n/a</i>                   | 52.91 $\pm$ <i>n/a</i>                   | —                                  | —                                  |
| MTask w/ MAE           | 47.17 $\pm$ 1.16                   | 27.89 $\pm$ 1.21                   | 43.88 $\pm$ 1.06                         | 26.29 $\pm$ 1.71                         | 45.53 $\pm$ 1.23                   | 29.30 $\pm$ 1.49                   |
| MTask w/ CE            | <b>50.44 <math>\pm</math> 0.81</b> | <b>34.38 <math>\pm</math> 0.86</b> | 49.43 $\pm$ 1.47                         | 33.11 $\pm$ 1.57                         | 49.72 $\pm$ 1.45                   | 33.21 $\pm$ 1.67                   |

Table 4.10: Transfer learning results on different test splits of the PDTB 3.0 on single-label classification (weighted F1-score). The Lin and Ji results are the average of three different runs with random starts, while the Cross results were averaged across all 12 folds. Values in bold show the best score for each level.

As shown in Table 4.10, the MTask model obtains lower results compared to models trained directly on the PDTB 3.0. This outcome is expected, given that MTask was trained exclusively on DiscoGeM and had no exposure to PDTB 3.0 data during training. However,

the performance gap is somewhat unexpected and can be partially attributed to annotation discrepancies between the two corpora. As shown in Table 4.3, the majority sense in DiscoGeM only matches the PDTB 3.0 sense in 30.5% of the overlapping instances. This also explains the performance contrast observed between MTask and the model of Yung et al. (2022) in Tables 4.9 and 4.10. While MTask achieves higher performance when evaluated on DiscoGeM, the model of Yung et al. (2022) performs better when evaluated on PDTB 3.0. Despite the lower performance of MTask in the transfer learning setting, the results offer a valuable insight into the current limitations of cross-corpus generalization in IDRR. This analysis serves as a first step toward understanding the challenges and opportunities in transfer learning between the DiscoGeM and the PDTB 3.0 corpora.

## 4.7 Further Analysis

To provide deeper insights into the performance and reliability of the proposed MTask model, this section presents two complementary analyses. The first focuses on the ability of the model to accurately predict individual senses at level-1 and level-2, providing a fine-grained evaluation beyond aggregate performance metrics. The second examines the coherence of the predictions across sense levels by analyzing how well the predicted level-1 and level-2 senses align with each other based on the PDTB 3.0 sense hierarchy. For easier interpretation, all analyses in this section are conducted using single-label predictions.

### 4.7.1 Individual Sense Results

To better understand the performance of the model across individual senses, this section evaluates the single-label predictions of MTask for each sense at level-1 and level-2. Table 4.11 reports the weighted F1-score for each sense on the test set of DiscoGeM and on different test splits of the PDTB 3.0. The individual sense results on the test set of DiscoGeM follow the distribution of sense labels in the corpus (see Table 4.1). The senses `CONDITION`, `PURPOSE`, `EQUIVALENCE` and `MANNER` never occur as majority labels in DiscoGeM (see Figure 4.1) and are therefore not present in the test set (see Table 4.2) of DiscoGeM at level-2. For the remaining senses that are present in the test set of DiscoGeM, Table 4.11 shows that MTask has difficulties predicting under-represented senses such as `SYNCHRONOUS`, `CONTRAST`, `SIMILARITY` and `SUBSTITUTION`. This suggests that the overall performance of the model could be improved by exploring targeted data augmentation techniques for under-represented senses (Omura et al., 2024; Yung et al., 2025). The performance on the different test splits of the PDTB 3.0 shown in Table 4.11 is lower with

| Level-1     | F1-Score ↗   |              |              |              | Level-2         | F1-Score ↗   |              |               |               |
|-------------|--------------|--------------|--------------|--------------|-----------------|--------------|--------------|---------------|---------------|
|             | DiscoGeM     | PDTB 3.0     |              |              |                 | DiscoGeM     | PDTB 3.0     |               |               |
|             | Test         | Lin          | Ji           | Cross        |                 | Test         | Lin          | Ji            | Cross         |
| TEMPORAL    | 61.27 ± 1.34 | 4.32 ± 1.58  | 15.75 ± 4.50 | 16.23 ± 4.86 | SYNCHRONOUS     | 0.00 ± 0.00  | 0.00 ± 0.00  | 0.00 ± 0.00   | 0.00 ± 0.00   |
|             |              |              |              |              | ASYNCHRONOUS    | 57.87 ± 0.84 | 6.01 ± 2.05  | 22.02 ± 4.42  | 23.16 ± 6.65  |
| CONTINGENCY | 58.67 ± 1.83 | 36.12 ± 7.17 | 34.27 ± 8.24 | 41.08 ± 2.87 | CAUSE           | 65.36 ± 0.12 | 47.21 ± 1.93 | 46.40 ± 2.47  | 48.93 ± 2.07  |
|             |              |              |              |              | CONDITION       | —            | 0.00 ± 0.00  | 0.00 ± 0.00   | 0.00 ± 0.00   |
|             |              |              |              |              | PURPOSE         | —            | 0.00 ± 0.00  | 0.00 ± 0.00   | 0.00 ± 0.00   |
| COMPARISON  | 39.31 ± 3.15 | 32.65 ± 1.69 | 37.67 ± 5.20 | 33.72 ± 2.34 | CONCESSION      | 36.24 ± 4.55 | 29.70 ± 5.09 | 27.92 ± 1.75  | 19.46 ± 3.70  |
|             |              |              |              |              | CONTRAST        | 4.13 ± 2.94  | 11.10 ± 7.92 | 12.70 ± 9.35  | 10.43 ± 6.61  |
|             |              |              |              |              | SIMILARITY      | 0.00 ± 0.00  | 0.00 ± 0.00  | 0.00 ± 0.00   | 0.00 ± 0.00   |
| EXPANSION   | 75.39 ± 1.15 | 64.44 ± 0.66 | 62.69 ± 0.38 | 63.48 ± 1.16 | CONJUNCTION     | 52.56 ± 0.66 | 42.30 ± 1.26 | 39.91 ± 1.97  | 44.45 ± 3.04  |
|             |              |              |              |              | EQUIVALENCE     | —            | 0.00 ± 0.00  | 0.00 ± 0.00   | 0.00 ± 0.00   |
|             |              |              |              |              | INSTANTIATION   | 44.84 ± 0.18 | 12.15 ± 9.12 | 17.90 ± 12.72 | 15.62 ± 10.49 |
|             |              |              |              |              | LEVEL-OF-DETAIL | 46.85 ± 3.28 | 33.61 ± 2.90 | 29.32 ± 2.70  | 29.27 ± 3.06  |
|             |              |              |              |              | MANNER          | —            | 0.00 ± 0.00  | 0.00 ± 0.00   | 0.00 ± 0.00   |
|             |              |              |              |              | SUBSTITUTION    | 0.00 ± 0.00  | 0.00 ± 0.00  | 0.00 ± 0.00   | 0.00 ± 0.00   |

Table 4.11: Individual sense results for each sense in level-1 and level-2 on the test set of DiscoGeM and on different test splits of the PDTB 3.0 on single-label classification (weighted F1-score). Senses marked with “—” were not present in the DiscoGeM test set. The DiscoGeM and the PDTB 3.0 Lin and Ji results are the average of three different runs with random starts, while the PDTB 3.0 Cross results were averaged across all 12 folds.

respect to the performance on the test set of DiscoGeM, which is consistent with the results presented in Tables 4.9 and 4.10 in the Section 4.6. This drop in performance is more pronounced for some senses, such as TEMPORAL at level-1 and ASYNCHRONOUS and INSTANTIATION at level-2. Conversely, the model performs better at the level-2 CONTRAST sense. Since the model was not trained on any PDTB 3.0 data, these discrepancies likely reflect annotation inconsistencies between both corpora as shown by the low agreement in Table 4.3.

To provide further details into the individual sense performance of MTask on the test set of DiscoGeM, Figures 4.4 and 4.5 present confusion matrices of single-label IDRR predictions at level-1 and level-2, respectively. At level-1, Figure 4.4 shows that, with the exception of COMPARISON, the most frequently predicted label for each true label corresponds to the correct label. However, the model is more biased toward the more frequent senses CONTINGENCY and EXPANSION (see Table 4.1). At level-2, the confusion matrix in Figure 4.5 confirms that the model fails to predict under-represented senses such as SYNCHRONOUS, CONTRAST, SIMILARITY and SUBSTITUTION. For the remaining senses, the model often identifies the correct label as the most frequent prediction, indicating stronger performance on better represented senses.

| DiscoGeM (Test): Level-1 |            |               |              |             |
|--------------------------|------------|---------------|--------------|-------------|
| Temporal -               | 67         | 12            | 5            | 27          |
| Contingency -            | 13         | 228           | 17           | 124         |
| Comparison -             | 4          | 26            | 46           | 58          |
| Expansion -              | 19         | 104           | 36           | 581         |
|                          | Temporal - | Contingency - | Comparison - | Expansion - |
| Predicted Labels         |            |               |              |             |

Figure 4.4: Confusion matrix for single-label IDRR at level-1 on the test set of DiscoGeM.

| DiscoGeM (Test): Level-2 |               |                |         |              |            |              |               |                 |                   |                |
|--------------------------|---------------|----------------|---------|--------------|------------|--------------|---------------|-----------------|-------------------|----------------|
| Synchronous -            | 0             | 6              | 1       | 3            | 0          | 0            | 1             | 0               | 2                 | 0              |
| Asynchronous -           | 0             | 62             | 15      | 5            | 0          | 0            | 13            | 1               | 7                 | 0              |
| Cause -                  | 0             | 18             | 357     | 26           | 3          | 0            | 59            | 6               | 46                | 0              |
| Concession -             | 0             | 4              | 41      | 39           | 0          | 0            | 11            | 3               | 6                 | 0              |
| Contrast -               | 0             | 1              | 3       | 14           | 1          | 0            | 3             | 0               | 5                 | 0              |
| Similarity -             | 0             | 0              | 1       | 0            | 0          | 0            | 9             | 0               | 0                 | 0              |
| Conjunction -            | 0             | 12             | 84      | 17           | 4          | 0            | 171           | 3               | 45                | 0              |
| Instantiation -          | 0             | 1              | 5       | 1            | 0          | 0            | 13            | 19              | 14                | 0              |
| Level-of-Detail -        | 0             | 2              | 66      | 3            | 0          | 0            | 30            | 0               | 101               | 0              |
| Substitution -           | 0             | 1              | 2       | 0            | 0          | 0            | 1             | 0               | 0                 | 0              |
|                          | Synchronous - | Asynchronous - | Cause - | Concession - | Contrast - | Similarity - | Conjunction - | Instantiation - | Level-of-Detail - | Substitution - |
| Predicted Labels         |               |                |         |              |            |              |               |                 |                   |                |

Figure 4.5: Confusion matrix for single-label IDRR at level-2 on the test set of DiscoGeM.

### 4.7.2 Sense Coherence across Levels

This section evaluates the coherence of the single-label predictions of MTask across sense levels based on the PDTB 3.0 sense hierarchy. Table 4.12 presents the percentage of times a sense at level-1 was predicted with a coherent sense at level-2 (and vice-versa) on the test set of DiscoGeM and on the different test sets of the PDTB 3.0. As Table 4.12 shows, most senses are often predicted with a coherent sense at the other level. For instance, the CONTINGENCY sense at level-1 is consistently predicted with a coherent level-2 sense (*i.e.*, CAUSE, CONDITION or PURPOSE) across all test sets. Similarly, INSTANTIATION, LEVEL-OF-DETAIL and CONTRAST are very often predicted with a coherent level-1 sense. An interesting exception is the CONJUNCTION level-2 sense. While it is very often predicted with a coherent level-1 sense on most test sets, for the PDTB 3.0 Ji test set it is almost never predicted with a coherent level-1 sense. It is also worth noting that the COMPARISON level-1 sense and the CAUSE level-2 sense have a significant drop in coherent predictions when evaluated on the different test sets of the PDTB 3.0, suggesting possible inconsistencies in the annotations across both corpora. Despite these cases, the overall results in Table 4.12 indicate that the single-label predictions of MTask generally exhibit strong coherence across sense levels. This coherence can be attributed to the shared weights of the linear layer across all classification heads (see Figure 4.3), as well as to the aggregated loss function used during training - which jointly optimizes predictions at each individual level. To further enhance cross-level coherence, a possible extension would be to propagate the output distributions of lower-level classification heads to inform the predictions of higher-level ones - for example, feeding the predictions of classification head 1 into heads 2 and 3, and the predictions of head 2 into head 3.

## 4.8 Discussion

This chapter introduced a novel multi-label approach to IDRR, addressing one of the most challenging tasks in discourse analysis. This approach is based on the multi-task classification model, MTask, which was trained on the DiscoGeM corpus to jointly learn multi-label representations of discourse relations across all three sense levels in the PDTB 3.0 framework. To enable comparisons with prior work, the model was also adapted to the traditional single-label setting by selecting the most probable sense from the predicted multi-label distribution. Extensive experiments were conducted to determine the optimal model configuration and loss functions in the multi-label and single-label setting. The results in Table 4.7 suggest that the representational advantages of using larger PLMs such

| Level-1     | DiscoGeM         | PDTB 3.0          |                   |                   | Level-2         | DiscoGeM          | PDTB 3.0          |                   |                   |
|-------------|------------------|-------------------|-------------------|-------------------|-----------------|-------------------|-------------------|-------------------|-------------------|
|             | Test (%)         | Lin (%)           | Ji (%)            | Cross (%)         |                 | Test (%)          | Lin (%)           | Ji (%)            | Cross (%)         |
| TEMPORAL    | 92.94 $\pm$ 2.93 | 93.27 $\pm$ 6.31  | 90.93 $\pm$ 4.51  | 92.48 $\pm$ 4.93  | SYNCHRONOUS     | <i>n/a</i>        | <i>n/a</i>        | <i>n/a</i>        | <i>n/a</i>        |
|             |                  |                   |                   |                   | ASYNCHRONOUS    | 90.00 $\pm$ 1.44  | 82.74 $\pm$ 6.76  | 82.44 $\pm$ 11.34 | 88.30 $\pm$ 8.08  |
| CONTINGENCY | 98.91 $\pm$ 0.20 | 100.00 $\pm$ 0.00 | 99.74 $\pm$ 0.45  | 100.00 $\pm$ 0.00 | CAUSE           | 62.37 $\pm$ 4.92  | 38.19 $\pm$ 6.45  | 38.42 $\pm$ 8.27  | 29.93 $\pm$ 1.32  |
|             |                  |                   |                   |                   | CONDITION       | —                 | <i>n/a</i>        | <i>n/a</i>        | <i>n/a</i>        |
|             |                  |                   |                   |                   | PURPOSE         | —                 | <i>n/a</i>        | <i>n/a</i>        | <i>n/a</i>        |
| COMPARISON  | 86.31 $\pm$ 1.22 | 73.81 $\pm$ 13.21 | 68.09 $\pm$ 10.47 | 58.45 $\pm$ 4.12  | CONCESSION      | 81.19 $\pm$ 2.83  | 89.15 $\pm$ 8.10  | 89.36 $\pm$ 8.98  | 97.93 $\pm$ 1.52  |
|             |                  |                   |                   |                   | CONTRAST        | 91.67 $\pm$ 14.43 | 90.77 $\pm$ 10.09 | 90.77 $\pm$ 11.11 | 92.68 $\pm$ 8.19  |
|             |                  |                   |                   |                   | SIMILARITY      | <i>n/a</i>        | <i>n/a</i>        | <i>n/a</i>        | <i>n/a</i>        |
| EXPANSION   | 70.70 $\pm$ 1.26 | 37.91 $\pm$ 10.82 | 65.02 $\pm$ 12.30 | 53.01 $\pm$ 2.21  | CONJUNCTION     | 97.61 $\pm$ 0.71  | 97.14 $\pm$ 0.03  | 3.20 $\pm$ 0.88   | 96.69 $\pm$ 0.96  |
|             |                  |                   |                   |                   | EQUIVALENCE     | —                 | <i>n/a</i>        | <i>n/a</i>        | <i>n/a</i>        |
|             |                  |                   |                   |                   | INSTANTIATION   | 100.00 $\pm$ 0.00 | 100.00 $\pm$ 0.00 | 100.00 $\pm$ 0.00 | 100.00 $\pm$ 0.00 |
|             |                  |                   |                   |                   | LEVEL-OF-DETAIL | 99.82 $\pm$ 0.31  | 99.80 $\pm$ 0.35  | 100.00 $\pm$ 0.00 | 100.00 $\pm$ 0.00 |
|             |                  |                   |                   |                   | MANNER          | —                 | <i>n/a</i>        | <i>n/a</i>        | <i>n/a</i>        |
|             |                  |                   |                   |                   | SUBSTITUTION    | <i>n/a</i>        | <i>n/a</i>        | <i>n/a</i>        | <i>n/a</i>        |

Table 4.12: Percentage of the instances that were classified with coherent senses at level-1 and level-2 on the test set of DiscoGeM and on different test splits of the PDTB 3.0 on single-label classification. Senses marked with “—” were not present in the DiscoGeM test set and senses marked with “*n/a*” were never predicted by the model. The DiscoGeM and the PDTB 3.0 Lin and Ji results are the average of three different runs with random starts, while the PDTB 3.0 Cross results were averaged across all 12 folds.

as RoBERTa<sub>base</sub> outweigh the benefits previously associated with PLMs trained using the NSP objective (Shi and Demberg, 2019). The results also show that the choice of loss function plays an important role in IDRR. Models trained with MAE, MSE and the Huber loss function consistently achieve lower JS distances in the multi-label setting, while model trained with CE loss function achieve higher F1-scores in the single-label setting.

Based on these findings, MTask establishes the first benchmark for multi-label IDRR and achieves SOTA performance on single-label IDRR using DiscoGeM, as shown in Table 4.9. Further analysis of the individual sense results in Table 4.11 for single-label IDRR using DiscoGeM reveals that the performance of MTask is directly correlated with the frequency of each sense in the training data. For instance, MTask has difficulties predicting under-represented senses such as SYNCHRONOUS, CONTRAST, SIMILARITY and SUBSTITUTION (see Table 4.1). This trend is further illustrated in the confusion matrices in Figures 4.4 and 4.5, where the model is more biased towards better represented senses such as the level-1 CONTINGENCY and EXPANSION causes and the level-2 CAUSE, CONJUNCTION and LEVEL-OF-DETAIL senses. These findings suggest that overall performance of the model could be improved from targeted data augmentation techniques for under-represented senses (Omura et al., 2024; Yung et al., 2025). Conversely, the results in Table 4.12 indicate that the single-label predictions of MTask are generally coherent

across sense levels. This coherence is likely supported by the shared weights of the linear layer across all classification heads of MTask (see Figure 4.3), as well as to the aggregated loss function used during training - which jointly optimizes predictions at each individual level. To further enhance cross-level coherence, a possible improvement in model design would be to propagate the output distributions of lower-level classification heads to inform the predictions of higher-level ones - for example, feeding the predictions of classification head 1 into heads 2 and 3, and the predictions of head 2 into head 3.

Although the model was trained exclusively on the DiscoGeM corpus, it was also evaluated on the PDTB 3.0 corpus in the single-label setting through direct transfer learning. This evaluation was conducted to present a first analysis on the potential of transfer learning between both discourse annotated corpora. However, as shown in Table 4.10, the performance gap between MTask and existing SOTA models trained directly on PDTB 3.0 suggests that effective transfer learning across these corpora remains a challenge. A contributing factor to this limitation is the mismatch in label sets. While single-label IDRR in the PDTB 3.0 is typically evaluated using the 14-label set of level-2 senses proposed by Kim et al. (2020), the CAUSE+BELIEF sense from this label set is not annotated in DiscoGeM and had to be replaced by the SIMILARITY sense to maintain a 14-label set (see Section 4.2.1). Ideally, it would have been replaced by another sense under the same CONTINGENCY level-1 sense, but all the other senses were already included in the label set. This small discrepancy in the annotation of senses poses a limitation in cross-corpora training.

Future research should explore targeted data augmentation techniques to improve model performance on under-represented senses in the DiscoGeM corpus. Another promising direction involves investigating whether the coherence of the predictions of the model can be further improved by cascading information from higher-level to lower-level classification heads within the model architecture. Additionally, to more effectively evaluate the potential of cross-corpus transfer learning for single-label IDRR, future work should consider intermediate fine-tuning strategies, such as fine-tuning a multi-label model trained on DiscoGeM using the single-label annotations of PDTB 3.0. Given recent advancements in LLMs and their increasingly strong capabilities in natural language understanding, further research should also explore the applicability of prompt-engineering techniques in both multi-label and single-label IDRR. Finally, it is important to acknowledge that the extensive experimental analysis presented in this chapter would not have been possible without access to high-performance computing resources. While this enabled the current study, the associated non-negligible carbon footprint was not monitored and should be taken into consideration in future research.

## Chapter 5

# Multilingual Implicit Discourse Relation Recognition

This chapter improves on the multi-label approach to IDRR presented in Chapter 4 and extends it beyond the English language to address the third research objective of this thesis introduced in Chapter 1: **to develop a multilingual IDRR classification model capable of capturing multiple implicit senses across different languages**. Although the previous approach established a first benchmark for multi-label IDRR, it was limited to the English language alone. This linguistic constraint brings into focus the need for models that can generalize across languages while capturing the ambiguity of implicit discourse relations. Thus, this chapter proposes a multilingual and multi-label IDRR classification model designed to explicitly leverage the hierarchical dependencies between senses in the PDTB 3.0 framework. This chapter is based on the paper published at the Special Interest Group on Discourse and Dialogue (SIGDIAL) 2025 (Costa and Kosseim, 2025b).

Despite significant advances of NLP models, most of the research in this field continues to focus exclusively on the English language (Joshi et al., 2020; Faisal et al., 2022; Ranathunga and de Silva, 2022). This linguistic focus overlooks the remarkable diversity of the roughly 7,000 languages in the world (Levisen, 2019; Blasi et al., 2022). To develop models that more accurately capture the complexity of human reasoning, it is essential to expand NLP research beyond the English language and build more inclusive models capable of generalizing across multiple languages.

In recent years, some progress has been made in multilingual computational discourse analysis, particularly in discourse parsing (Braud et al., 2017; Sluyter-Gäthje et al., 2020; Liu et al., 2020, 2021; Bourgonje and Lin, 2024). However, research on IDRR across languages remains sparse. Evaluation of multilingual data has shown that the interpretation of

implicit discourse relations may vary between translations of the same text (Zeyrek et al., 2022; Tomaszewska et al., 2024; Zufferey and Degand, 2024), which amplifies the complexity of the already ambiguous IDRR task (Hoek and Zufferey, 2015; Kurfalı and Östling, 2019; Lapshinova-Koltunski and Carl, 2022; Yung et al., 2023). The DISRPT shared task (Zeldes et al., 2021; Braud et al., 2023) marked an important step toward multilingual discourse classification, introducing datasets and benchmarks in multiple languages. However, participating models in the shared tasks have been restricted to single-label classification and evaluated on each language independently (Liu et al., 2023b; Metheniti et al., 2023; Anuranjana, 2023). As a result, the development of models capable of capturing multiple implicit senses across languages remains an unexplored area of research.

This research gap is largely due to the scarcity of multilingual discourse annotated corpora. While some individual corpora exist for languages other than English, until the recent release of DiscoGeM 2.0 only the TED-MDB corpus (Zeyrek et al., 2020) provided parallel multilingual annotation of discourse relations. However, the corpus is relatively small as it only contains 1,286 implicit discourse relations distributed between 6 different languages (roughly 200 relations per language) - all of which are annotated with single-labels. Thus, to address the lack of linguistic diversity in IDRR, this chapter introduces a hierarchical multi-label classification model, HArch, capable of learning multi-label representation of implicit discourse relations across the available languages in the DiscoGeM 2.0 corpus: English, German, French and Czech. In line with the recent trend in NLP, this chapter also explores a few-shot prompting framework to leverage LLMs on multilingual and multi-label IDRR.

## 5.1 Previous Work

Research on multilingual IDRR remains limited, with existing work stemming from the two most recent editions of the DISRPT shared task (Zeldes et al., 2021; Braud et al., 2023). In the discourse relation classification track of this shared task, participants have to develop models capable of assigning a single label to discourse relations across a variety of corpora. In total, the task includes 26 corpora annotated following 4 different frameworks and covering 13 languages. Models are evaluated individually on each corpus and a mean score is then calculated for each model. For the 9 corpora following the PDTB 3.0 framework, only level-2 senses are considered and these are the only corpora for which performance on explicit and implicit discourse relations is presented separately. For all other corpora, results on both relation types are aggregated. Although none of the models were evaluated on a unified multilingual corpus, their performance on 12 languages other

than English represents an important first step toward multilingual IDRR. Below is a review of the models that participated in the most recent published edition of the DISRPT shared task (Liu et al., 2023b; Metheniti et al., 2023; Anuranjana, 2023), along with a recent IDRR model developed for the low-resource Nigerian Pidgin language (Saeed et al., 2025) based on the DiscoNaija corpus (Scholman et al., 2025).

- Liu et al. (2023b) proposed multiple models to the discourse classification track of DISRPT 2023. The models consisted of a PLM encoder followed by a linear classification layer and they were fine-tuned individually on each corpus of the shared task. The models were trained following an adversarial strategy (Miyato et al., 2017) and the choice of the PLM encoder depended on the language of the corpus. For corpora in English they used RoBERTa<sub>large</sub>, in Chinese they used MacBERT<sub>large</sub> (Cui et al., 2020), in French they used CamemBERT<sub>large</sub> (Martin et al., 2020), in Basque they used BERTeUS<sub>large</sub> (Agerri et al., 2020) and for all the other languages they used XLM-RoBERTa<sub>large</sub> (Conneau et al., 2020). Their proposed methodology won the shared task having achieved a mean accuracy of 62.36% on the test sets.
- Metheniti et al. (2023) proposed a model based on a single PLM encoder to the discourse classification track of DISRPT 2023. Their model consisted of a multilingual mBERT<sub>base</sub> (Devlin et al., 2019) PLM encoder followed by a linear classification layer. The model was trained on each language available in the shared task following an adapters methodology (Houlsby et al., 2019), which exposes only a subset of parameters in transformer-based models to fine-tuning. Their proposed methodology finished second on the shared task having achieved a mean accuracy of 54.44% on the test sets.
- Anuranjana (2023) proposed a model based on Flan-T5 (Chung et al., 2024) to the discourse classification track of DISRPT 2023. Their model was prompt-tuned on each language available in the shared task following a series of instruction based prompts. A post-processing step tried to match the output of the model to an existing label or select the majority label in the respective corpus if the output could not be mapped. Their proposed methodology finished last on the shared task having achieved a mean accuracy of 31.21% on the test sets.
- Saeed et al. (2025) presented an IDRR classification model for the Nigerian Pidgin language following the PDTB 2.0 sense hierarchy. Nigerian Pidgin has nearly 100 million speakers, but despite this, it has very little support in terms of NLP resources and corpora. They explored different methods and found that fine-tuning the

DiscoPrompt model (Chan et al., 2023) (see Section 4.1) on the DiscoNaija corpus (Scholman et al., 2025) led to the best results. They achieved an F1-score of 46.1 at level-1 and 32.7 at level-2 on single-label classification.

These models represent an important first step toward multilingual IDRR and, with the exception of Anuranjana (2023) which adopts a prompt-tuning strategy, the majority are based on fine-tuning encoder-based models. In parallel, the increasing natural language understanding capabilities of LLMs have motivated recent studies to explore their effectiveness on single-label IDRR through prompt-engineering strategies (Chan et al., 2024; Yung et al., 2024a). However, the evaluation of LLMs for IDRR remains limited to the English language and to the single-label classification setting. More precisely:

- Chan et al. (2024) evaluated the performance of ChatGPT on IDRR using three different prompting methods. The simplest one formulates the classification task as a multiple-choice question-answering problem, the second one uses a manually designed prompt template based on expert knowledge and the third one uses in-context learning. They found that the latter approach achieved the best results, yet it still struggled in the IDRR task - achieving results far below the fine-tuned smaller encoder-based SOTA models (see Section 4.1). Their best model achieved an F1-score of 36.11 at level-1 and 16.20 at level-2 on single-label classification on the PDTB 2.0 corpus.
- Yung et al. (2024a) replicated the methodology of Chan et al. (2024) to evaluate the performance of GPT-4 (OpenAI, 2023) on IDRR and explored the benefits of decomposing the classification problem into multiple prompts each containing a binary choice between sense labels. Despite their efforts, their results also show that prompt-engineering strategies using LLMs for IDRR still fall short of fine-tuned smaller encoder-based SOTA models. Their best model achieved an F1-score of 47.53 at level-1 and 25.77 at level-2 on single-label classification on the PDTB 3.0 corpus. They also performed experiments at a smaller scale using LLaMa (Touvron et al., 2023), but found the performance was substantially worse.

It is worth noting that both of the prompt-engineering approaches above perform worse than the MTask model described in Chapter 4, which achieved an F1-score of  $49.43 \pm 1.47$  at level-1 and  $33.11 \pm 1.57$  at level-2 on single-label classification on the PDTB 3.0 corpus despite being trained exclusively on DiscoGeM and evaluated in a transfer learning setting (see Section 4.6.2). These findings are consistent with recent work showing that fine-tuning smaller encoder-based model still outperforms prompting LLMs in text classification task

when moderate amounts of training data are available (Edwards and Camacho-Collados, 2024; Bucher and Martini, 2024). This chapter will evaluate if the same observations also hold for multilingual and multi-label IDRR by comparing a fine-tuned hierarchical classification model, HArch, against recent LLMs via few-shot prompting across different languages.

## 5.2 Data Preparation

To address multilingual and multi-label IDRR, the work presented in this chapter will mostly be conducted on the DiscoGeM 2.0 corpus. However, the DiscoGeM corpus will also be used to compare the hierarchical approach presented in this chapter against the approach presented in Chapter 4 for multi-label IDRR in the English language. As described in Section 2.2.2, the DiscoGeM corpus consists of 6,807 implicit discourse relations annotated with multiple senses in English, while the DiscoGeM 2.0 corpus consists of 13,063 parallel implicit discourse relations annotated with multiple senses across four languages: 5,847 in English, 2,588 in German, 2,628 in French and 2,000 in Czech (see Section 2.2.3). While Chapter 4 only considered an adapted set of 14 level-2 senses in the DiscoGeM corpus to enable transfer learning with the PDTB 3.0 corpus (see Section 4.2.1), this chapter retains all of the level-2 senses in the DiscoGeM 2.0 corpus. Table 5.1 shows the distribution of level-1 senses and Table 5.2 shows the distribution of level-2 and level-3 senses per language in the DiscoGeM 2.0 corpus. Note that the level-2 senses SYNCHRONOUS, CONTRAST, SIMILARITY, CONJUNCTION, DISJUNCTION and EQUIVALENCE do not have corresponding level-3 senses. In order to conduct IDRR at level-3, their level-3 sense label was defaulted to their corresponding level-2 label - similar to what was done in Chapter 4. Finally, the DiscoGeM 2.0 corpus provides a suggested split of its data for training, validation and testing. To ensure reproducibility of the results presented in this chapter, the suggested splits are adopted to train and evaluate the proposed models.

## 5.3 Methodology

Although significant progress has been done in the field of IDRR, most existing approaches treat discourse sense labels as flat and independent (see Section 4.1). This neglects the rich structural information embedded in discourse frameworks like the PDTB 3.0. The MTask model described in Chapter 4 (see Figure 4.3) included a shared linear layer across

| Level-1     | English | German  | French  | Czech   | All     |
|-------------|---------|---------|---------|---------|---------|
| Temporal    | 556.8   | 438.4   | 400.7   | 458.8   | 1,854.6 |
| Contingency | 1,695.4 | 772.3   | 811.0   | 681.1   | 3,959.8 |
| Comparison  | 796.3   | 346.5   | 525.0   | 227.1   | 1,894.9 |
| Expansion   | 2,798.6 | 1,030.8 | 891.3   | 632.9   | 5,353.7 |
| Total       | 5,847.1 | 2,588.0 | 2,628.0 | 1,999.9 | 13,063  |

Table 5.1: Distribution of level-1 senses per language in the DiscoGeM 2.0 corpus. Each value represents the sum of the corresponding sense across all instances in the specific language set of the corpus.

| Level-2         | English | German | French | Czech | All     | Level-3              | English | German | French | Czech | All     |
|-----------------|---------|--------|--------|-------|---------|----------------------|---------|--------|--------|-------|---------|
| SYNCHRONOUS     | 99.4    | 191.2  | 139.8  | 254.6 | 685.0   | SYNCHRONOUS*         | 99.4    | 191.2  | 139.8  | 254.6 | 685.0   |
| ASYNCHRONOUS    | 457.4   | 247.1  | 260.9  | 204.2 | 1169.6  | PRECEDENCE           | 420.7   | 222.7  | 240.0  | 193.7 | 1,077.1 |
|                 |         |        |        |       |         | SUCCESSION           | 36.7    | 24.4   | 20.9   | 10.5  | 92.5    |
| CAUSE           | 1,690.7 | 457.5  | 608.6  | 576.7 | 3,333.5 | REASON               | 386.9   | 243.9  | 237.6  | 281.3 | 1,149.8 |
|                 |         |        |        |       |         | RESULT               | 1,303.8 | 213.6  | 371.0  | 295.4 | 2,183.7 |
| CONDITION       | 1.2     | 126.1  | 93.2   | 42.0  | 262.5   | ARG1-AS-COND         | 0.1     | 109.4  | 87.9   | 25.1  | 222.4   |
|                 |         |        |        |       |         | ARG2-AS-COND         | 1.1     | 16.7   | 5.4    | 16.9  | 40.1    |
| NEG-CONDITION   | 1.1     | 14.0   | 14.0   | 12.9  | 42.0    | ARG1-AS-NEGCOND      | 1.0     | 7.8    | 10.9   | 10.9  | 30.6    |
|                 |         |        |        |       |         | ARG2-AS-NEGCOND      | 0.1     | 6.2    | 3.1    | 2.0   | 11.5    |
| PURPOSE         | 2.3     | 174.8  | 95.2   | 49.5  | 321.8   | ARG1-AS-GOAL         | 1.4     | 114.8  | 91.6   | 46.5  | 254.2   |
|                 |         |        |        |       |         | ARG2-AS-GOAL         | 0.9     | 60.0   | 3.6    | 3.1   | 67.6    |
| CONCESSION      | 505.1   | 161.3  | 261.3  | 160.9 | 1088.6  | ARG1-AS-DENIER       | 167.5   | 61.9   | 91.3   | 39.9  | 360.5   |
|                 |         |        |        |       |         | ARG2-AS-DENIER       | 337.7   | 99.4   | 169.9  | 121.0 | 728.1   |
| CONTRAST        | 187.4   | 98.7   | 173.3  | 40.8  | 500.2   | CONTRAST*            | 187.4   | 98.7   | 173.3  | 40.8  | 500.2   |
| SIMILARITY      | 103.7   | 86.5   | 90.4   | 25.5  | 306.1   | SIMILARITY*          | 103.7   | 86.5   | 90.4   | 25.5  | 306.1   |
| CONJUNCTION     | 1,298.8 | 260.0  | 121.8  | 155.9 | 1,836.5 | CONJUNCTION*         | 1,298.8 | 260.0  | 121.8  | 155.9 | 1,836.5 |
| DISJUNCTION     | 2.7     | 6.5    | 3.0    | 3.1   | 15.3    | DISJUNCTION*         | 2.7     | 6.5    | 3.0    | 3.1   | 15.3    |
| EQUIVALENCE     | 19.0    | 154.7  | 101.5  | 57.4  | 332.6   | EQUIVALENCE*         | 19.0    | 154.7  | 101.5  | 57.4  | 332.6   |
| EXCEPTION       | 1.8     | 26.8   | 41.0   | 6.8   | 76.4    | ARG1-AS-EXCEPTION    | 0.3     | 14.7   | 30.7   | 2.8   | 48.5    |
|                 |         |        |        |       |         | ARG2-AS-EXCEPTION    | 1.4     | 12.2   | 10.3   | 4.0   | 27.8    |
| INSTANTIATION   | 352.4   | 99.3   | 124.2  | 84.4  | 660.3   | ARG1-AS-INSTANCE     | 16.5    | 0.0    | 35.6   | 25.9  | 78.0    |
|                 |         |        |        |       |         | ARG2-AS-INSTANCE     | 335.9   | 99.3   | 88.6   | 58.5  | 582.3   |
| LEVEL-OF-DETAIL | 1079.1  | 393.2  | 363.1  | 278.8 | 2114.1  | ARG1-AS-DETAIL       | 154.9   | 88.7   | 51.8   | 110.1 | 405.6   |
|                 |         |        |        |       |         | ARG2-AS-DETAIL       | 924.1   | 304.5  | 311.3  | 168.6 | 1708.5  |
| MANNER          | 4.5     | 33.7   | 124.0  | 37.2  | 199.4   | ARG1-AS-MANNER       | 1.3     | 26.6   | 112.3  | 22.7  | 163.0   |
|                 |         |        |        |       |         | ARG2-AS-MANNER       | 3.1     | 7.0    | 11.7   | 14.6  | 36.5    |
| SUBSTITUTION    | 40.5    | 56.6   | 12.7   | 9.3   | 119.1   | ARG1-AS-SUBSTITUTION | 0.0     | 7.6    | 6.0    | 1.3   | 14.9    |
|                 |         |        |        |       |         | ARG2-AS-SUBSTITUTION | 40.5    | 49.0   | 6.7    | 8.0   | 104.1   |

Table 5.2: Distribution of level-2 and level-3 senses per language in the DiscoGeM 2.0 corpus. Each value represents the sum of the corresponding sense across all instances in the specific language set of the corpus. Level-3 senses that did not have a sense label (marked with \*) were defaulted to their level-2 sense as replacement.

all of the classification heads to learn the dependencies between sense levels through indirect parameter sharing. However, further analysis of the performance of the model (see Section 4.7.2) suggested that a more explicit representation of hierarchical dependencies could further improve its performance. Therefore, this section presents a hierarchical multi-task IDRR classification model, HArch, fine-tuned to learn multi-label representations of implicit discourse relations in a multilingual setting across all three sense levels in the PDTB 3.0 framework. In line with recent work on single-label IDRR (see Section 5.1), this section also presents a few-shot prompting framework to leverage LLMs on multilingual and multi-label IDRR.

### 5.3.1 Fine-Tuned Classification Model

The HArch model improves on MTask by explicitly modeling hierarchical dependencies across sense levels. Figure 5.1 presents the architecture of the model. The model takes as input a concatenated pair of discourse arguments, which is encoded using a PLM. The resulting contextual representation is passed through a shared linear and dropout layer. This shared representation is then passed to three separate classification heads - one for each sense level. The level-1 classification head predicts a probability distribution over the 4 possible level-1 senses using a linear layer followed by a softmax activation - as shown in red on the bottom-right side of Figure 5.1. This 4-dimensional output vector is then projected onto the same embedding space of the PLM through an augmentation block. As shown in blue on the upper-right side of Figure 5.1, this augmentation is done in two steps: the output of the classification head is first mapped to a hidden dimension of half the size of the embedding space and then, in a second step, further projected to match the full embedding size. A GELU (Hendrycks and Gimpel, 2016) activation function after each incremental step introduces non-linearity, while helping with sparsity and preventing exploding activations in the dimensionality projection. Based on early experiments, this incremental dimensional projection ensured a smoother transition in feature space. The augmented level-1 output is then combined with the shared representation through a weighted sum defined by Equation 5, where  $\alpha$  is a learnable parameter of the model, to yield the input of the level-2 classification head.

$$\mathbf{W}_\alpha = \alpha \cdot \mathbf{h}_{\text{Aug}_1} + (1 - \alpha) \cdot \mathbf{h}_{\text{Dropout}} \quad (5)$$

The level-2 classification head then predicts a probability distribution over the 17 possible level-2 senses. Similar to level-1, the level-2 output is passed through a dimensionality

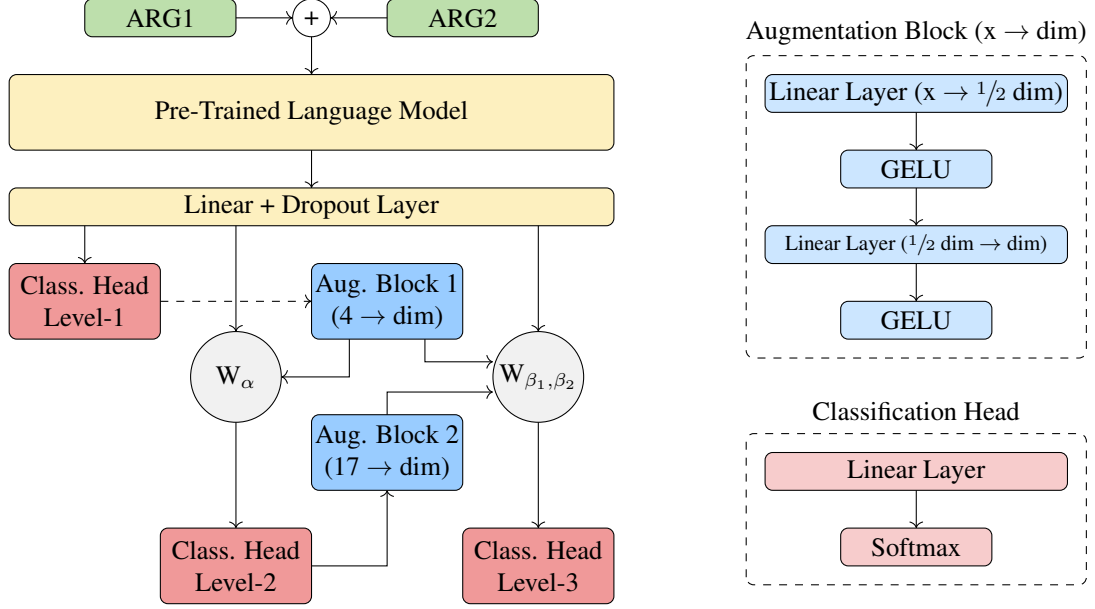


Figure 5.1: Architecture of the hierarchical multi-task model, HArch, for multilingual and multi-label IDRR. Given a pair of discourse arguments as input, each classification head of the model generates a probability distribution over the senses at the corresponding sense level in the PDTB 3.0 framework. The predictions of the lower-level classification heads are projected onto the embedding space of the encoder via augmentation blocks and then combined with the shared representation using learnable weighted sums.

augmentation block to project it into the same embedding space of the PLM. The augmented level-1 and level-2 outputs are then added to the shared representation through a second weighted sum defined by Equation 6, where  $\beta_1$  and  $\beta_2$  are learnable parameters. The output of this weighted sum is then fed into the level-3 classification head, which predicts a probability distribution over the level-3 senses.

$$\mathbf{W}_{\beta_1, \beta_2} = \beta_1 \cdot \mathbf{h}_{\text{Aug}_1} + \beta_2 \cdot \mathbf{h}_{\text{Aug}_2} + (1 - \beta_1 - \beta_2) \cdot \mathbf{h}_{\text{Dropout}} \quad (6)$$

This cascading of information ensures that the predictions of lower-level classification heads are used to help inform the prediction of higher-level classification heads that require more fine-grained information. The model is trained using the Adam optimizer (Kingma and Ba, 2015) and the loss function is calculated as the sum of the individual classification losses of each head. Based on previous experiments with different loss functions (see Section 4.5.1), the loss of each classification head is computed using the mean absolute error (MAE) loss function. Similar to the evaluation of MTask in the multi-label setting, HArch is evaluated using the Jensen-Shannon (JS) distance (Lin, 1991) to measure the similarity between the predicted and reference sense probability distributions.

### 5.3.2 LLM Prompting Framework

The prompting framework for multilingual and multi-label IDRR adopts a few-shot role-playing strategy. Figure 5.2 presents the few-shot conversational prompt template used when evaluating the framework in the English language. When evaluating the framework in the other languages individually, the prompt was translated into the corresponding language and the few-shot examples were drawn from the corresponding language subset of the DiscoGeM 2.0 corpus. In the multilingual evaluation setting, where the framework was evaluated on all of the languages simultaneously, the main prompt text remained in English and the examples were drawn from all languages. Due to space constraints, Figure 5.2 only shows two examples of discourse relations in the initial *User Message* prompt as well as only two subsequent *User* and *Assistant* interactions. In practice, five examples were used in the initial *User Message* and it was followed by five *User* and *Assistant* interactions in all language settings. All examples were taken from the training split of the corpus.

The choice of a role-playing prompting strategy was driven by the need for consistent output formatting. While techniques such as chain-of-thought prompting may enhance reasoning in tasks requiring intermediate logical steps, they often produce verbose and unstructured outputs. In contrast, the role-playing strategy combined with few-shot examples, reinforces the need for a consistent output formatting to facilitate the evaluation of the predictions. Meanwhile, the prompt itself was formulated to closely replicate the annotation methodology of DiscoGeM 2.0. The annotators of the corpus were provided with a list of connectives that represented each of the possible senses in the PDTB 3.0 framework (see Section 2.2.3), from which they had to choose the connective that best expressed the semantic relation between the arguments of the relation being annotated - irrespective of syntactic adjustments. The annotations were then averaged to provide a distribution over the possible senses which corresponds to the desired output of the prompting framework for each relation. The prompting framework was evaluated using GPT-4o (OpenAI, 2024) and Llama-4-Maverick on the test split of DiscoGeM 2.0 and it was allowed up to 5 retries per relation when prompt outputs did not match the expected format. To ensure consistency, the prompting temperature was set to 0. Prompting costs were approximately 5.00\$ for GPT-4o and 0.49\$ for Llama-4-Maverick per million input tokens.

## 5.4 Experiments

This section evaluates the performance of different PLMs as encoders of the HArch model (see Figure 5.1) for multi-label IDRR in both multilingual and English-only settings

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><b>Prompt template in English</b></p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| <p><b>System Message:</b></p> <p>You're an expert linguist with the task of identifying the best discourse connectives to fill in the gap between two arguments. Below you have the ordered list of 28 discourse connectives you can use.</p> <p>["at the same time", "then", "after", "because", "as a result", "in that case", "if", "if not", "unless", "for that purpose", "so that", "even though", "nonetheless", "on the other hand", "similarly", "also", "or", "in other words", "other than that", "an exception is that", "this illustrates that", "for example", "in short", "in more detail", "thereby", "as if", "rather than", "instead"]</p> <p>For each pair of arguments given to you, you have to choose the connectives from this list that could be used to connect the two arguments. Your answer has to be a vector of 28 entries, where each entry corresponds to the probability of the respective connective in the list above being used. The sum of all probabilities in the vector must be equal to 1.</p> |
| <p><b>User Message:</b></p> <p>Here are a few examples.</p> <p>Argument 1: There, comrades, is the answer to all our problems. It is summed up in a single word - Man.<br/>Argument 2: Man is the only real enemy we have. Remove Man from the scene, and the root cause of hunger and overwork is abolished for ever.</p> <p>Answer: [0.0, 0.1, 0.0, 0.3, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.1, 0.0, 0.0, 0.0, 0.0, 0.3, 0.2, 0.0, 0.0, 0.0, 0.0]</p> <p>... (x3)</p> <p>Argument 1: It had been a good one. There had been a flying motorcycle in it.<br/>Argument 2: He had a funny feeling he'd had the same dream before. His aunt was back outside the door.</p> <p>Answer: [0.0, 0.0, 0.0, 0.0, 0.111, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.111, 0.0, 0.0, 0.667, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.111, 0.0, 0.0, 0.0, 0.0]</p>                                                                                                                                                                        |
| <p><b>Assistant Message:</b></p> <p>Got it! Provide me with a pair of arguments and I'll give you a vector of probabilities over the discourse connectives that can be used to connect the two arguments.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
| <p><b>User Message:</b></p> <p>Argument 1: Everybody seemed to hear a rustling outside the door. There was no sound of footsteps.<br/>Argument 2: It was like light silk sliding over the panel. Then it stopped.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| <p><b>Assistant Message:</b></p> <p>[0.0, 0.0, 0.0, 0.0, 0.1, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.1, 0.1, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.2, 0.0, 0.0, 0.0, 0.5]</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| <p><b>User Message:</b></p> <p>Argument 1: It pretends to respect human rights. It pretends to persecute no one.<br/>Argument 2: It pretends to fear nothing. It pretends to pretend nothing.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| <p><b>Assistant Message:</b></p> <p>[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.3, 0.7, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |

Figure 5.2: Conversational prompt template of the few-shot prompting framework for multi-label IDRR in the English language.

on the validation split of the DiscoGeM 2.0 corpus. Table 5.3 presents the results of these experiments. All of the results represent the average performance across three independent runs, each initialized with a different random seed. All models were trained for 10 epochs with a batch size of 16 and following the optimal hyperparameter configuration previously found for MTask in the multi-label classification setting (see Section 4.5.1): learning rate of  $1 \times 10^{-5}$  with a cosine annealing decay function. In addition, results obtained using LLMs with the proposed few-shot prompting framework are also presented.

In the English-only setting, HArch was evaluated using three different PLM encoders: RoBERTa<sub>base</sub>, the recently released ModernBERT<sub>base</sub> (Warner et al., 2024) and the multilingual XLM-RoBERTa<sub>base</sub> (Conneau et al., 2020). Despite ModernBERT claiming superior performance over the BERT family of models, ModernBERT-HArch was consistently the worst performing fine-tuned model across all three sense levels in Table 5.3. Even XLM-RoBERTa-HArch, despite its multilingual background, outperformed ModernBERT-HArch in terms of JS distance. Similar to the previous experiments with MTask (see Section 4.5.1), the HArch model employing RoBERTa<sub>base</sub> as its PLM encoder achieved the best results across all three sense levels. To isolate the effect of explicitly modeling hierarchical dependencies in multi-label IDRR, the performance of RoBERTa-HArch was compared against MTask (which also uses RoBERTa<sub>base</sub>) after training it on the English subset of DiscoGeM 2.0. As shown in Table 5.3, RoBERTa-HArch achieved lower JS distances at level-2 and level-3 compared to MTask. However, at level-1, the performance of both models is comparable - which is to be expected as level-1 predictions do not benefit from lower-level inputs. These experimental results show an improvement of performance of the new HArch architecture over MTask for higher-level sense prediction.

In the multilingual setting, HArch was evaluated with three multilingual PLM encoders: XLM-RoBERTa<sub>base</sub>, FLAN-T5<sub>base</sub> (Chung et al., 2024) and the recently released EuroBERT<sub>210m</sub> (Boizard et al., 2025). As shown in Table 5.3, XLM-RoBERTa-HArch achieved the highest overall performance, outperforming both configuration using the much more recent FLAN-T5<sub>base</sub> and EuroBERT<sub>210m</sub> models. It is worth noting that performance in this setting was consistently lower than in the English-only experiments, which can be attributed to the added complexity of multilingual IDRR. Given the relatively lower number of annotated instances in German, French and Czech in the DiscoGeM 2.0 corpus (see Section 2.2.3), no experiments were conducted with language specific PLM encoders for these languages.

In addition to fine-tuned models, the proposed few-shot prompting framework was also evaluated with GPT-4o and Llama-4-Maverick in both multilingual and English-only settings. As shown in the rows shaded in gray of Table 5.3, GPT-4o consistently outperforms

| Language |       | Model             | JS Distance ↘                       |                                     |                                     | Params |
|----------|-------|-------------------|-------------------------------------|-------------------------------------|-------------------------------------|--------|
| Test     | Train |                   | Level-1                             | Level-2                             | Level-3                             |        |
| Eng      | Eng   | MTask             | $0.329 \pm 0.005$                   | $0.498 \pm 0.004$                   | $0.569 \pm 0.005$                   | 125M   |
|          |       | RoBERTa-HArch     | <b><math>0.327 \pm 0.004</math></b> | <b><math>0.478 \pm 0.005</math></b> | <b><math>0.541 \pm 0.005</math></b> | 125M   |
|          |       | ModernBERT-HArch  | $0.364 \pm 0.004$                   | $0.507 \pm 0.005$                   | $0.592 \pm 0.006$                   | 149M   |
|          |       | XLM-RoBERTa-HArch | $0.343 \pm 0.004$                   | $0.487 \pm 0.006$                   | $0.564 \pm 0.005$                   | 125M   |
|          |       | GPT-4o            | $0.373 \pm 0.002$                   | $0.559 \pm 0.003$                   | $0.657 \pm 0.003$                   | 200B   |
|          |       | Llama-4-Maverick  | $0.423 \pm 0.004$                   | $0.641 \pm 0.006$                   | $0.711 \pm 0.007$                   | 400B   |
| All      | All   | XLM-RoBERTa-HArch | <b><math>0.356 \pm 0.005</math></b> | <b><math>0.531 \pm 0.006</math></b> | <b><math>0.605 \pm 0.006</math></b> | 125M   |
|          |       | EuroBERT-HArch    | $0.414 \pm 0.006$                   | $0.569 \pm 0.007$                   | $0.641 \pm 0.006$                   | 210M   |
|          |       | Flan-T5-HArch     | $0.364 \pm 0.003$                   | $0.550 \pm 0.003$                   | $0.623 \pm 0.004$                   | 220M   |
|          |       | GPT-4o            | $0.438 \pm 0.002$                   | $0.671 \pm 0.002$                   | $0.709 \pm 0.003$                   | 200B   |
|          |       | Llama-4-Maverick  | $0.534 \pm 0.004$                   | $0.774 \pm 0.004$                   | $0.839 \pm 0.005$                   | 400B   |

Table 5.3: Results of experimenting with different PLMs as encoders of the HArch model. The results are the average of three different runs with random starts. Results are reported on the English language and on all languages simultaneously. Values in bold show the best score at each level in each language setting. Rows shaded in gray present the results of the LLMs with the few-shot prompting framework. The final column reports the approximate number of parameters of each model configuration.

Llama-4-Maverick across all sense levels, following a pattern already observed between GPT-4 and LLaMa on single-label IDRR in the English language (Yung et al., 2024a). However, despite its stronger performance, GPT-4o still does not surpass any of the fine-tuned smaller models at any sense level in either language setting. Similar to the fine-tuned models, there is a significant drop in performance in the LLM models in the multilingual setting compared to the English-only setting - particularly for Llama-4-Maverick at level-2 and level-3 classification.

## 5.5 Results

Before presenting the multilingual results on DiscoGeM 2.0, HArch was first evaluated on DiscoGeM under the same conditions as MTask (see Section 4.6.1). Since DiscoGeM contains only English annotations, RoBERTa<sub>base</sub> was used as the encoder. Table 5.4 presents the results of both models. Similar to the results presented in Table 5.3, RoBERTa-HArch achieved better performance at level-2 and level-3 than MTask, but obtained comparable results at level-1. While level-1 predictions do not benefit from lower-level inputs, the

| Model         | JS Distance ↘        |                      |                      |
|---------------|----------------------|----------------------|----------------------|
|               | Level-1              | Level-2              | Level-3              |
| MTask         | <b>0.299 ± 0.002</b> | 0.446 ± 0.003        | 0.523 ± 0.002        |
| RoBERTa-HArch | 0.302 ± 0.003        | <b>0.435 ± 0.003</b> | <b>0.506 ± 0.004</b> |

Table 5.4: Final results of the MTask and the HArch models on the test set of DiscoGeM. The results are the average of three different runs with random starts. Values in bold show the best score for each metric.

superior performance of HArch at level-2 and level-3 reinforces the previous findings that explicitly modeling hierarchical dependencies in multi-label IDRR improves performance for higher-level sense prediction.

Having compared the performance of MTask and HArch on multi-label IDRR using the DiscoGeM corpus annotated in English, Table 5.5 presents the results of HArch and the few-shot prompting framework on multi-label IDRR using the DiscoGeM 2.0 corpus across different languages. The models were evaluated in a multilingual setting using all languages of the corpus simultaneously as well on each of its languages individually. For individual languages, the HArch models were trained either on the full multilingual corpus or on the evaluated language alone. Based on its performance in the multilingual setting on the validation split of DiscoGeM 2.0 (see Table 5.3), XLM-RoBERTa-HArch was used in the multilingual setting and for all individual languages, while RoBERTa-HArch was also used for the English language. The few-shot prompting framework was evaluated on all language settings using the multilingual GPT-4o Llama-4-Maverick - the two most recent LLMs at the time of these results.

The results in Table 5.5 show that the HArch models consistently outperform both LLMs using the few-shot prompting framework across all languages and all sense levels. The performance gap is particularly pronounced in all settings with the exception of the English language. Notably, XLM-RoBERTa-HArch achieves substantially lower JS distances than both LLMs at level-2 and level-3 when evaluated on the multilingual setting using all languages and when evaluated on the German and Czech languages alone. Consistent with earlier results in Table 5.3, GPT-4o outperforms Llama-4-Maverick across all sense levels and language configurations, but does not outperform any of the smaller fine-tuned HArch models. The performance of Llama-4-Maverick in the German and in the Czech language is particularly low - especially for the latter. These findings reinforce the idea that model size and complexity are not sufficient substitute for application-specific training data in text classification (Chan et al., 2024; Yung et al., 2024a; Edwards and Camacho-Collados, 2024;

| Language |       | Model             | JS Distance ↘        |                      |                      |
|----------|-------|-------------------|----------------------|----------------------|----------------------|
| Test     | Train |                   | Level-1              | Level-2              | Level-3              |
| All      | All   | XLM-RoBERTa-HArch | <b>0.353 ± 0.005</b> | <b>0.529 ± 0.005</b> | <b>0.606 ± 0.04</b>  |
|          | —     | GPT-4o            | 0.434 ± 0.002        | 0.678 ± 0.003        | 0.705 ± 0.003        |
|          | —     | Llama-4-Maverick  | 0.529 ± 0.004        | 0.769 ± 0.005        | 0.830 ± 0.005        |
| Eng      | Eng   | RoBERTa-HArch     | <b>0.331 ± 0.004</b> | <b>0.477 ± 0.004</b> | <b>0.548 ± 0.005</b> |
|          | All   | XLM-RoBERTa-HArch | 0.349 ± 0.004        | 0.493 ± 0.006        | 0.568 ± 0.005        |
|          | Eng   |                   | 0.347 ± 0.005        | 0.486 ± 0.006        | 0.561 ± 0.006        |
|          | —     | GPT-4o            | 0.368 ± 0.001        | 0.551 ± 0.001        | 0.647 ± 0.002        |
|          | —     | Llama-4-Maverick  | 0.420 ± 0.001        | 0.634 ± 0.002        | 0.702 ± 0.002        |
|          | —     |                   |                      |                      |                      |
| Ger      | All   | XLM-RoBERTa-HArch | <b>0.360 ± 0.007</b> | <b>0.566 ± 0.006</b> | <b>0.630 ± 0.006</b> |
|          | Ger   |                   | 0.393 ± 0.006        | 0.586 ± 0.006        | 0.656 ± 0.007        |
|          | —     | GPT-4o            | 0.411 ± 0.002        | 0.644 ± 0.002        | 0.725 ± 0.002        |
|          | —     | Llama-4-Maverick  | 0.509 ± 0.003        | 0.732 ± 0.006        | 0.809 ± 0.007        |
| Fre      | All   | XLM-RoBERTa-HArch | <b>0.367 ± 0.006</b> | <b>0.553 ± 0.005</b> | <b>0.628 ± 0.006</b> |
|          | Fre   |                   | 0.382 ± 0.004        | 0.576 ± 0.006        | 0.655 ± 0.007        |
|          | —     | GPT-4o            | 0.406 ± 0.001        | 0.611 ± 0.001        | 0.689 ± 0.002        |
|          | —     | Llama-4-Maverick  | 0.473 ± 0.003        | 0.667 ± 0.003        | 0.748 ± 0.004        |
| Cze      | All   | XLM-RoBERTa-HArch | <b>0.371 ± 0.010</b> | <b>0.555 ± 0.009</b> | <b>0.647 ± 0.009</b> |
|          | Cze   |                   | 0.404 ± 0.009        | 0.584 ± 0.011        | 0.689 ± 0.010        |
|          | —     | GPT-4o            | 0.549 ± 0.011        | 0.820 ± 0.010        | 0.854 ± 0.012        |
|          | —     | Llama-4-Maverick  | 0.613 ± 0.005        | 0.864 ± 0.004        | 0.890 ± 0.005        |

Table 5.5: Final results of the HArch model and of the LLMs with the few-shot prompting framework on the test set of DiscoGeM 2.0. Results are reported on all of the languages simultaneously and on each language individually. The results are the average of three different runs with random starts. Values in bold show the best score for each metric. Rows shaded in gray present the results of the LLMs with the few-shot prompting framework.

Bucher and Martini, 2024). Thus, for multi-label IDRR, fine-tuning smaller encoder-based model is still preferable to LLM prompting.

Examining the performance of the models in each language setting reveals the existing bias in PLMs and LLMs towards the English language (Joshi et al., 2020; Faisal et al., 2022; Ranathunga and de Silva, 2022; Levisen, 2019; Blasi et al., 2022). Models fine-tuned or prompted in English consistently achieve substantially higher performance than in any other language setting. This bias is also evident in the multilingual configuration, where models evaluated on all languages simultaneously outperform their counterparts evaluated on the German, French and Czech languages alone. Among these three languages, XLM-RoBERTa-HArch delivers relatively consistent results, whereas both LLMs perform best in French and worst in Czech. This is particularly more visible at higher-level senses, where the distinctions between sense labels are more fine-grained. These differences likely reflect the previous exposure of these LLMs to training data in these languages during pre-training - with French being better represented than Czech.

Finally, XLM-RoBERTa-HArch achieves higher performance on German, French and Czech when trained on all languages in DiscoGeM 2.0 compared to when trained on the evaluated language alone. This indicates that multilingual training can help offset the reduced annotated data in these languages. Conversely, in the English language for which there is more annotated data, XLM-RoBERTa-HArch performs slightly better when fine-tuned exclusively on the English subset of the corpus. However, in both cases, it fails to outperform RoBERTa-HArch - which remains the strongest overall model for English. When comparing the performance of RoBERTa-HArch on DiscoGeM in Table 5.4 and on DiscoGeM 2.0 in Table 5.5, there is a drop across all sense levels on the latter. This drop can be attributed to the label set used for evaluation on DiscoGeM (see Section 4.2.1). The reduced set of 14 level-2 senses, and consequent level-3 senses, lowers the complexity of the multi-label IDRR task compared to DiscoGeM 2.0 where all sense labels are kept.

## 5.6 Further Analysis

While the main results show that HArch outperforms both LLMs using the few-shot prompting framework for multilingual and multi-label IDRR, further analysis is needed to better understand the factors driving its performance. In particular, it is important to understand the contribution of the multi-task architecture, as well as to examine how the hierarchical information is actually leveraged during training. This section provides two complementary analyses to address these questions and provide further insights into the practical benefits of the proposed model architecture.

| Language |       | Model                  | JS Distance $\searrow$              |                                     |                                     | Time    |
|----------|-------|------------------------|-------------------------------------|-------------------------------------|-------------------------------------|---------|
| Test     | Train |                        | Level-1                             | Level-2                             | Level-3                             |         |
| Eng      | Eng   | RoBERTa-HArch          | <b>0.331 <math>\pm</math> 0.004</b> | <b>0.477 <math>\pm</math> 0.004</b> | <b>0.548 <math>\pm</math> 0.005</b> | 9h 09m  |
|          |       | RoBERTa-Individual     | 0.342 $\pm$ 0.004                   | 0.513 $\pm$ 0.004                   | 0.589 $\pm$ 0.005                   | 21h 56m |
| All      | All   | XLM-RoBERTa-HArch      | <b>0.353 <math>\pm</math> 0.005</b> | <b>0.529 <math>\pm</math> 0.005</b> | <b>0.606 <math>\pm</math> 0.04</b>  | 20h 31m |
|          |       | XLM-RoBERTa-Individual | 0.365 $\pm$ 0.004                   | 0.552 $\pm$ 0.005                   | 0.641 $\pm$ 0.004                   | 59h 14m |

Table 5.6: Results evaluating the contribution of the multi-task architecture of HArch. The results are the average of three different runs with random starts. Results are reported on the English language and on all languages simultaneously. Values in bold show the best score at each level in each language setting. The final column reports the total time required to compute all of the averaged results in each row.

### 5.6.1 Multi-Task Architecture

To evaluate the contribution of the multi-task architecture in the multi-label IDRR task, this section compares the performance of HArch at each sense level against individual models trained exclusively on each sense level. While the HArch model (see Figure 4.3) jointly learns how to predict probability distributions for all three sense levels leveraging a shared linear layer between classification heads and aggregated loss function, each individual model consists of a PLM encoder followed by a linear and dropout layer and a single classification head trained to predict probability distributions for a single sense level. To ensure a fair comparison, the architecture of the individual models corresponds to a reduced version of HArch containing only a single classification head and no augmentation blocks. Table 5.6 compares the results of the three individual models at each sense level against RoBERTa-HArch in the English-only setting and XLM-RoBERTa-HArch in the multilingual setting using the test split of DiscoGeM 2.0.

As shown in Table 5.6, the HArch models consistently outperform the individual models across all three sense levels. The performance gains are especially pronounced at level-2 and level-3, suggesting that these finer-grained senses benefit significantly from the shared representations and hierarchical cues provided by joint training. While level-1 predictions also improve slightly, the main advantage of multi-task learning comes from its ability to propagate information through the sense hierarchy. This study isolates the impact of the multi-task architecture and confirms that jointly modeling the hierarchy enhances multi-label IDRR performance. In addition to yielding better performance, HArch is also more efficient - as shown by the total time it took to calculate the averaged results in each row of Table 5.6 with a 32-core compute node with 512GB of RAM. This highlights both the effectiveness and efficiency of the proposed multi-task architecture.

| Language |       | Model             | $\alpha$ (L1 $\rightarrow$ L2) | $\beta_1$ (L1 $\rightarrow$ L3) | $\beta_2$ (L2 $\rightarrow$ L3) |
|----------|-------|-------------------|--------------------------------|---------------------------------|---------------------------------|
| Test     | Train |                   |                                |                                 |                                 |
| Eng      | Eng   | RoBERTa-HArch     | $0.5178 \pm 0.0002$            | $0.2693 \pm 0.0004$             | $0.2613 \pm 0.0001$             |
| All      | All   | XLM-RoBERTa-HArch | $0.5427 \pm 0.0005$            | $0.2945 \pm 0.0001$             | $0.2890 \pm 0.0002$             |

Table 5.7: Evaluation of the learnable parameters in the weighted sums controlling the hierarchical learning of HArch. The results are the average of three different runs with random starts. Results are reported on the English language and on all languages simultaneously.

### 5.6.2 Hierarchical Learning

To examine how hierarchical information is leveraged during training, this section analyzes the learned weights controlling the cascading of information across sense levels in HArch (see Section 5.3.1). The model has two weighted sums to combine lower-level predictions with the shared representation of the encoded input arguments. The first, defined in Equation 5, quantifies the influence of the level-1 predictions on level-2 predictions through a learnable parameter  $\alpha$ . The second, defined in Equation 6, quantifies the contribution of both level-1 and level-2 predictions on level-3 predictions through the learnable parameters  $\beta_1$  and  $\beta_2$ , respectively. These parameters are adjusted jointly with the rest of the model during training. Table 5.7 presents their final values for the models used to obtain the results in Section 5.5 in the English-only and multilingual settings.

The values of the learnable parameters  $\alpha$ ,  $\beta_1$  and  $\beta_2$  in Table 5.7 show that the predictions of higher-level classification heads are significantly influenced from lower-level predictions. This cascading of information is the main difference between the MTask and the HArch models and the driving factor behind the improved performance of HArch in Table 5.4. It is interesting to note that the parameters converge to similar values in both language settings, suggesting a consistent learning pattern. In both cases,  $\alpha$  indicates that slightly more than half of the information for level-2 predictions comes from level-1 predictions, while  $\beta_1$  and  $\beta_2$  show that nearly two-thirds of the information for level-3 predictions comes from level-1 and level-2 predictions, respectively. These values highlight the importance of explicitly modeling hierarchical dependencies across sense levels in multi-label IDRR.

## 5.7 Discussion

This chapter introduced HArch, the first multilingual and multi-label classification model for IDRR. Building upon the previously proposed multi-label model MTask, HArch

explicitly models hierarchical dependencies across sense levels in the PDTB 3.0 framework. The model was trained on the DiscoGeM 2.0 corpus to learn multi-label representations of discourse relations in the English, German, French and Czech language in a multilingual setting. To allow a direct comparison with MTask, the HArch model was also separately trained on the DiscoGeM corpus using only the English language. In all experiments, HArch was fine-tuned using the optimal hyperparameter configuration previously established for MTask in the multi-label setting (Section 4.5.2) with the MAE loss function (Section 4.5.1). Results in Table 5.4 show that HArch achieves superior performance at levels 2 and 3 while maintaining comparable results at level 1 to MTask using the same RoBERTa<sub>base</sub> encoder. While level-1 predictions do not directly benefit from lower-level inputs, the performance at higher-level sense prediction demonstrate the advantage of explicitly modeling hierarchical dependencies for multi-label IDRR.

In the DiscoGeM 2.0 corpus, the proposed HArch model establishes the first benchmark across all language settings. Following the experimental results in Table 5.3, XLM-RoBERTa<sub>base</sub> was used as the PLM encoder when training and evaluating HArch on each language individually as well as in the multilingual setting. In addition, when trained and evaluated on the English language, RoBERTa<sub>base</sub> is also used as the PLM encoder of HArch. The results in Table 5.5 show that XLM-RoBERTa-HArch achieves its best performance when evaluated exclusively in the English subset of DiscoGeM 2.0. However, its performance on all of the other languages individually is lower than on the multilingual setting - which can be attributed to the presence of English annotations in the latter. Moreover, with the exception of the English language, the XLM-RoBERTa-HArch model consistently performs better when trained on all languages simultaneously than when trained solely on the language being evaluated. This hints at the existing bias in multilingual models towards the English language (Joshi et al., 2020; Faisal et al., 2022; Ranathunga and de Silva, 2022; Levisen, 2019; Blasi et al., 2022). Despite its strong performance on the English language, the XLM-RoBERTa-HArch model is outperformed by RoBERTa-HArch - highlighting the benefits of using PLMs pre-trained specifically on the evaluated language. Due to the reduced number of annotated discourse relations in the other languages of the DiscoGeM 2.0 corpus (see Section 2.2.3), it was not feasible to use language-specific PLMs for the German, French and Czech languages.

In addition to the encoder-based HArch model, this chapter also evaluated a few-shot prompting framework for multilingual and multi-label IDRR using the two most recent LLMs: GPT-4o and Llama-4-Maverick. The results in Table 5.5 show that both LLMs perform best when evaluated on the English language. In contrast to the HArch model, both LLMs only perform better in the multilingual setting when compared against the Czech

language alone. When evaluated on the French and German language they achieve better results than in the multilingual setting. Despite having fewer parameters and being slightly older (see Table 5.3), GPT-4o consistently outperforms Llama-4-Maverick. However, neither LLM surpasses the smaller fine-tuned HArch model in any language setting. The performance gap is particularly pronounced at level-2 and level-3 - particularly for the German, Czech and multilingual setting. The performance of Llama-4-Maverick in the German and Czech languages is particularly low - especially for the latter. These findings show the continued importance of application-specific fine-tuning in text classification, despite the increasing model size and complexity of LLMs (Chan et al., 2024; Yung et al., 2024a; Edwards and Camacho-Collados, 2024; Bucher and Martini, 2024). Thus, for multi-label IDRR, fine-tuning smaller encoder-based models remains the more effective and cost-efficient approach - particularly when considering the significantly higher monetary and environmental costs of LLM prompting.

Should future versions of the DiscoGeM 2.0 corpus include additional annotations in German, French and Czech, it would be interesting to evaluate the performance of encoder-based models for multi-label IDRR using PLMs pre-trained specifically on these language. Future work should also explore fine-tuning LLMs for multilingual and multi-label IDRR through parameter-efficient techniques such as Low-Rank Adaptation (Hu et al., 2022). This would allow a more direct comparison between the benefits and drawbacks of fine-tuning and prompting for this specific task. In addition, the evaluation of the prompting framework presented in this chapter was performed under limited experimental conditions. Given the few previous works in this field using LLMs, future research should explore different prompting techniques with more exhaustive experiments to better evaluate the performance of LLMs in IDRR. Finally, another promising line of work for multi-label IDRR would be to train Graph Neural Networks to capture the co-occurrence patterns and hierarchical relationships between discourse senses in multi-label discourse annotated corpora.

# Chapter 6

## Conclusion

Computational discourse analysis allows models to move beyond sentence-level understanding and capture the logical and structured flow of ideas that shape meaning across larger stretches of text. It enables applications such as summarization, machine translation, question-answering and dialogue systems to produce more coherent and contextually appropriate responses. The work presented in this thesis seeks to address different sources of divergence currently hindering progress in computational discourse analysis. More precisely, this thesis explored (i) how discourse relations can be mapped across frameworks (Chapter 3), (ii) how ambiguous implicit discourse relations can be modeled with multiple senses (Chapter 4) and (iii) how discourse modeling can be extended to multilingual contexts (Chapter 5). To address these challenges, this thesis has proposed new methodologies, developed novel model architectures and established benchmarks that together advance the goal of harmonizing divergence in computational discourse analysis.

### 6.1 Contributions

The structure of the thesis was organized around three main research objectives, each targeting a different source of divergence in computational discourse analysis. This section summarizes the methodology and main contributions of the research conducted to address each of these objectives.

**Objective 1.** Map explicit and implicit discourse relations between the RST and the PDTB frameworks.

Progress in computational discourse analysis is constrained by the absence of a consensual theoretical framework, which creates obstacles for the comparison, evaluation and

interoperability of discourse models. To address this gap, Chapter 3 proposes an empirical mapping between the RST and the PDTB 3.0 discourse frameworks based on the paper published at Recent Advances in Natural Language Processing (RANLP) 2023 (Costa et al., 2023c). The proposed mapping methodology was able to map 6,510 (80.0%) of the 8,134 explicit and implicit discourse relations in the 385 overlapping Wall Street Journal articles annotated in the two most widely used corpora for each framework: the RST-DT and the PDTB 3.0. More precisely, the mapping algorithm was able to map 3,073 (73.7%) of the 4,169 explicit relations and 3,437 (86.7%) of the 3,965 implicit relations. Despite their lower mapping percentage, the mappings obtained for explicit discourse relation found clearer one-to-one correspondences between both frameworks compared to implicit discourse relations - which can be attributed to their inherently more subjective nature. The main contributions of this work are:

- First mapping of explicit and implicit discourse relations between the RST and the PDTB 3.0 discourse frameworks.
- Total mapping coverage of 6,510 (80.0%) of the 8,134 explicit and implicit discourse relations annotated in Wall Street Journal articles following both frameworks.
- Mapping coverage of 3,073 (73.7%) of the 4,169 explicit relations with high one-to-one mapping between frameworks.
- Mapping coverage of 3,437 (86.7%) of the 3,965 implicit relations with low one-to-one mapping between frameworks.

The mapping methodology proposed in this thesis has been used in subsequent research efforts to improve interoperability between discourse frameworks (Fu, 2024; Eichin et al., 2025). For instance, Fu (2024) introduce a fully automatic approach to map discourse relations between the RST-DT and the PDTB 3.0 corpora using contrastive learning to generate label embeddings. These embeddings are used to map the discourse relations between the two corpora without the need to align their arguments. They compare the results of mapping explicit discourse relations using their method against the one proposed in this thesis and achieve less consistent one-to-one mappings. Nevertheless, automatically mapping discourse relations using embeddings is a promising research direction. Conversely, Eichin et al. (2025) build on recent mapping efforts to define a unified set of 17 discourse relations to enable cross-framework interoperability. This unified set of discourse relations is currently being used in the ongoing edition of DISRPT 2025 to allow evaluation across different corpora in their discourse classification track.

**Objective 2.** Develop a multi-label IDRR classification model capable of capturing multiple implicit senses.

Traditional approaches to IDRR follow the assumption that each discourse relation should be assigned a single sense. However, the observed ambiguity in the interpretation of implicit discourse relations observed in Chapter 3 motivates the need for models capable of moving beyond single-label classification and capture the multiple possible senses of implicit relations. To address this challenge, Chapter 4 introduces a novel multi-label approach to IDRR based on the papers published at the Workshop on Computational Approaches to Discourse (CODI) 2024 (Costa and Kosseim, 2024) and at the Special Interest Group on Discourse and Dialogue (SIGDIAL) 2025 (Costa and Kosseim, 2025a). The approach is based on the multi-task classification model, MTask, which was trained on the DiscoGeM corpus to jointly learn multi-label representations of discourse relations across all three sense levels in the PDTB 3.0 framework. The model was also adapted to the single-label setting where it was evaluated on the DiscoGeM and the PDTB 3.0 corpus in a transfer learning setting. The main contributions of this work are:

- First multi-label approach to IDRR based on the MTask model. The model can be adapted to learn either multi-label or single-label representations of discourse relations and is trained on the DiscoGeM corpus.
- Extensive experiments comparing different PLMs as encoders of the MTask model and evaluating how different loss functions impact performance on multi-label and single-label IDRR.
- First benchmark on multi-label IDRR and SOTA results on single-label IDRR using DiscoGeM.
- First in-depth analysis on the potential of transfer learning between the DiscoGeM and the PDTB 3.0 corpus on single-label IDRR.

**Objective 3.** Develop a multilingual IDRR classification model capable of capturing multiple implicit senses across different languages.

Most research in computational discourse analysis is centered around the English language, which overlooks the wide diversity of languages in the world. To more accurately capture the complexity of human reasoning, it is important to develop discourse models

capable of generalizing across multiple languages. Therefore, Chapter 5 extends the multi-label IDRR approach to different languages based on the paper published at the Special Interest Group on Discourse and Dialogue (SIGDIAL) 2025 (Costa and Kosseim, 2025b). The multilingual approach is based on the HArch classification model which was trained on the DiscoGeM 2.0 corpus to learn multi-label representations of discourse relations in different languages. The model improves on the previous MTask model by explicitly modeling hierarchical dependencies across sense levels in the PDTB 3.0 framework. Chapter 5 also explores a few-shot prompting framework to leverage LLMs on multilingual and multi-label IDRR. The main contributions of this work are:

- First multilingual approach to IDRR based on the HArch model. The model improves on the previous MTask model by explicitly modeling hierarchical dependencies across sense levels.
- HArch achieved SOTA results on multi-label IDRR using DiscoGeM.
- First benchmark on multi-label IDRR using the DiscoGeM 2.0 corpus in both the multilingual setting and on the English, German, French and Czech languages individually.
- Comparison of different PLMs as encoders of the HArch model in different language settings for multi-label IDRR.
- Evaluation of the GPT-4o and Llama-4-Maverick LLMs using few-shot prompting for multi-label IDRR.

Collectively, the contributions of this thesis offer a unified response to different sources of divergence in computational discourse analysis. First, by establishing an empirical mapping between the RST and the PDTB 3.0 frameworks, the thesis offers meaningful cross-framework comparisons and contributed to recent efforts on the interoperability between annotation methodologies (Fu, 2024; Eichin et al., 2025). Second, by reframing IDRR as a multi-label classification problem, it challenges the limitations of single-label assumptions by capturing ambiguity of implicit discourse relations rather than discarding it. It also establishes new benchmarks and provides practical guidance on the choice of architectures and loss functions for IDRR. Third, by extending IDRR to a multilingual setting and proposing hierarchical architectures alongside prompting strategies, the thesis lays the groundwork for more inclusive and generalizable discourse modeling approaches across

languages. Taken together, these contributions advance the overarching goal of harmonizing divergence in computational discourse analysis and offer more generalizable and inclusive methods for discourse modeling across frameworks, ambiguity and languages.

## 6.2 Future Work

While the proposed mapping between the RST and the PDTB 3.0 frameworks marks an important step toward cross-framework interoperability, there are a few limitations. Future work should consider incorporating additional relation types such as AltLex and AltLexC in the PDTB 3.0 framework. Despite being much less frequent, they are also annotated with discourse senses and offer opportunities to broaden the mapping coverage. Moreover, the current methodology has been evaluated exclusively on a limited set of overlapping annotations of gold-standard corpora. As high-quality discourse parsers for the RST and the PDTB 3.0 continue to improve, it will be important to evaluate the robustness of the proposed mapping on automatically parsed texts. This would not only validate its applicability in practical settings but also extend its utility to downstream tasks. Finally, expanding the mapping methodology to include other theoretical frameworks such as SDRT and eRST would ultimately contribute toward a more unified representation of discourse.

In the domain of multi-label IDRR, several promising research directions emerge from this thesis. For instance, exploring targeted data augmentation techniques to reduce class imbalance and enhance model performance on under-represented discourse senses. Additionally, when considering only the English language, further exploration of intermediate fine-tuning from DiscoGeM to PDTB 3.0 in the transfer learning setting, should lead to better cross-corpus generalizability. For multilingual and multi-label IDRR, should additional annotations in the German, French and Czech language be available, it would make it possible to evaluate the HArch model using PLMs pre-trained specifically on these languages and potentially improve language-specific performance. Additionally, parameter-efficient fine-tuning techniques such as Low-Rank Adaptation could also enable a different adaptation of LLMs to this task other than prompting, providing a more direct comparison between fine-tuning and prompting approaches to multi-label IDRR. Another promising avenue is to use Graph Neural Networks to capture co-occurrence patterns and hierarchical relationships between discourse senses.

Finally, across all research directions, it is important to be mindful of the environmental and computational costs associated with large-scale experimentation with PLMs and LLMs. While access to high-performance computing resources enables exhaustive experimentation, it also raises important concerns about the sustainability of such research

practices. Although often overlooked, these practices lead to a non-negligible carbon footprint. As the field progresses, future work should prioritize more efficient training regimes, leverage parameter-efficient fine-tuning techniques and report energy consumption whenever possible. Incorporating these practices not only aligns with broader efforts to reduce the environmental impact of artificial intelligence but also promotes greater accountability and reproducibility in computational discourse analysis.

# Bibliography

- Agerri, R., San Vicente, I., Campos, J. A., Barrena, A., Saralegi, X., Soroa, A., and Agirre, E. (2020). Give your Text Representation Models some Love: the Case for Basque. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC'20)*, pages 4781–4788, Marseille, France. European Language Resources Association (ELRA).
- Aktas, B. and Roth, M. (2025). Clarifying Underspecified Discourse Relations in Instructional Texts. In *Findings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL'25)*, pages 12237–12256, Vienna, Austria. Association for Computational Linguistics (ACL).
- Anuranjana, K. (2023). DiscoFlan: Instruction Fine-tuning and Refined Text Generation for Discourse Relation Label Classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT'23)*, pages 22–28, Toronto, Ontario, Canada. Association for Computational Linguistics (ACL).
- Asher, N. and Lascarides, A. (2003). *The Logical Form of Discourse*, chapter 4, pages 111–178. Cambridge University Press.
- Basile, V., Fell, M., Fornaciari, T., Hovy, D., Paun, S., Plank, B., Poesio, M., and Uma, A. (2021). We Need to Consider Disagreement in Evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics (ACL).
- Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., and Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences*, 26(12):1153–1170.
- Boizard, N., Gisserot-Boukhlef, H., Alves, D. M., Martins, A., Hammal, A., Corro, C., Hudelot, C., Malherbe, E., Malaboeuf, E., Jourdan, F., Hautreux, G., Alves, J., El-Haddad, K., Faysse, M., Peyrard, M., Guerreiro, N. M., Fernandes, P., Rei, R., and

- Colombo, P. (2025). EuroBERT: Scaling Multilingual Encoders for European Languages. In *Proceedings of the 2nd Conference on Language Modelling (COLM'25)*, Montreal, Quebec, Canada.
- Bourgonje, P. and Lin, P.-J. (2024). Projecting Annotations for Discourse Relations: Connective Identification for Low-Resource Languages. In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI'24)*, pages 39–49, St. Julian's, Malta. Association for Computational Linguistics (ACL).
- Braud, C., Coavoux, M., and Søgaard, A. (2017). Cross-lingual RST Discourse Parsing. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL'17)*, pages 292–304, Valencia, Spain. Association for Computational Linguistics (ACL).
- Braud, C., Liu, Y. J., Metheniti, E., Muller, P., Rivière, L., Rutherford, A., and Zeldes, A. (2023). The DISRPT 2023 Shared Task on Elementary Discourse Unit Segmentation, Connective Detection, and Relation Classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT'23)*, pages 1–21, Toronto, Ontario, Canada. Association for Computational Linguistics (ACL).
- Bucher, M. J. J. and Martini, M. (2024). Fine-Tuned 'Small' LLMs (Still) Significantly Outperform Zero-Shot Generative AI Models in Text Classification. *arXiv preprint arXiv:2406.08660*.
- Bunt, H., Alexandersson, J., Choe, J.-W., Fang, A. C., Hasida, K., Petukhova, V., Popescu-Belis, A., and Traum, D. (2012). ISO 24617-2: A semantically-based standard for dialogue annotation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 430–437, Istanbul, Turkey. European Language Resources Association (ELRA).
- Bunt, H., Petukhova, V., Gilmartin, E., Pelachaud, C., Fang, A., Keizer, S., and Prévot, L. (2020). The ISO Standard for Dialogue Act Annotation, Second Edition. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC'20)*, pages 549–558, Marseille, France. European Language Resources Association (ELRA).
- Bunt, H. and Prasad, R. (2016). ISO DR-Core (ISO 24617-8): Core Concepts for the Annotation of Discourse Relations. In *Proceedings of the 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA'16)*, pages 45–54, Portorož, Slovenia. European Language Resources Association (ELRA).

- Carlson, L., Marcu, D., and Okurovsky, M. E. (2001). Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In *Proceedings of the 2nd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL'01)*, Aalborg, Denmark. Association for Computational Linguistics (ACL).
- Carlson, L., Marcu, D., and Okurowski, M. E. (2003). *Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory*, pages 85–112. Springer.
- Cartoni, B. and Meyer, T. (2012). Extracting Directional and Comparable Corpora from a Multilingual Corpus for Translation Studies. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 2132–2137, Istanbul, Turkey. European Language Resources Association (ELRA).
- Chan, C., Jiayang, C., Wang, W., Jiang, Y., Fang, T., Liu, X., and Song, Y. (2024). Exploring the Potential of ChatGPT on Sentence Level Relations: A Focus on Temporal, Causal, and Discourse Relations. In *Findings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL'24)*, pages 684–721, St. Julian's, Malta. Association for Computational Linguistics (ACL).
- Chan, C., Liu, X., Cheng, J., Li, Z., Song, Y., Wong, G. Y., and See, S. (2023). Disco-Prompt: Path Prediction Prompt Tuning for Implicit Discourse Relation Recognition. In *Findings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL'23)*, pages 35–57, Toronto, Ontario, Canada. Association for Computational Linguistics (ACL).
- Chen, J. and Yang, D. (2021). Structure-Aware Abstractive Conversation Summarization via Discourse and Action Graphs. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'21)*, pages 1380–1391, Online. Association for Computational Linguistics (ACL).
- Chen, T., Jiang, Z., Poliak, A., Sakaguchi, K., and Van Durme, B. (2020a). Uncertain Natural Language Inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL'20)*, pages 8772–8779, Online. Association for Computational Linguistics (ACL).
- Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020b). A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning (PMLR'20)*, pages 1597–1607, Online. Journal of Machine Learning Research (JMLR).

- Chi, T.-C. and Rudnicky, A. (2022). Structured Dialogue Discourse Parsing. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL'22)*, pages 325–335, Edinburgh, United Kingdom. Association for Computational Linguistics (ACL).
- Chiarcos, C. (2012). Ontologies of Linguistic Annotation: Survey and perspectives. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 303–310, Istanbul, Turkey. European Language Resources Association (ELRA).
- Chiarcos, C. (2014). Towards interoperable discourse annotation. Discourse features in the Ontologies of Linguistic Annotation. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*, pages 4569–4577, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Chiarcos, C. and Sukhareva, M. (2015). OLiA – Ontologies of Linguistic Annotation. *Semantic Web*, 6(4):379–386.
- Choubey, P. K. and Huang, R. (2021). Profiling News Discourse Structure Using Explicit Subtopic Structures Guided Critics. In *Findings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP'21)*, pages 1594–1605, Punta Cana, Dominican Republic. Association for Computational Linguistics (ACL).
- Choubey, P. K., Lee, A., Huang, R., and Wang, L. (2020). Discourse as a Function of Event: Profiling Discourse Structure in News Articles around the Main Event. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL'20)*, pages 5374–5386, Online. Association for Computational Linguistics (ACL).
- Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al. (2024). Scaling Instruction-Finetuned Language Models. *Journal of Machine Learning Research (JMLR)*, 25(70):1–53.
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL'20)*, pages 8440–8451, Online. Association for Computational Linguistics (ACL).

- Costa, N. F., Cheng, Y., Muermans, T. C., Hanel, B., and Kosseim, L. (2023a). Automatic Identification of Chinese Paired Discourse Connectives. In *Proceedings of the 17th International Conference on Semantic Computing (ICSC'23)*, pages 114–117, Laguna Hills, California, USA. Institute of Electrical and Electronics Engineers (IEEE).
- Costa, N. F., Hamilton, B., and Kosseim, L. (2023b). CLaC at SemEval-2023 Task 3: Language Potluck RoBERTa Detects Online Persuasion Techniques in a Multilingual Setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval'23)*, pages 1613–1618, Toronto, Ontario, Canada. Association for Computational Linguistics (ACL).
- Costa, N. F. and Kosseim, L. (2024). Exploring Soft-Label Training for Implicit Discourse Relation Recognition. In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI'24)*, pages 120–126, St. Julians, Malta. Association for Computational Linguistics (ACL).
- Costa, N. F. and Kosseim, L. (2025a). A Multi-Task and Multi-Label Classification Model for Implicit Discourse Relation Recognition. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL'25)*, Avignon, France. Association for Computational Linguistics (ACL).
- Costa, N. F. and Kosseim, L. (2025b). Multi-Lingual Implicit Discourse Relation Recognition with Multi-Label Hierarchical Learning. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL'25)*, Avignon, France. Association for Computational Linguistics (ACL).
- Costa, N. F., Sheikh, N., and Kosseim, L. (2023c). Mapping Explicit and Implicit Discourse Relations between the RST-DT and the PDTB 3.0. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing (RANLP'23)*, pages 344–352, Varna, Bulgaria.
- Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., and Hu, G. (2020). Revisiting Pre-Trained Models for Chinese Natural Language Processing. In *Findings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP'20)*, pages 657–668, Online. Association for Computational Linguistics (ACL).
- Das, D. and Taboada, M. (2018). RST Signalling Corpus: A corpus of signals of coherence relations. *Language Resources and Evaluation*, 52(1):149–184.

- Demberg, V., Scholman, M. C., and Asr, F. T. (2019). How compatible are our discourse annotation frameworks? Insights from mapping RST-DT and PDTB annotations. *Dialogue & Discourse*, 10(1):87–135.
- Devatine, N., Muller, P., and Braud, C. (2022). Predicting Political Orientation in News with Latent Discourse Structure to Improve Bias Understanding. In *Proceedings of the 3rd Workshop on Computational Approaches to Discourse (CODI’22)*, pages 77–85, Gyeongju, Republic of Korea. International Conference on Computational Linguistics (COLING).
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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 (NAACL-HLT’19)*, pages 4171–4186, Minneapolis, Minnesota, USA. Association for Computational Linguistics (ACL).
- Dong, Y., Mircea, A., and Cheung, J. C. K. (2021). Discourse-Aware Unsupervised Summarization for Long Scientific Documents. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL’21)*, pages 1089–1102, Online. Association for Computational Linguistics (ACL).
- Du, H., Feng, Y., Li, C., Li, Y., Lan, Y., and Zhao, D. (2023). Structure-Discourse Hierarchical Graph for Conditional Question Answering on Long Documents. In *Findings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL’23)*, pages 6282–6293, Toronto, Ontario, Canada. Association for Computational Linguistics (ACL).
- Edwards, A. and Camacho-Collados, J. (2024). Language Models for Text Classification: Is In-Context Learning Enough? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING’24)*, pages 10058–10072, Torino, Italia. European Language Resources Association (ELRA) and International Conference on Computational Linguistics (COLING).
- Eichin, F., Liu, Y. J., Plank, B., and Hedderich, M. A. (2025). Probing LLMs for Multilingual Discourse Generalization Through a Unified Label Set. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL’25)*, pages 18665–18684, Vienna, Austria. Association for Computational Linguistics (ACL).

- Faisal, F., Wang, Y., and Anastasopoulos, A. (2022). Dataset Geography: Mapping Language Data to Language Users. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL’22)*, pages 3381–3411, Dublin, Ireland. Association for Computational Linguistics (ACL).
- Fornaciari, T., Uma, A., Paun, S., Plank, B., Hovy, D., and Poesio, M. (2021). Beyond Black & White: Leveraging Annotator Disagreement via Soft-Label Multi-Task Learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT’21)*, pages 2591–2597, Online. Association for Computational Linguistics (ACL).
- Fu, Y. (2024). Automatic Alignment of Discourse Relations of Different Discourse Annotation Frameworks. In *Proceedings of the 20th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA’24)*, pages 27–38, Torino, Italia. European Language Resources Association (ELRA) and International Conference on Computational Linguistics (COLING).
- Guan, Z., Liang, X., and Zhang, S. (2025). Enhancing Long-Term Capabilities of Large Language Models via Discourse Sub-graph Analysis. In *Proceedings of the 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP’25)*, Hyderabad, Telangana, India. Institute of Electrical and Electronics Engineers (IEEE).
- Hendrycks, D. and Gimpel, K. (2016). Gaussian Error Linear Units (GELUs). *arXiv preprint arXiv:1606.08415*.
- Hobbs, J. R. (1979). Coherence and Coreference. *Cognitive Science*, 3(1):67–90.
- Hoek, J., Scholman, M. C., and Sanders, T. J. (2021). Is there less annotator agreement when the discourse relation is underspecified? In *Proceedings of the 1st Workshop on Integrating Perspectives on Discourse Annotation*, pages 1–6, Online. Association for Computational Linguistics (ACL).
- Hoek, J. and Zufferey, S. (2015). Factors Influencing the Implication of Discourse Relations across Languages. In *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA’15)*, England, United Kingdom. Association for Computational Linguistics (ACL).
- Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning (PMLR’19)*,

- pages 2790–2799, Long Beach, California, USA. Journal of Machine Learning Research (JMLR).
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the 10th International Conference on Learning Representations (ICLR’22)*, Online.
- Hu, X. and Wan, X. (2023). Exploring Discourse Structure in Document-level Machine Translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP’23)*, pages 13889–13902, Singapore. Association for Computational Linguistics (ACL).
- Huang, Y., Fang, M., Cao, Y., Wang, L., and Liang, X. (2021). DAGN: Discourse-Aware Graph Network for Logical Reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT’21)*, pages 5848–5855, Online. Association for Computational Linguistics (ACL).
- Huang, Y. J. and Kurohashi, S. (2021). Extractive Summarization Considering Discourse and Coreference Relations based on Heterogeneous Graph. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL’21)*, pages 3046–3052, Online. Association for Computational Linguistics (ACL).
- Huber, P. and Carenini, G. (2020). From Sentiment Annotations to Sentiment Prediction through Discourse Augmentation. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING’21)*, pages 185–197, Online. International Conference on Computational Linguistics (COLING).
- Jahangir, K., Muller, P., and Braud, C. (2024). Complex question generation using discourse-based data augmentation. In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI’24)*, pages 105–119, St. Julians, Malta. Association for Computational Linguistics (ACL).
- Jayaweera, C. and Dorr, B. (2025). From Disagreement to Understanding: The Case for Ambiguity Detection in NLI. *arXiv preprint arXiv:2507.15114*.
- Ji, Y. and Eisenstein, J. (2015). One Vector is Not Enough: Entity-Augmented Distributed Semantics for Discourse Relations. *Transactions of the Association for Computational Linguistics (TACL)*, 3:329–344.

- Jiang, N.-J. and de Marneffe, M.-C. (2022). Investigating Reasons for Disagreement in Natural Language Inference. *Transactions of the Association for Computational Linguistics (TACL)*, 10:1357–1374.
- Jiang, Y., Zhang, L., and Wang, W. (2023a). Global and Local Hierarchy-aware Contrastive Framework for Implicit Discourse Relation Recognition. In *Findings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL’23)*, pages 8048–8064, Toronto, Ontario, Canada. Association for Computational Linguistics (ACL).
- Jiang, Y. E., Liu, T., Ma, S., Zhang, D., Sachan, M., and Cotterell, R. (2023b). Discourse-Centric Evaluation of Document-level Machine Translation with a New Densely Annotated Parallel Corpus of Novels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL’23)*, pages 7853–7872, Toronto, Ontario, Canada. Association for Computational Linguistics (ACL).
- Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL’20)*, pages 6282–6293, Online. Association for Computational Linguistics (ACL).
- Jurafsky, D. and Martin, J. H. (2008). *Speech and Language Processing*. Prentice Hall.
- Kim, N., Feng, S., Gunasekara, C., and Lastras, L. (2020). Implicit Discourse Relation Classification: We Need to Talk about Evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL’20)*, pages 5404–5414, Online. Association for Computational Linguistics (ACL).
- Kingma, D. P. and Ba, J. (2015). Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR’15)*, pages 1–15, San Diego, California, USA.
- Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of Machine Translation Summit X (MTSummit’05)*, pages 79–86, Phuket, Thailand.
- Kurfalı, M. and Östling, R. (2019). Zero-shot transfer for implicit discourse relation classification. In *Proceedings of the 20th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL’19)*, pages 226–231, Stockholm, Sweden. Association for Computational Linguistics (ACL).

- Lapshinova-Koltunski, E. and Carl, M. (2022). Using Translation Process Data to Explore Explicitation and Implication through Discourse Connectives. In *Proceedings of the 3rd Workshop on Computational Approaches to Discourse (CODI'22)*, pages 42–47, Gyeongju, Republic of Korea and Online. International Conference on Computational Linguistics (COLING).
- Lascarides, A. and Asher, N. (2007). *Segmented Discourse Representation Theory: Dynamic Semantics With Discourse Structure*, chapter 5, pages 87–124. Springer.
- Lei, Y. and Huang, R. (2023). Discourse Structures Guided Fine-grained Propaganda Identification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP'23)*, pages 331–342, Singapore. Association for Computational Linguistics (ACL).
- Lei, Y., Huang, R., Wang, L., and Beauchamp, N. (2022). Sentence-level Media Bias Analysis Informed by Discourse Structures. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP'22)*, pages 10040–10050, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics (ACL).
- Levisen, C. (2019). Biases we live by: Anglocentrism in linguistics and cognitive sciences. *Language Sciences*, 76(1–13).
- Li, C., Braud, C., Amblard, M., and Carenini, G. (2024a). Discourse Relation Prediction and Discourse Parsing in Dialogues with Minimal Supervision. In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI'24)*, pages 161–176, St. Julians, Malta. Association for Computational Linguistics (ACL).
- Li, C., Huber, P., Xiao, W., Amblard, M., Braud, C., and Carenini, G. (2023). Discourse Structure Extraction from Pre-Trained and Fine-Tuned Language Models in Dialogues. In *Findings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL'23)*, pages 2562–2579, Dubrovnik, Croatia. Association for Computational Linguistics (ACL).
- Li, C., Yin, Y., and Carenini, G. (2024b). Dialogue Discourse Parsing as Generation: A Sequence-to-Sequence LLM-based Approach. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL'24)*, pages 1–14, Kyoto, Japan. Association for Computational Linguistics (ACL).
- Li, M. and Huang, R. (2023). Semi-supervised News Discourse Profiling with Contrastive Learning. In *Proceedings of the 13th International Joint Conference on Natural*

- Language Processing and 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL'23)*, pages 771–781, Nusa Dua, Bali, Indonesia. Association for Computational Linguistics (ACL).
- Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.
- Lin, Z., Kan, M.-Y., and Ng, H. T. (2009). Recognizing Implicit Discourse Relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP'09)*, pages 343–351, Singapore. Association for Computational Linguistics (ACL).
- Liu, D. and Demberg, V. (2024). RST-LoRA: A Discourse-Aware Low-Rank Adaptation for Long Document Abstractive Summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'24)*, pages 2200–2220, Mexico City, Mexico. Association for Computational Linguistics (ACL).
- Liu, D., Wang, Y., and Demberg, V. (2023a). Incorporating Distributions of Discourse Structure for Long Document Abstractive Summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL'23)*, pages 5574–5590, Toronto, Ontario, Canada. Association for Computational Linguistics (ACL).
- Liu, W., Fan, Y., and Strube, M. (2023b). HITS at DISRPT 2023: Discourse Segmentation, Connective Detection, and Relation Classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT'23)*, pages 43–49, Toronto, Ontario, Canada. Association for Computational Linguistics (ACL).
- Liu, W. and Strube, M. (2023). Annotation-Inspired Implicit Discourse Relation Classification with Auxiliary Discourse Connective Generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL'23)*, pages 15696–15712, Toronto, Ontario, Canada. Association for Computational Linguistics (ACL).
- Liu, W., Wan, S., and Strube, M. (2024). What Causes the Failure of Explicit to Implicit Discourse Relation Recognition? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'24)*, pages 2738–2753, Mexico City, Mexico. Association for Computational Linguistics (ACL).

- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.
- Liu, Z., Shi, K., and Chen, N. (2020). Multilingual Neural RST Discourse Parsing. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING'21)*, pages 6730–6738, Online. International Conference on Computational Linguistics (COLING).
- Liu, Z., Shi, K., and Chen, N. (2021). DMRST: A Joint Framework for Document-Level Multilingual RST Discourse Segmentation and Parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse (CODI'21)*, pages 154–164, Online. Association for Computational Linguistics (ACL).
- Long, W., Siddharth, N., and Webber, B. (2024). Multi-Label Classification for Implicit Discourse Relation Recognition. In *Findings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL'24)*, pages 8437–8451, Bangkok, Thailand. Association for Computational Linguistics (ACL).
- Long, W. and Webber, B. (2022). Facilitating Contrastive Learning of Discourse Relational Senses by Exploiting the Hierarchy of Sense Relations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP'22)*, pages 10704–10716, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics (ACL).
- Long, W. and Webber, B. (2024). Leveraging Hierarchical Prototypes as the Verbalizer for Implicit Discourse Relation Recognition. *arXiv preprint arXiv:2411.14880*.
- López Cortez, S. M. and Jacobs, C. L. (2023). Incorporating Annotator Uncertainty into Representations of Discourse Relations. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL'23)*, pages 530–537, Prague, Czechia. Association for Computational Linguistics (ACL).
- Loshchilov, I. and Hutter, F. (2017). SGDR: Stochastic Gradient Descent with Warm Restarts. In *Proceedings of the 5th International Conference on Learning Representations (ICLR'17)*, Toulon, France.
- Mann, W. C. and Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

- Marchal, M., Scholman, M., Yung, F., and Demberg, V. (2022). Establishing Annotation Quality in Multi-label Annotations. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING’22)*, pages 3659–3668, Gyeongju, Republic of Korea. International Conference on Computational Linguistics (COLING).
- Marcu, D. (2000). *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press.
- Marcu, D., Amorrortu, E., and Romera, M. (1999). Experiments in Constructing a Corpus of Discourse Trees. In *Proceedings of Workshop on Towards Standards and Tools for Discourse Tagging*, pages 48–57, College Park, Maryland, USA. Association for Computational Linguistics (ACL).
- Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., and Sagot, B. (2020). CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL’20)*, pages 7203–7219, Online. Association for Computational Linguistics (ACL).
- Mehndiratta, A. and Asawa, K. (2025). Enhancing Dialogue Systems with Discourse-Level Understanding Using Deep Canonical Correlation Analysis. *arXiv preprint arXiv:2504.09094*.
- Meissner, J. M., Thumwanit, N., Sugawara, S., and Aizawa, A. (2021). Embracing Ambiguity: Shifting the Training Target of NLI Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP’21)*, pages 862–869, Online. Association for Computational Linguistics (ACL).
- Metheniti, E., Braud, C., Muller, P., and Rivière, L. (2023). DisCut and DiscReT: MELODI at DISRPT 2023. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT’23)*, pages 29–42, Toronto, Ontario, Canada. Association for Computational Linguistics (ACL).
- Miltsakaki, E., Prasad, R., Joshi, A., and Webber, B. (2004). The Penn Discourse Treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC’04)*, pages 2237–2240, Lisbon, Portugal. European Language Resources Association (ELRA).

- Miyato, T., Dai, A. M., and Goodfellow, I. (2017). Adversarial Training Methods for Semi-Supervised Text Classification. In *Proceedings of the 5th International Conference on Learning Representations (ICLR’17)*, Toulon, France.
- Nie, Y., Zhou, X., and Bansal, M. (2020). What Can We Learn from Collective Human Opinions on Natural Language Inference Data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP’20)*, pages 9131–9143, Online. Association for Computational Linguistics (ACL).
- Omura, K., Cheng, F., and Kurohashi, S. (2024). An Empirical Study of Synthetic Data Generation for Implicit Discourse Relation Recognition. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING’24)*, pages 1073–1085, Torino, Italia. European Language Resources Association (ELRA) and International Conference on Computational Linguistics (COLING).
- OpenAI (2023). GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.
- OpenAI (2024). GPT-4o System Card. *arXiv preprint arXiv:2410.21276*.
- Pavlick, E. and Kwiatkowski, T. (2019). Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics (TACL)*, 7:677–694.
- Plank, B. (2022). The “Problem” of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP’22)*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics (ACL).
- Polakova, L., Mírovský, J., and Synková, P. (2017). Signalling implicit relations: A PDTB - RST comparison. *Dialogue & Discourse*, 8(2):225–248.
- Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse TreeBank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC’08)*, pages 2961–2968, Marrakech, Morocco. European Language Resources Association (ELRA).
- Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., Robaldo, L., and Webber, B. (2007). The Penn Discourse Treebank 2.0 Annotation Manual. Technical report, University of Pennsylvania.

- Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., and Webber, B. (2006). The Penn Discourse Treebank 1.0 Annotation Manual. Technical report, University of Pennsylvania.
- Pyatkin, V., Klein, A., Tsarfaty, R., and Dagan, I. (2020). QADiscourse - Discourse Relations as QA Pairs: Representation, Crowdsourcing and Baselines. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP'20)*, pages 2804–2819, Online. Association for Computational Linguistics (ACL).
- Pyatkin, V., Yung, F., Scholman, M. C. J., Tsarfaty, R., Dagan, I., and Demberg, V. (2023). Design Choices for Crowdsourcing Implicit Discourse Relations: Revealing the Biases Introduced by Task Design. *Transactions of the Association for Computational Linguistics (TACL)*, 11:1014–1032.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research (JMLR)*, 21(140):1–67.
- Ranathunga, S. and de Silva, N. (2022). Some Languages are More Equal than Others: Probing Deeper into the Linguistic Disparity in the NLP World. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (AACL-IJCNLP'22)*, pages 823–848, Online. Association for Computational Linguistics (ACL).
- Rehbein, I., Scholman, M., and Demberg, V. (2016). Annotating Discourse Relations in Spoken Language: A Comparison of the PDTB and CCR Frameworks. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*, pages 1039–1046, Portorož, Slovenia. European Language Resources Association (ELRA).
- Rohde, H., Dickinson, A., Schneider, N., Clark, C. N. L., Louis, A., and Webber, B. (2016). Filling in the Blanks in Understanding Discourse Adverbials: Consistency, Conflict, and Context-Dependence in a Crowdsourced Elicitation Task. In *Proceedings of the 10th Linguistic Annotation Workshop (LAW'16)*, pages 49–58, Berlin, Germany. Association for Computational Linguistics (ACL).
- Ru, D., Qiu, L., Qiu, X., Zhang, Y., and Zhang, Z. (2023). Distributed Marker Representation for Ambiguous Discourse Markers and Entangled Relations. In *Proceedings*

- of the 61st Annual Meeting of the Association for Computational Linguistics (ACL'23), pages 5334–5351, Toronto, Ontario, Canada. Association for Computational Linguistics (ACL).
- Saeed, M., Bourgonje, P., and Demberg, V. (2025). Implicit Discourse Relation Classification For Nigerian Pidgin. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING'25)*, pages 2561–2574, Abu Dhabi, UAE. International Conference on Computational Linguistics (COLING).
- Sanders, T. J., Demberg, V., Hoek, J., Scholman, M. C., Asr, F. T., Zufferey, S., and Evers-Vermeul, J. (2021). Unifying dimensions in coherence relations: How various annotation frameworks are related. *Corpus Linguistics and Linguistic Theory*, 17(1):1–71.
- Sanders, T. J., Spooren, W. P., and Noordman, L. G. (1992). Toward a taxonomy of coherence relations. *Discourse Processes*, 15(1):1–35.
- Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Scheffler, T. and Stede, M. (2016). Mapping PDTB-style connective annotation to RST-style discourse annotation. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS'16)*, pages 242–247, Bochum, Germany. Bochumer Linguistische Arbeitsberichte.
- Scholman, M. and Demberg, V. (2017). Examples and Specifications that Prove a Point: Identifying Elaborative and Argumentative Discourse Relations. *Dialogue & Discourse*, 8(2):56–83.
- Scholman, M., Dong, T., Yung, F., and Demberg, V. (2022a). DiscoGeM: A Crowdsourced Corpus of Genre-Mixed Implicit Discourse Relations. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC'22)*, pages 3281–3290, Marseille, France. European Language Resources Association (ELRA).
- Scholman, M., Pyatkin, V., Yung, F., Dagan, I., Tsarfaty, R., and Demberg, V. (2022b). Design Choices in Crowdsourcing Discourse Relation Annotations: The Effect of Worker Selection and Training. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC'22)*, pages 2148–2156, Marseille, France. European Language Resources Association (ELRA).

- Scholman, M. C., Marchal, M., Brown, A., and Demberg, V. (2025). DiscoNaija: a discourse-annotated parallel Nigerian Pidgin-English corpus. *Language Resources and Evaluation*, 59(2):1–37.
- Shi, W. and Demberg, V. (2017). On the Need of Cross Validation for Discourse Relation Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL’17)*, pages 150–156, Valencia, Spain. Association for Computational Linguistics (ACL).
- Shi, W. and Demberg, V. (2019). Next Sentence Prediction helps Implicit Discourse Relation Classification within and across Domains. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP’19)*, pages 5790–5796, Hong Kong, China. Association for Computational Linguistics (ACL).
- Sluyter-Gäthje, H., Bourgonje, P., and Stede, M. (2020). Shallow Discourse Parsing for Under-Resourced Languages: Combining Machine Translation and Annotation Projection. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC’20)*, pages 1044–1050, Marseille, France. European Language Resources Association (ELRA).
- Stede, M. (2008). Disambiguating Rhetorical Structure. *Research on Language and Computation*, 6(3):311–332.
- Stede, M. and Neumann, A. (2014). Potsdam Commentary Corpus 2.0: Annotation for Discourse Research. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC’14)*, pages 925–929, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Taboada, M. and Mann, W. C. (2006). Rhetorical Structure Theory: Looking back and moving ahead. *Discourse Studies*, 8(3):423–459.
- Tomaszewska, A., Silvano, P., Leal, A., and Amorim, E. (2024). ISO 24617-8 Applied: Insights from Multilingual Discourse Relations Annotation in English, Polish, and Portuguese. In *Proceedings of the 20th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA’24)*, pages 99–110, Torino, Italia. European Language Resources Association (ELRA) and International Conference on Computational Linguistics (COLING).

- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.
- Uma, A. N., Fornaciari, T., Hovy, D., Paun, S., Plank, B., and Poesio, M. (2021). Learning from Disagreement: A Survey. *Journal of Artificial Intelligence Research*, 72(1):1385–1470.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention Is All You Need. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS’17)*, Long Beach, California, USA. Curran Associates, Inc.
- Wang, L., Lyu, C., Ji, T., Zhang, Z., Yu, D., Shi, S., and Tu, Z. (2023). Document-Level Machine Translation with Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP’23)*, pages 16646–16661, Singapore. Association for Computational Linguistics (ACL).
- Warner, B., Chaffin, A., Clavié, B., Weller, O., Hallström, O., Taghadouini, S., Gallagher, A., Biswas, R., Ladhak, F., Aarsen, T., Cooper, N., Adams, G., Howard, J., and Poli, I. (2024). Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. *arXiv preprint arXiv:2412.13663*.
- Webber, B., Prasad, R., Lee, A., and Joshi, A. (2019). The Penn Discourse Treebank 3.0 Annotation Manual. Technical report, University of Pennsylvania.
- Wu, H., Zhou, H., Lan, M., Wu, Y., and Zhang, Y. (2023). Connective Prediction for Implicit Discourse Relation Recognition via Knowledge Distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL’23)*, pages 5908–5923, Toronto, Ontario, Canada. Association for Computational Linguistics (ACL).
- Xu, F., Li, J. J., and Choi, E. (2022). How Do We Answer Complex Questions: Discourse Structure of Long-form Answers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL’22)*, pages 3556–3572, Dublin, Ireland. Association for Computational Linguistics (ACL).

- Xu, J., Jiang, F., Gao, A., Fernando D’Haro, L., and Li, H. (2024). Unsupervised Mutual Learning of Dialogue Discourse Parsing and Topic Segmentation in Dialogue. *arXiv preprint arXiv:2405.19799*.
- Yung, F., Ahmad, M., Scholman, M., and Demberg, V. (2024a). Prompting Implicit Discourse Relation Annotation. In *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 150–165, St. Julian’s, Malta. Association for Computational Linguistics (ACL).
- Yung, F., Anuranjana, K., Scholman, M., and Demberg, V. (2022). Label distributions help implicit discourse relation classification. In *Proceedings of the 3rd Workshop on Computational Approaches to Discourse (CODI’22)*, pages 48–53, Gyeongju, Republic of Korea. International Conference on Computational Linguistics (COLING).
- Yung, F. and Demberg, V. (2025). On Crowdsourcing Task Design for Discourse Relation Annotation. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 12–19, Abu Dhabi, UAE. International Conference on Computational Linguistics (COLING).
- Yung, F., Demberg, V., and Scholman, M. (2019). Crowdsourcing Discourse Relation Annotations by a Two-Step Connective Insertion Task. In *Proceedings of the 13th Linguistic Annotation Workshop (LAW’19)*, pages 16–25, Florence, Italy. Association for Computational Linguistics (ACL).
- Yung, F., Scholman, M., Lapshinova-Koltunski, E., Pollkläsener, C., and Demberg, V. (2023). Investigating Explicitation of Discourse Connectives in Translation using Automatic Annotations. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL’23)*, pages 21–30, Prague, Czechia. Association for Computational Linguistics (ACL).
- Yung, F., Scholman, M., Zikanova, S., and Demberg, V. (2024b). DiscoGeM 2.0: A Parallel Corpus of English, German, French and Czech Implicit Discourse Relations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING’24)*, pages 4940–4956, Torino, Italia. European Language Resources Association (ELRA) and International Conference on Computational Linguistics (COLING).
- Yung, F., Suresh, V., Reza, Z., Ahmad, M., and Demberg, V. (2025). Synthetic Data Augmentation for Cross-domain Implicit Discourse Relation Recognition. *arXiv preprint arXiv:2503.20588*.

- Zeldes, A., Aoyama, T., Liu, Y. J., Peng, S., Das, D., and Gessler, L. (2025). eRST: A Signaled Graph Theory of Discourse Relations and Organization. *Computational Linguistics (CL)*, 51(1):23–72.
- Zeldes, A., Das, D., Maziero, E. G., Antonio, J., and Iruskieta, M. (2019). The DISRPT 2019 Shared Task on Elementary Discourse Unit Segmentation and Connective Detection. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking (DISRPT’19)*, pages 97–104, Minneapolis, Minnesota, USA. Association for Computational Linguistics (ACL).
- Zeldes, A., Liu, Y. J., Iruskieta, M., Muller, P., Braud, C., and Badene, S. (2021). The DISRPT 2021 Shared Task on Elementary Discourse Unit Segmentation, Connective Detection, and Relation Classification. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT’21)*, pages 1–12, Online. Association for Computational Linguistics (ACL).
- Zeng, L., He, R., Sun, H., Xu, J., Liu, C., and Wang, B. (2024). Global and Local Hierarchical Prompt Tuning Framework for Multi-level Implicit Discourse Relation Recognition. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING’24)*, pages 7760–7773, Torino, Italia. European Language Resources Association (ELRA) and International Conference on Computational Linguistics (COLING).
- Zeyrek, D., Mendes, A., Grishina, Y., Kurfalı, M., Gibbon, S., and 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.
- Zeyrek, D., Mendes, A., Oleškevičienė, G. V., and Özer, S. (2022). An Exploratory Analysis of TED Talks in English and Lithuanian, Portuguese and Turkish Translations. *Contrastive Pragmatics*, 3(3):452–479.
- Zhang, H., Liu, X., and Zhang, J. (2023). Contrastive Hierarchical Discourse Graph for Scientific Document Summarization. In *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI’23)*, pages 37–47, Toronto, Ontario, Canada. Association for Computational Linguistics (ACL).
- Zhao, H., He, R., Xiao, M., and Xu, J. (2023). Infusing Hierarchical Guidance into Prompt Tuning: A Parameter-Efficient Framework for Multi-level Implicit Discourse Relation

- Recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL'23)*, pages 6477–6492, Toronto, Ontario, Canada. Association for Computational Linguistics (ACL).
- Zhong, Y. and Litman, D. (2025). Discourse-Driven Evaluation: Unveiling Factual Inconsistency in Long Document Summarization. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'25)*, pages 2050–2073, Albuquerque, New Mexico. Association for Computational Linguistics (ACL).
- Zufferey, S. and Degand, L. (2024). *Discourse Relations and Connectives across Languages and Genres*, chapter 7, pages 142–162. Cambridge University Press.