Discourse Analysis of Argumentative Essays of English Learners based on their CEFR Level

Blaise Hanel

A Thesis

in

The Department

of

Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Computer Science at

Concordia University

Montréal, Québec, Canada

July 2023

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Blaise Hanel**

Entitled: **Discourse Analysis of Argumentative Essays of English Learners based on their CEFR Level**

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

Master of 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. Juergen Rilling

______________________________ Examiner
Dr. Essam Mansour

______________________________ Supervisor
Dr. Leila Kosseim

Approved by ______________________________
Joey Paquet, Chair
Department of Computer Science and Software Engineering

__________ 2023

____
Mourad Debabbi, Dean
Faculty of Engineering and Computer Science

Abstract

Discourse Analysis of Argumentative Essays of English Learners based on their CEFR Level

Blaise Hanel

This thesis aims to explore the relationship between discourse information and the CEFR-level (Common European Framework of Reference for Languages) in argumentative English learner essays. The study leverages two prominent frameworks: the Rhetorical Structure Theory (RST) and the Penn Discourse TreeBank (PDTB), to analyze essays obtained from The International Corpus Network of Asian Learners (ICNALE) and the Corpus and Repository of Writing (CROW). The research investigates the influence of different discourse relations and connectives on the language proficiency level of the writers, and further explores the potential of using discourse information as additional features for automated CEFR-level determination.

The analysis of the collected essays reveals significant findings regarding the utilization of discourse relations by English learners. Notably, the RST relations of EXPLANATION and BACKGROUND are statistically used more often by writers with a CEFR level below fluency. In addition, as the CEFR level increases, the use of the PDTB relation of CONTINGENCY decreases. These results provide empirical evidence of the relationship between discourse relations and language proficiency, highlighting the differential usage patterns among learners at various CEFR levels.

To validate these findings computationally, discourse relations and connectives are employed as supplementary features for machine learning models. The experimental results indicate that incorporating discourse information into the automated CEFR-level determination process leads to a mild increase in performance compared to relying solely on lexical and grammatical features. However, it is important to note that the proposed approach does not outperform the use of large language models, such as RoBERTa, which have demonstrated superior performance in various natural language processing tasks.

Nevertheless, this study contributes valuable insights into the relationship between discourse relations and argumentative English learner essays. The findings highlight the potential influence of discourse relations on language proficiency and suggest avenues for further research and development in language assessment methodologies.

Acknowledgments

I feel incredibly grateful for all the wonderful people I've come across who have supported me, not only academically, but during my time in Montréal as a whole. This work would not have been possible without them.

Fellow members of the CLaC lab: Andrei Neagu, Nelson Filipe Costa, and Shamanth Nayak, have given me invaluable help and feedback through all the dry-runs leading up to my final thesis defence. Past lab members, notably Thomas Chapados Muermans, Andrés Lou, Pavel Khloponin, Farhood Farahnak, and Parsa Bagherzadeh gave me help and support as I navigated through my first year of coursework. I would like to express my appreciation to Dr. Juergen Rilling and Dr. Essam Mansour for taking the time to review my thesis and examine my thesis defence.

Je veux dire un grand merci à Montréal International pour m'avoir accepté dans le programme *Nous travaillons en Français*. Je remercie spécifiquement ma professeure de français Anne-Marie Rivard et mon mentor Maxime Daigle pour tout leur soutien.

I am incredibly grateful for my counsellor, Marc Leger, who helped me navigate my experience in Montréal. I am also grateful for my lovely friends of the C Squad, sending me love and support from down in the states: Danyell Monk, Logan Hebert, Jess Fernandes, and Tess Friedel. I am grateful for my friend/roommate Jasmine Louisy, and for my dear friend Ewelina Nierzwicka, who, in addition to giving me support throughout my time working on this thesis, took the time to listen to my thesis presentation and gave me feedback. I would be remiss not to mention Sara Ouarda, whose kind heart drove me to keep working throughout this year, and I'm very grateful for all of their help prior to my defence, from listening to my dry-runs to helping me go through my flashcards.

. . . and I think that's everyone

just kidding :)

Of course, I can not thank Dr. Leila Kosseim enough for all the help she's given me. So I'll at least do it twice. Thank you, thank you! She took a chance on me and accepted me into this program, and provided endless support, patience, and knowledge throughout my coursework and thesis work. Dr. Kosseim meticulously looked through my 30-some iterations of my thesis, as well as several iterations for my RANLP and NLDB paper submissions, to ensure everything was as strong as it could be. I certainly wouldn't be here today, writing this acknowledgement on my completed thesis, if not for her.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In a world where over 7,000 languages are used, much research has focused on improving methods to teach and learn natural languages. In particular, the field of Natural Language Processing (NLP) has a long history of developing tools to assist language learners and reduce learning barriers. Previous works on computational discourse analysis, such as Webber (2009), Bachand, Davoodi, and Kosseim (2014), and Abdalla, Rudzicz, and Hirst (2018) have shown significant differences in discourse usage across textual genre and simplicity level. However, to our knowledge, very few studies have investigated the relationship between discourse structures and language learning. This thesis attempts to fill this gap by investigating the usage of discourse relations and connectives to identify trends in argumentative texts written by English learners across various proficiency levels.

1.1 Goal of the Thesis

Over the past few decades, the field of Natural Language Processing (NLP) has made remarkable progress in various areas such as language translation, sentiment analysis, and text classification. One of the most exciting areas of NLP research has been the development of computational tools to assess the quality of student essays written by English Language Learners (ELL)[1].

The importance of assessing the quality of ELL essays cannot be overstated, as it plays a critical

[1] In this thesis, we will frequently use the term ELL as coined by Lacelle-Peterson and Rivera (1994), as commonly used terms ESL (English as a Second Language) and L2 imply that English is the second language of the learners, which may not always be the case.

role in improving students' writing skills and enhancing their language proficiency. Traditional assessment methods have relied on human graders, which is time-consuming, expensive, and prone to subjectivity. On the other hand, NLP methods have the potential to automate the assessment process, providing an objective and efficient way of evaluating student essays.

Rhetorical Structure Theory (RST) and Penn Discourse TreeBank (PDTB) are two important frameworks that have been used to analyze the discourse structure and coherence of texts. RST aims to identify the rhetorical relations that connect different parts of a text, such as CAUSE-EFFECT, ELABORATION, and CONTRAST, while PDTB focuses on the identification of explicit and implicit discourse connectives (e.g. *but, if... then, because*) and labelling the relation they convey (e.g. COMPARISON.CONTRAST, CONTINGENCY.CONDITION, and CONTINGENCY.CAUSE). By leveraging these frameworks, NLP methods can be used to empirically analyze the difference in usage of discourse information in student essays of ELL.

The goal of this thesis is to investigate the usage of discourse relations and connectives to discover trends in their usage in argumentative texts by English learners across various proficiency levels. Specifically, we investigate the usage of discourse relations using the Rhetorical Structure Theory (Mann & Thompson, 1988) and the Penn Discourse TreeBank (Prasad et al., 2008) frameworks, as well as discourse connectives from the Penn Discourse TreeBank, to discover trends in their usage in argumentative texts by English learners across various proficiency levels. Ultimately, this thesis aims to contribute to ongoing efforts to develop more accurate and effective NLP-based tools for assessing ELL essays, thereby enhancing the quality of education for ELL students.

1.2 Motivation

Corpus research on the use of discourse structures across different CEFR levels can provide valuable insights into language learners' ability to organize and convey their ideas in written or spoken language. This analysis can also identify common patterns of language use that prove challenging for learners at different CEFR levels, thereby facilitating the development of more effective teaching materials and strategies tailored to learners' specific needs (Aoyama, 2022). Additionally, it can help reduce the workload of human graders (Mieskes & Padó, 2018). Findings can also inform the development of more reliable assessment tools that accurately measure learners' proficiency in the use of discourse structures. Accurate assessment is essential for learners to identify their strengths and weaknesses and make informed decisions about their language learning goals and strategies.

With the rise of Generative AI models such as ChatGPT[2], there is a growing need for more intelligent and context-aware AI chatbots that can effectively communicate with users across different language proficiency levels. While existing chatbot systems offer valuable assistance, they often lack the ability to accurately gauge and adapt to the user's language skills, resulting in communication gaps and potential frustration. By analyzing the discourse structure changes in essays written by English learners, we can gain valuable insights into the linguistic progress from novice to fluency, which can be harnessed to develop improved language assessment models for AI chatbots.

Tyen, Brenchley, Caines, and Buttery (2022) discuss methods to adjust the difficulty of chatbot generated responses based on the Common European Framework of Reference (CEFR) levels. They propose various decoding strategies that take into account vocabulary restriction, re-ranking of candidate messages, and sub-token penalties, among others. While their focus is on adjusting the generated text difficulty, our thesis expands upon their work by exploring the underlying changes in discourse structures that occur as English learners progress through different proficiency levels, allowing for the possibility of a more automatic system to determine when difficulty adjustment is necessary.

1.3 Methodology

The goal of this thesis is to investigate differences in the usage of discourse relations and connectives between English learner and English native texts. In order to achieve this goal, we used the following methodology. First, we parsed essays from different CEFR levels with four discourse parsers (two from RST-DT and two from PDTB). Then, we used the agreement between the two

[2]https://openai.com/blog/chatgpt

RST parsers to determine the frequency of RST relations among different levels of English proficiency. Using the agreement between the two PDTB parsers, we repeated this step for PDTB level 1 relations, and for discourse connectives. Then, we used this information to create features for Random Forest, Support Vector Machine, and Logistic Regression models to classify learner texts. A flowchart of our full methodology, starting with the raw text from the datasets, is shown in Figure 1.1.

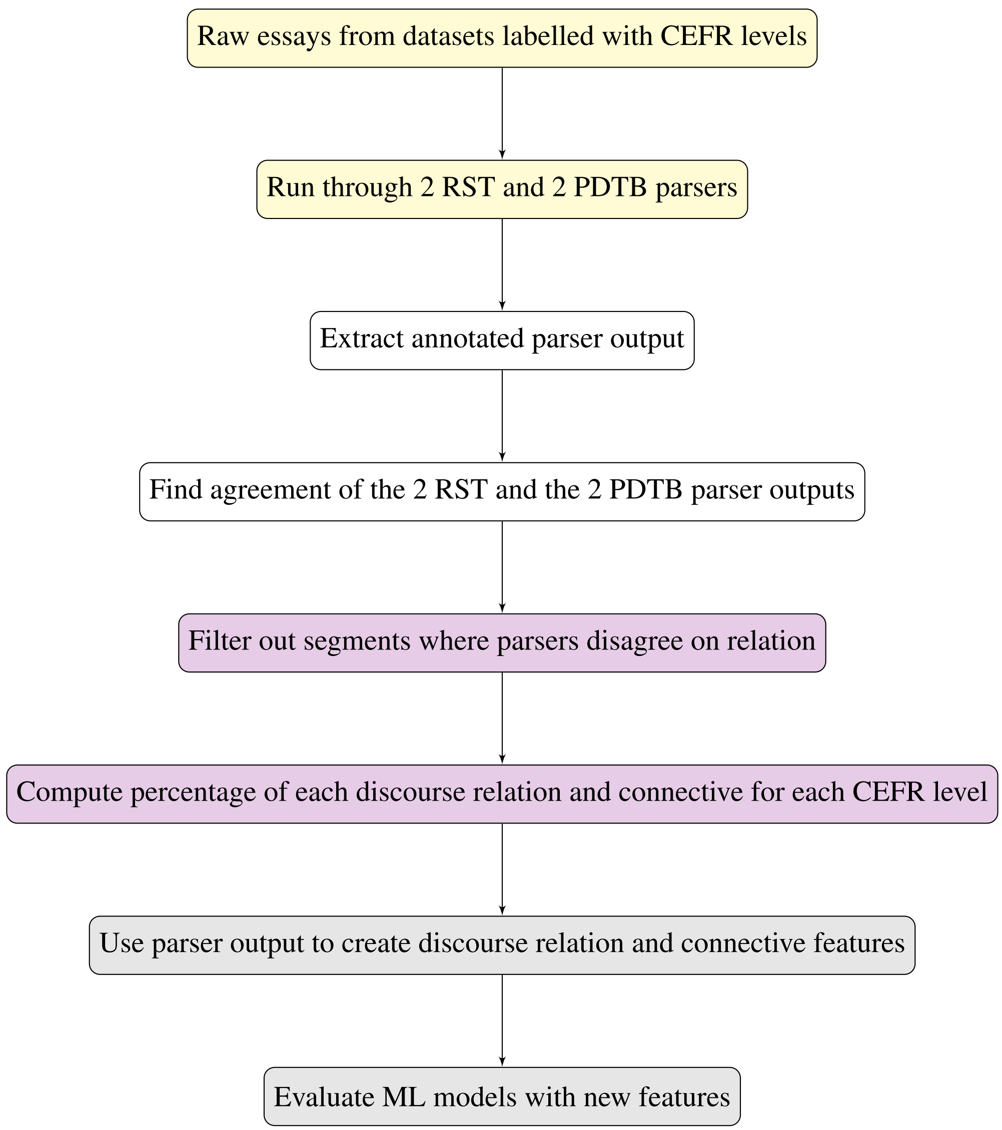

Figure 1.1: Flowchart showing our methodology, beginning with raw text from English learner datasets.

1.4 Contributions

This thesis makes several contributions to the field of Natural Language Processing (NLP) and language assessment for English learners:

(1) **Discourse Analysis of English Learner Essays:** In Chapter 3, we provide a comprehensive analysis of discourse structures in English learner essays, focusing on the Rhetorical Structure Theory (RST) and Penn Discourse TreeBank (PDTB) frameworks. By examining the usage of discourse relations and connectives across CEFR levels, we identify patterns and trends that are indicative of different levels of language proficiency.

(2) **Discourse Parser Agreement:** In Section 3.2.2 and Section 3.3.1, we investigate the agreement between commonly-used PDTB and RST parsers. This could provide useful insight to future users of these discourse parsers.

(3) **Insights into Language Learning Challenges:** Our analysis of RST and PDTB discourse relation patterns in Chapter 3 uncovers specific discourse patterns that pose challenges for English learners at different proficiency levels. This information can inform the development of targeted teaching materials and strategies to address these challenges and enhance language learning outcomes.

(4) **Automated Language Assessment:** In Chapter 4, we explore the feasibility of using discourse relations and connectives as features for machine learning models to automatically assess proficiency levels of English learner essays. Our experiments with Random Forest, Support Vector Machine, and Logistic Regression models measured the potential of these features to enhance accurate and efficient language assessment. While, in recent years, this assessment has begun to be taken over by Large Language Models (LLMs), and has yielded effective results on automated essay scoring (Naismith, Mulcaire, & Burstein, 2023), traditional machine learning models remain relevant for this task. Firstly, LLMs are computationally expensive and require significant computational resources to train and deploy. Thus, systems with limited computing power must rely on lower-cost models, such as traditional

machine learning models, to complete tasks. Additionally, applications of traditional machine learning models can be used as baselines for the further study of LLMs.

On a more general point of view, by analyzing the discourse structure changes in essays written by English learners, we contribute to the development of more intelligent and context-aware AI chatbots. Our findings can be used to create language assessment models that adapt to users' proficiency levels, resulting in improved communication and user experience.

Through these contributions, this thesis aims to advance the field of NLP, facilitate language learning and assessment, and contribute to the development of more effective and intelligent language technologies.

1.5 Thesis Structure

This chapter presented our motivation to work on discourse analysis for language learning and our contributions to the field of NLP. Chapter 2 will provide an overview of previous work and the required background related to our research. Chapter 3 will provide an in-depth discussion of our discourse analysis process and results. Chapter 4 will discuss the methodology and results of our work to use machine learning to empirically validate the use of the findings of Chapter 3. Finally, Chapter 5 will summarize the thesis and present directions for future work.

Chapter 2

Literature Review

This chapter provides an overview of previous work and background information related to our research.

Section 2.1 covers linguistic background, which includes language proficiency scores and text complexity metrics. We discuss how these measures can be used to analyze texts and how they can aid in the development of language models. This section also discusses the assessment of textual complexity, reviewing previous work that analyzes the language and structure of a text to determine its level of difficulty

Section 2.2 gives an overview of discourse analysis, including a description of the most important framework; Rhetorical Structure Theory and Discourse Lexicalized Tree-Adjoining Grammar. These frameworks help identify and analyze the discourse structure of a text, including the relationships between different textual spans. The section also describes the mapping of discourse relations across different frameworks. This is important because different frameworks have different ways of representing discourse relations, and mapping them helps identify similarities and differences between the frameworks. This section also reviews classic work on computational discourse analysis, in which past research has found differences in discourse features in a wide variety of texts.

Finally, Section 2.3 discusses machine learning classifiers, which are used to classify texts based on different criteria. We cover transformer-based methods and classic models, which are widely used in natural language processing.

2.1 Assessment of Textual Complexity and Proficiency

In this section, we describe standard scores and formulae used to grade complexity and proficiency of English texts. We then discuss previous corpus research aimed to assign these scores to texts automatically.

2.1.1 Language Proficiency Scores

To assess language proficiency, several measures have been developed. In particular, the Common European Framework of Reference for Languages (CEFR), and the Test of English as a Foreign Language (TOEFL).

CEFR defines six proficiency reference levels: A1, A2, B1, B2, C1, and C2, which represent a progression from basic understanding of a language (A1) to full fluency (C2). Each level of the CEFR provides a general description of what a learner should be able to accomplish to achieve that level, in terms of writing, reading, speaking, and listening proficiency. The TOEFL score, meanwhile, is given to a language learner as a result of taking an official test in English. The test consists of four sections, one of which involves writing an essay based on a reading passage, or based on opinions and personal experiences. A score between 0 (low proficiency) and 120 (full fluency) is given.

The CEFR and TOEFL levels have become standards to evaluate English proficiency, and several datasets of texts have been labelled with these measures. To facilitate their interoperability, in 2010, the Educational Testing Service (ETS), a non-profit organization that develops and administers standardized tests, proposed a metric[1] for mapping TOEFL scores directly to CEFR levels. Figure 2.1 shows this mapping as used by the International Corpus Network of Asian Learners of English (Ishikawa, 2013) (see Section 3.1)[2]. C2 levels are not listed, as only ELLs had taken a TOEFL test.

Essays in the datasets that we used (see Section 3.1) are labelled with both TOEFL and CEFR scores in order to assess the English language proficiency of the reader. These scores will be used

[1] https://language.sakura.ne.jp/icnale/images/about/toefl_mapping.pdf

[2] The International Corpus Network of Asian Learners of English separates B1 into a lower and upper level (B1_1 and B1_2). For our analysis, we will combine them into a singular B1 label.

CEFR Level	TOEFL Internet-based Test
A2 (Waystage)	-56
B1_1 (Threshold: Lower)	57+
B1_2 (Threshold: Upper)	72+
B2+ (Vantage or higher)	87+

Table 2.1: Mapping of CEFR levels to TOEFL scores according to the Educational Testing Service.

to analyze and evaluate these texts based on the writer's language proficiency.

2.1.2 Text Complexity Metrics

Text Complexity Metrics have been created in order to convert surface-level information such as word and character counts into a metric that can be used to give a general complexity and readability score to a text. These readability scores are generally known to be imperfect, but are designed to be simple and fast to calculate. These scores can be used as a baseline, and in conjunction with more informative features, for machine learning classification of learner texts (see Section 4). The main formulae are described below.

Flesch Reading Ease (Flesch, 1943) relies entirely on the total number of words, sentences, and syllables per word. The formula for computing Flesch Reading Ease is shown below:

$$FRE = 206.835 - 1.015\left(\frac{total\ words}{total\ sentences}\right) - 84.6\left(\frac{total\ syllables}{total\ words}\right) \tag{1}$$

The possible values in the Flesch Reading Ease score range from 0 to 100. A higher score indicates a text which is easier to understand. However, to some, a number sans context between 0 and 100 means little at first glance. Using this reasoning, J. Kincaid and a team working with the United States Army created the **Flesch–Kincaid Grade Level** (or Flesch-Kincaid Index) (Kincaid, Fishburne, Rogers, & Chissom, 1975). Similarly to Flesch Reading Ease, this score relies entirely on the total number of words, sentences, and syllables per word, however the scale is limited to a score of 0 to 18:

$$FKGL = 0.39\left(\frac{total\ words}{total\ sentences}\right) + 11.8\left(\frac{total\ syllables}{total\ words}\right) - 15.59 \tag{2}$$

This creates a scale where the readability score is meant to map to the average reading ability of a student in US grade level 0-12, as well as up to six years of post-secondary study in the topic of the essay. Though more factors other than word length, sentence length, and syllables should be considered, this index is generally accepted as an official framework for text readability in legal contexts. In some US States, insurance policies cannot exceed a certain Flesch-Kincaid Index (Texas Department of Insurance, 1992) in order to ensure that clients can understand the policies.

The **Dale-Chall Readability Score** took inspiration from the Flesch Reading Ease, but disregarded the assumption that a word with more syllables is always more complex. Dale and Chall (1948) argued that the original formula of Flesch (1943) did not efficiently predict the readability of certain types of words, such as affixes and proper nouns. In the original version of the formula, Dale and Chall used a list called the *Dale List*, containing 769 words that were understood by at least 80 percent of 4th graders. Any word outside this list was considered a difficult word, the count of which was applied to this formula:

$$DCRS = 0.1579\left(\frac{difficult\ words}{words}\right) + 0.0496\left(\frac{words}{sentences}\right) \tag{3}$$

The **Gunning Fog Index** was developed by businessman Robert Gunning in 1952 (Gunning, 1952). The formula takes a similar idea of complex words being 3 or more syllables, yet adds on the idea that proper nouns and compound words should not be counted as complex. In addition, common suffixes such as *-es, -ed,* and *-ing* are ignored in syllable counts. The equation relies solely on words, sentence count, and complex words that do not fall into the above categories:

$$GFI = 0.4\left(\frac{words}{sentences} + 100\left(\frac{complex\ words}{words}\right)\right) \tag{4}$$

The **Linsear Write Formula** (sometimes also called the Lensear Write Formula) was first proposed in the incredibly-named book *Gobbledygook Has Got To Go* (O'Hayre, 1966). This formula is stated to be concerned not so much with the reader as with the writer. In the version of the formula used by the Textstat library[3], the equation is as follows, where hard words are defined as words with

[3] `https://pypi.org/project/textstat/`

3 or more syllables, and easy words are defined as words with 2 or fewer syllables.

$$r = \frac{easy\ words + 3(complex\ words)}{sentences} \tag{5}$$

$$LWF = \begin{cases} r > 20 & \frac{r}{2} \\ r <= 20 & \frac{r}{2} - 1 \end{cases}$$

The Linsear Write Formula once again gives a score equivalent to a US grade level, reflecting the estimated years of education needed to read the text fluently.

The **Automated Readability Index** (Smith & Senter, 1967) and the **Coleman Liau Index** (Coleman & Liau, 1975) both rely on characters per word, rather than syllables. In creating the Automated Readability Index, Smith and Senter (1967) gave a random passage to a selection of 65 college students, asking them to count the number of syllables in the passage. The students returned a mean character count with a standard deviation of 17.52, showing that there is discrepancy, even among human annotators, regarding syllable count. The Coleman-Liau Index was created for mechanical scanners, so a simple method was needed that did not require knowledge of syllabification. The output of both indices is similar to that of the Flesch-Kincaid Index, in which the value corresponds to a US Grade Level.

The Automated Readability Index is computed as:

$$ARI = 4.71\left(\frac{characters}{words}\right) + 0.5\left(\frac{words}{sentences}\right) - 21.43 \tag{6}$$

The Coleman-Liau Index is:

$$CLI = 0.0588\left(\frac{characters}{words}\right) + 0.296\left(\frac{sentences}{words}\right) - 15.8 \tag{7}$$

Finally, a simple text complexity metric is the **SMOG Index** (Simple Measure Of Gobbledygook) (McLaughlin, 1969). In its simplest form, this index only takes into account the number of polysyllabic words (defined as 3 or more syllables) in a 30-sentence sample from the text, takes the square root of this number, and adds 3. This index is considered the easiest to be computed by mental math, and is generally found most useful for medical documents.

$$SMOG = \sqrt{polysyllabic\ words} + 3 \tag{8}$$

These readability scores, though naïve, have been used as a baseline for CEFR-level classification (Montgomerie, 2021), and will be combined with more informative features for machine learning classifiers (see Section 4.1).

2.1.3 Corpus Research in Assessment of Textual Complexity and Proficiency

Previous works have used machine learning to assess learner texts. Browning (2017) compared the performances of Gaussian Naive Bayes, Logistic Regression, Decision Trees, and Random Forest models to identify texts written by native English speakers. They used standard linguistic features, such as part-of-speech tags, and frequency of grammatical and spelling errors. Results showed that lexical and syntactic usage are strong indicators of language competency. For example, native speakers tend to produce more right-branching syntactic trees, avoid adverbs at the end of sentences, and use passive voice. However, discourse structures have not been analyzed.

Aoyama (2022) discovered that contextualized word embeddings (CWEs) from BERT (see Section 2.3.2) could be used to find differences in word usage between L1 and ELL writers of English. They used the EFCAMDAT (Geertzen, Alexopoulou, Korhonen, et al., 2013) dataset for language learners, and the LOCNESS (Granger, 1998) dataset for native English speakers. Aoyama (2022) found a steady decrease in CWE distance (that is to say, the distance between a word and the words providing its context) as proficiency level increases. Though this can provide an insight into grammatical structure, this work does not look specifically into discourse analysis and relations.

More recently, Schmalz and Brutti (2021) used BERT embeddings on 2-3 sentence inputs by English learners to train a model to automatically assess CEFR levels (see Section 2.1.1) based on frequencies of errors. They used a 100,000 essay sample from the EFCAMDAT and the CLC-FCE (Yannakoudakis, Briscoe, & Medlock, 2011) datasets, with essays written by adult English language learners. Schmalz and Brutti (2021) was able to show strong performance in CEFR-level classification, with both manual and automatic error detection. Montgomerie (2021) performed a similar task in which Logistic Regression, SVM, and Random Forest classifiers were used to classify reading practice tests from LearnEnglish[4] by the British Council into CEFR levels. Syntactic features such as parse tree depth, average frequency of part-of-speech tags, and text complexity metrics (see Section 2.1.2) were used as features for the classifier. With this dataset, the highest accuracy score reported for standard machine learning models was 68.2% for a Random Forest classifier. Again, discourse relation information was not considered.

2.2 Discourse Analysis

Discourse analysis is often defined as going "beyond the sentence" (Rimmer, 2006), and looking at the relationships between sentences and the structures that they form. This differs from syntactic analysis, which focuses on the grammatical structure of a sentence, and semantic analysis, which determines intra-sentence word meanings without looking at the larger context. In computational discourse analysis, computational techniques are used to analyze and understand discourse structures, meaning, and coherence in large amounts of text (Jurafsky & Martin, 2023). To facilitate the development of such computational tools, several discourse frameworks have been proposed. In this section we will focus on two frameworks: Rhetorical Structure Theory (RST), proposed by Mann and Thompson (1988) and Discourse Lexicalized Tree-Adjoining Grammar (Webber, 2004) (DLTAG), the basis for the Penn Discourse TreeBank (Prasad et al., 2008).

2.2.1 Rhetorical Structure Theory

In order to model the discourse structure of a text, Rhetorical Structure Theory (Mann & Thompson, 1988) first segments the text into distinct text spans, and then connects these spans to each other with a rhetorical relation. A rhetorical relation is defined as a pragmatic function that one span fulfills with respect to another (Jasinskaja & Karagjosova, 2020). The spans that are related to each other can be further described as a Nucleus or a Satellite. Relations will always contain a Nucleus, but the second span can be either a Nucleus (in a multi-nucleic relation) or a Satellite. A nucleus carries the meaning of a relation, while a satellite cannot be understood without its attachment to

[4] https://learnenglish.britishcouncil.org/skills/reading

the nucleus. Example 1 from the RST-DT (Carlson, Marcu, & Okurowski, 2002) corpus shows a sentence segmented using RST.

> **(1)** [Seasonal swings in the auto industry this year aren't occurring at the same time as in the past][1], [because of production and pricing differences.][2]

Segment 1 of Example 1 is a nucleus, as it can exist on its own as a complete thought. Segment 2 is a satellite related to the nucleus via an EXPLANATION relation which provides, as the name suggests, an explanation as to why the nucleus is true.

It is worth noting that these relations can be nested within each other. A span of text acting as a nucleus or a satellite of a relation may have embedded nuclei and satellites relating to each other within it, thus creating a hierarchical structure of relations. This tree is mathematically laid out by Kornai and Tuza (1992). At the lowest leaf level, also known as Elementary Discourse Units (EDUs), text spans in RST are non-overlapping, however, they are combined into higher levels in the tree structure. Thus, the text spans within a particular higher level of the tree structure do not overlap, but the segmentation forms the leaf EDUs.

RST schemas are often illustrated with the satellite pointing to the nucleus it relates to. The depth of trees is often displayed with these schemas stacked on top of each other. Example 2 is parsed as shown in Figure 2.1. EDUs 1-2 are a BACKGROUND satellite relating to the nucleus containing EDUs 3-4. Further down the tree, EDU 2 is an ELABORATION satellite to EDU 1, while EDU 3 and EDU 4 are related with a CONTRAST relation.

> **(2)** [Mario is a plumber;][1] [Bowser is his enemy.][2] [As the game progressed, Mario grew more powerful,][3] [while Bowser got weaker.][4]

The RST discourse Treebank (RST-DT) (Carlson et al., 2002) is the main standard corpus used in RST-based research. It was built following the RST framework, using 385 manually annotated Wall Street Journal articles (347 for training and 38 for testing) which cover a variety of topics and contain over 176,000 words. The RST-DT uses a total of 16 relation classes, as shown in Table 2.2. Appendix C shows a further explanation of 12 of these RST relations.

Several state-of-the-art RST parsers have been trained on the RST-DT corpus in order to segment and label discourse relations in new texts, with slight variances in their approach. Hernault,

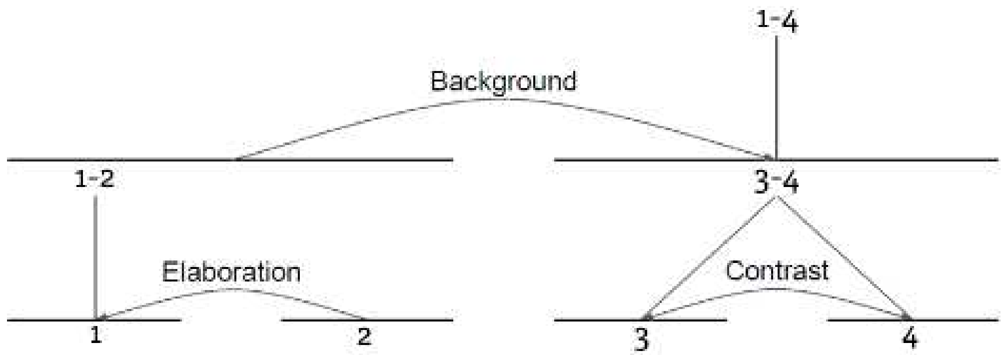

Figure 2.1: Example of an RST tree of Example 2, based on an visual format by Mann and Taboada (2005).

RST-DT Discourse Relations			
1	ATTRIBUTION	9	EVALUATION
2	BACKGROUND	10	EXPLANATION
3	CAUSE	11	JOINT
4	COMPARISON	12	MANNER-MEANS
5	CONDITION	13	SUMMARY
6	CONTRAST	14	TOPIC-COMMENT
7	ELABORATION	15	TEMPORAL
8	ENABLEMENT	16	TOPIC-CHANGE

Table 2.2: RST relations used in the RST-DT (Carlson & Marcu, 2001)

Prendinger, duVerle, and Ishizuka (2010) developed the HILDA classifier using two Support Vector Machine classifiers. The first is a binary classifier which decides whether two spans of text have a relation, while the second is a multi-class classifier for deciding the relation. Joty, Carenini, Ng, and Mehdad (2013) expanded on this classification by arguing for the importance of a classifier for both intra-sentential and inter-sentential relations, using a Conditional Random Field for both. Feng and Hirst (2014) used a greedy approach, where each step merges two existing spans, and then two Conditional Random Fields are used, one predicting the structure, and the other predicting the relation. Ji and Eisenstein (2014) use an approach in which predictions make incremental moves in feature-space to match the annotations of the training data in the RST-DT. Later, Heilman and Sagae (2015) used a transition-based approach, with multi-class logistic regression, and achieved similar results to the at-the-time state-of-the-art parsers while reporting an RST-DT test set document parsing time of 0.4 seconds per document, on an i7-4850HQ CPU at 2.30 GHz., compared to Feng and Hirst

(2014)'s reported 10.71 seconds on a system with "four duo-core 3.0 GHz processors." Li, Li, and Chang (2016) used a deep learning model attention-based hierarchical neural network to achieve similar results to previous parsers without the need for manual feature engineering. Y. Wang et al. (2017) used a two-stage approach, first using a transition-based model for identifying span and nuclearity, and a second stage where inter-sentence, cross-sentence, and cross-paragraph relations are classified with SVM classifiers. As will be discussed in Section 3.2.1, we used the Heilman and Sagae (2015) and Y. Wang et al. (2017) parsers given that they are the most recent open-source parsers. [5]

2.2.2 Discourse Lexicalized Tree-Adjoining Grammar

Apart from RST, Discourse Lexicalized Tree-Adjoining Grammar (DLTAG) constitutes the main framework for modelling textual discourse. DLTAG builds from its predecessor, Lexicalized Tree-Adjoining Grammar (LTAG) (XTAG Research Group, 1998), which in and of itself is a linguistic theory based on the formalism of tree-adjoining grammar. LTAG is a grammar in which initial or auxiliary trees are associated with each lexical item (single word, part of a word, or chain of words) in a language. Each of these trees has exactly one lexical anchor, meaning one lexical item carrying meaning.

Initial tree structures, often denoted by α, represent a simple tree built from small units of text, such as noun phrases or verb phrases, which contain only one non-terminal root with terminal leaves. On the other hand, auxiliary tree structures, often denoted by β, have a root and a foot which always match. Figure 2.2 shows a simple Auxiliary tree with a Verb Phrase (VP) root and a Verb Phrase (VP) foot. This allows auxiliary trees to be inserted into other trees via the process of substitution or adjunction. In substitution, a node is replaced by a tree with a root node of the same value. This operation requires an initial tree, or a derived tree with a non-terminal root, to take the place of the node being substituted. In adjunction, a new tree is created from an auxiliary tree and any other tree. This operation can only be applied to non-terminal nodes which have not previously been marked as substitution nodes. As an example, Figure 2.3 shows an auxiliary tree

[5] https://github.com/EducationalTestingService/rstfinder and https://github.com/yizhongw/StageDP

with a root and foot of a verb phrase (VP) attaching to another tree containing a verb phrase node. In Figure 2.3, (a) is the auxiliary tree to be attached to the VP node in the (b) tree. (c) is the derived tree produced by their adjunction. This appends an adverb in front of the previously existing verb phrase, retaining a grammatically correct sentence while altering the meaning.

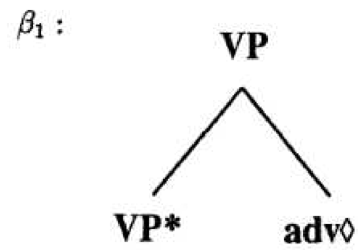

Figure 2.2: Example of an auxiliary tree showing a matching Verb Phrase foot and root, from Vijay-Shanker (1992).

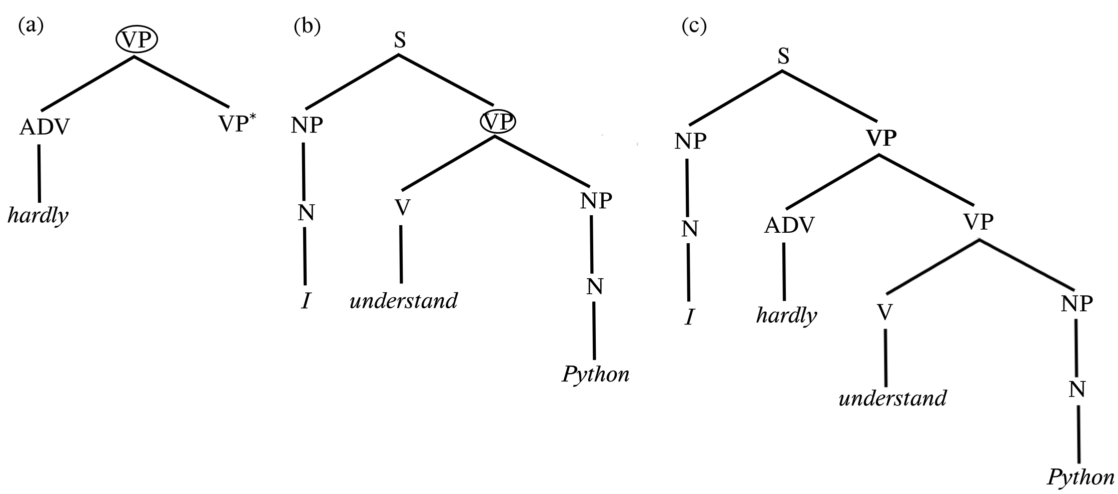

Figure 2.3: Example of the adjunction operation in LTAG.

While LTAG was developed for syntactic analysis, DLTAG is an extension of LTAG to model discourse. Much like in Rhetorical Structure Theory (see Section 2.2.1), the fundamental units of DLTAG consist of discourse segments separated by discourse connectives. These discourse connectives (e.g. *because, still, but, etc.*) may consist of subordinating conjunctions, coordinating conjunctions, adverbial phrases, prepositional phrases, or an implicit connective. DLTAG tree structures are based around these connectives, which can be singleton, in the case of a singular discourse connective, such as *so*, or parallel, in the case of a paired discourse connective, such as *on one hand* ... *on the other hand*. These paired discourse connectives are uncommon in English, unlike in other languages, such as Chinese Costa, Cheng, Muermans, Hanel, and Kosseim (2023). An example of an initial tree for a singular discourse connective is shown in Figure 2.4, while an example of an initial tree for a paired discourse connective is shown in Figure 2.5. In Figure 2.4, D_c indicates the discourse segment before and after the connective, while the down arrows indicate the points at which a substitution via an auxiliary tree may occur. In Figure 2.5, D_c indicates the discourse segments preceding each half of the paired discourse connective, while the down arrows indicate the points at which a substitution via an auxiliary tree may occur.

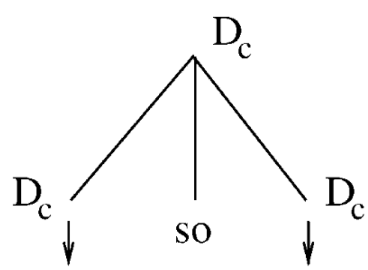

Figure 2.4: Initial tree structure for the singular discourse connective *so*, from Webber (2004).

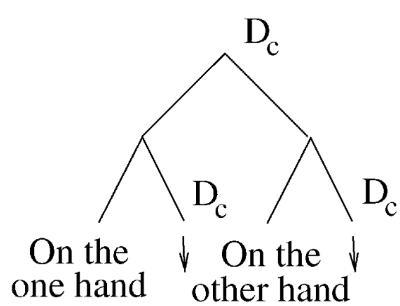

Figure 2.5: Initial tree structure for the paired discourse connective *on one hand ... on the other hand*, from Webber (2004).

DLTAG varies from RST based on its structure and classification. In the RST framework, the discourse structure of a text is represented via a tree (see Section 2.2.1), where two clauses with a relation to each other may be leaf nodes (EDUs) nested within the clause of another relation. In DLTAG, the discourse structure of a text is not hierarchical, and segments are separated by a discourse connective. If no discourse connective is explicitly stated, the two discourse segments are said to be related via an implicit discourse connective, with no lexical realization. Example 3 shows two arguments forming a COMPARISON connected by an explicit connective, *but*. This example uses the convention of Prasad, Forbes-Riley, and Lee (2017), in which Argument 1 is in italics, Argument 2 is in bold, and the discourse connective is underlined.

(3) *The Manhattan U.S. attorney's office stressed criminal cases from 1980 to 1987, averaging 43 for every 100,000 adults.* <u>But</u> **the New Jersey U.S. attorney averaged 16.**

Example 4 shows two arguments implicitly connected. This example, provided by Prasad et al. (2017), assumes that the implied connective is *For example*.

(4) *So far, the mega-issues are a hit with investors.* (Implicit=For example,) **Earlier this year, Tata Iron & Steel Co.'s offer of $355 million of convertible debentures was oversubscribed.**

The Penn Discourse TreeBank (Prasad, Webber, Lee, & Joshi, 2019) takes into account the lexical aspects of discourse provided by DLTAG when annotating and analyzing discourse relations. Three main versions have been developed, the PDTB-1.0 (Prasad et al., 2006), the PDTB-2.0 (Prasad et al., 2008), and the PDTB-3.0 (Prasad et al., 2019). As the PDTB-3.0 is very recent, most research has been done using PDTB-2.0. Figure 2.6 shows every discourse relation that appears in the PDTB-2.0. For this thesis, we will mainly be focusing on the four top-level relations, CONTINGENCY[6], EXPANSION, COMPARISON, and TEMPORAL. The PDTB-2.0 corpus contains over one million words of articles from the Wall Street Journal, with 53,631 manually annotated discourse relations.

[6]For sake of readability, RST relations are indicated in SMALL CAPS; while first-level PDTB relations are in CAPITAL letters.

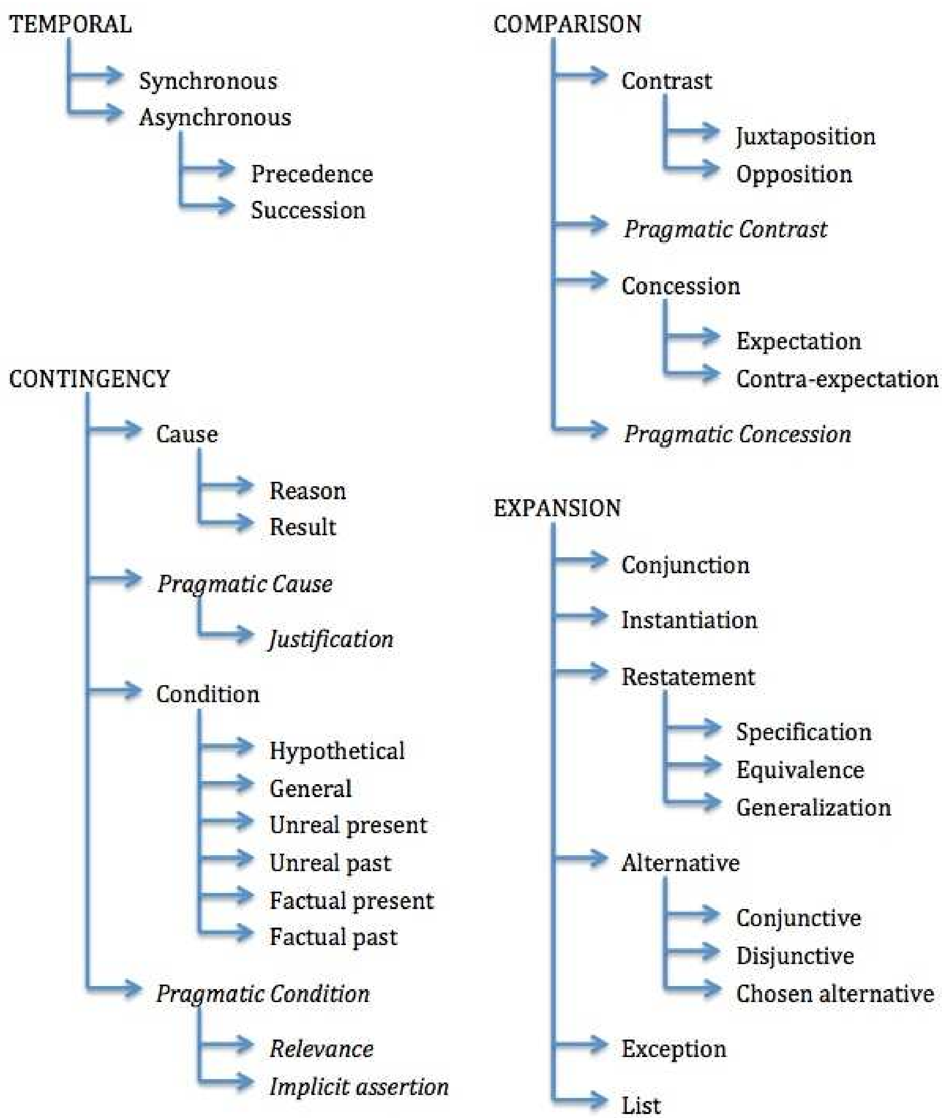

Figure 2.6: PDTB-2.0 relation hierarchy.
TEMPORAL, CONTINGENCY, COMPARISON, and EXCEPTION are the first-level relations which are further expanded into a tree of more specific relations.

RST-DT label \ PDTB label	Temp.		Cont.		Comp.		Expansion			Total
	Synch.	Asynch.	Cause	Condition	Contrast	Concession	Conjunction	Instantiation	List	
Temp.-same-time	<u>**63**</u>	1	1	1	4		5			75
Temporal-after		<u>**48**</u>			1		2			54
Sequence	2	<u>**29**</u>	1		1		19			52
Circumstance	<u>89</u>	79	29	21	7	4	10			259
Result	1	3	<u>**30**</u>		2		5			41
Consequence	5	6	<u>**45**</u>		4	1	16		1	95
Explanation-arg.	6		<u>**39**</u>		7	2	1			57
Reason		2	<u>**67**</u>							72
Condition	5	13	1	<u>**104**</u>	2	1	1			182
Contrast	1	1			<u>**160**</u>	23	17			208
Concession	5	4		3	101	<u>**53**</u>	4			182
Antithesis	1	3			<u>170</u>	<u>37</u>	10			243
Comparison	2	1			26	2	<u>**9**</u>			41
Elaboration-add.	4	3	1		30	8	<u>122</u>	3	3	192
Example	1				1		3	<u>**29**</u>		35
List	2	4	1		17	1	303		<u>**47**</u>	377
Total	187	197	215	129	533	132	527	32	51	

Figure 2.7: Explicit Discourse Relation mapping between RST and PDTB based on three linguistic approaches, from Demberg et al. (2017). Underlined and bold entries indicate that all proposals agree on the mapping, while underlined entries indicate an agreement of two approaches.

2.2.3 Mapping Discourse Relations Across Frameworks

Demberg et al. (2017) studied the compatibility of the two discourse annotation frameworks, RST and PDTB. The study found that while the two frameworks have some similarities, there are also differences in their annotations, that lead to compatibility issues. The authors provide insights into how to address these issues and improve the compatibility between the two frameworks for future research. They used a mapping of three different linguistic studies (Chiarcos (2014), Bunt and Prasad (2016), and Sanders et al. (2018)) and analyzed empirically the annotations of the overlapping texts between the RST-DT and the PDTB to propose a mapping between RST and PDTB-2.0 relations. Figure 2.7 shows this mapping, and will be the basis for conglomerating RST and PDTB data in Section 3. In the figure, underlined and bold entries indicate that all proposals agree on the mapping, while underlined entries indicate an agreement of two approaches.

2.2.4 Corpus Research in Discourse Analysis Across Textual Genres

Differences in discourse structures have been analyzed computationally across textual genres, text complexity, and cognitive abilities.

Webber (2009) and Bachand et al. (2014) showed that the genre of a text influences the choice of discourse relations. Bachand et al. (2014) used articles of various genres to look for common patterns of relations. The researchers observed, for example, that the RST relation of ATTRIBUTION is common in the newspaper article genre, JOINT is comparatively more frequent in online reviews, and TEMPORAL is more frequent in academic paper methodology sections.

Davoodi (2017) addressed a similar task of using both RST and PDTB relations to find to what degree these relations can be used as features to classify written texts, as well as exploring how the complexity level of a text influences its discourse-level linguistic choices. It was found, in the case of discourse relations, that there is no statistical difference in their explicit usage across levels of complexity, and that using discourse relations as features for classifying texts based on their complexity did not lead to better performance than the use of other linguistic features. However, the text complexity was shown to influence the usage of discourse connectives (e.g. *but, because*). Statistical differences were shown to exist; such as the more frequent usage of *because* as opposed to *thus* to signal CAUSE relations, and *but* as opposed to *while* to signal CONTRAST relations in simplified texts.

Abdalla et al. (2018) identified changes in the usage of discourse relations among patients with Alzheimer's disease. They used the RST parser of Feng and Hirst (2014) to analyze written material by patients with Alzheimer's, as well as a control group, using the DementiaBank (MacWhinney, Fromm, Forbes, & Holland, 2011) and CCC (Pope & Davis, 2011) datasets, which contain material from patients with Alzheimer's and a control group. Results showed that these two groups displayed a significant increase in ATTRIBUTION relations and a decrease in ELABORATION relations among writers with Alzheimer's disease.

While, to our knowledge, discourse structures have not been analyzed across English proficiency levels, some discourse relations, specifically PDTB level-1 relations, have been used to classify learner texts in Czech (Rysová, Rysová, & Mírovský, 2016), showing that the ratio of usage of these relations proved useful as features in Random Forest and Multilayer Perceptron classifiers.

2.3 Machine Learning Classification

As will be described in Section 4, we will be using a number of machine learning models in order to classify learner texts based on their CEFR level. Section 2.3.1 will describe a few of the traditional machine learning classification models used in our work. Section 2.3.2 will describe recent advancements in transformer-based models.

2.3.1 Classic Machine Learning Models

Many classification models have been used for text categorization. Below is a description of the main approaches. Each of these algorithms has its own advantages and disadvantages, and the choice of which to use depends on the specific requirements of the task at hand.

Support Vector Machines (SVMs) are a class of supervised learning algorithms that are widely used for classification and regression tasks (Cortes & Vapnik, 1995). The fundamental idea behind SVMs is to find a hyperplane, also known as a decision boundary, that separates the data into different classes with the maximum possible margin. The margin is defined as the distance between the closest points in the data, referred to as support vectors, and the hyperplane.

SVMs are particularly effective in dealing with high-dimensional data, where the number of features is much larger than the number of samples. They can handle non-linear decision boundaries by transforming the data into a higher dimensional space. SVMs are also relatively insensitive to the presence of noisy data. However, they can be computationally intensive, particularly for large datasets, and the choice of kernel can have a significant impact on the performance of the algorithm (Cortes & Vapnik, 1995).

One of the most commonly used kernels in SVMs is the Radial Basis Function (RBF) kernel. The RBF kernel maps the input data into a higher dimensional feature space through a non-linear transformation (Schölkopf & Smola, 2002). In this higher dimensional space, the decision boundary can be linear, even if the original data is not linearly separable. The RBF kernel is defined as the exponential of the negative Euclidean distance between two data points, multiplied by a parameter known as the bandwidth. The bandwidth parameter determines the smoothness of the decision boundary and can be optimized through cross-validation. The RBF function is shown below:

$$f(x) = \sum_{i=1}^{n} \alpha_i \exp(-\gamma ||x - x_i||^2). \tag{9}$$

In the equation, the variable $f(x)$ represents the output of the SVM classifier for a given input vector x. It serves as the decision function that predicts the class or assigns a score to the input. The coefficients α_i are associated with each training sample x_i and are determined during the SVM training phase. These coefficients indicate the importance or contribution of each training sample to the decision function.

The term $\exp(-\gamma ||x - x_i||^2)$ represents the radial basis function (RBF) kernel, which measures the similarity or distance between the input vector x and the support vectors x_i. The RBF kernel computes a weighted sum of exponential functions, where each term represents the similarity between the input vector and a support vector.

The parameter γ controls the width of the RBF kernel. It determines the influence or spread of the kernel. Higher values of γ result in a narrower kernel, meaning that only nearby support vectors have a significant impact on the decision function. Conversely, lower values of γ widen the kernel, allowing more support vectors to contribute to the decision function.

The term $||x - x_i||$ denotes the Euclidean distance between the input vector x and a support vector x_i. It quantifies the similarity or dissimilarity between the two vectors in the input feature space.

Lastly, the variable n represents the total number of support vectors used in the SVM model. These support vectors are a subset of the training samples that play a crucial role in defining the decision boundary.

The RBF kernel has advantages over the polynomial kernel in that it tests the hyperplane in an infinite number of dimensions, whereas the polynomial kernel is limited to a set number of dimensions.

Logistic Regression is a popular machine learning algorithm for solving classification problems, where the goal is to predict a categorical outcome based on one or more predictor variables. It is a statistical method that uses a logistic function to model the relationship between the predictors and the binary response variable. A main advantage of Logistic Regression is its speed.

The logistic function, also known as the sigmoid function, maps a real-valued number to a value between 0 and 1, which can be interpreted as the probability of the positive class. The parameters of the logistic regression model are estimated using maximum likelihood estimation, which maximizes the likelihood of observing the data given the model parameters (Nelder & Wedderburn, 1972).

One of the strengths of logistic regression is its interpretability. The coefficients of the model represent the change in the log-odds of the positive class for a one-unit increase in the predictor, while holding all other predictors constant. This allows for the identification of the most important predictors and the assessment of their individual and combined effect on the response (Nelder & Wedderburn, 1972).

Logistic regression is also relatively fast and computationally efficient, making it a good choice for large datasets. However, it is limited to linear relationships between the predictors and the response and may not be appropriate for datasets with complex non-linear relationships (Hastie, Trevor and Tibshirani, Robert and Friedman, Jerome, 2009).

Random Forest Classifiers solve both regression and classification problems. These classifiers operate by creating an ensemble of decision trees, each of which is trained on a random subset of the data and makes a prediction. The final prediction is made by aggregating the predictions of all the trees in the forest. This method not only improves the performance of the model but also reduces overfitting, which is a common issue with traditional decision trees (Breiman, 2001). Additionally, Random Forest classifiers have the ability to handle high dimensional data, missing values and are relatively easy to interpret compared to other complex models. Overall, Random Forest classifiers offer a robust and flexible solution for a variety of real-world problems and have proven to be effective in numerous applications across various domains (Liaw & Wiener, 2002).

2.3.2 Transformer-Based Methods

RoBERTa is a pre-trained transformer-based neural language model that uses bidirectional self-attention to perform a range of natural language processing tasks. In order to understand RoBERTa, it is necessary to understand its predecessor, BERT, as well as Transformers.

A transformer processes an ordered sequence of data, applies some neural-network based algorithm, and returns a sequence of outputs. The development of the Transformer architecture in 2017 (Vaswani et al., 2017) marked a significant milestone in the field of NLP. The Transformer allowed for parallel processing of sequences, making it suitable for processing long sequences of textual data for NLP tasks. This architecture has since become the foundation for many state-of-the-art models in NLP.

One such model is BERT (Bidirectional Encoder Representations from Transformers), introduced by Devlin, Chang, Lee, and Toutanova (2018). BERT is a pre-trained language model that has been fine-tuned for various NLP tasks such as sentiment analysis, named entity recognition, and question answering. BERT has achieved remarkable results on a wide range of NLP benchmarks, demonstrating the potency of the Transformer architecture in capturing contextual information in language data. Devlin et al. (2018) show that BERT achieved state-of-the-art performance on a range of natural language understanding tasks, including question answering, text classification, and named entity recognition. The authors also analyze the behaviour of BERT and show that it effectively captures contextual information in language data through its use of the Transformer architecture.

RoBERTa (Robustly Optimized BERT Approach), introduced by Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V (2019), builds upon the original BERT architecture and makes several modifications aimed at improving its performance. These modifications include a larger model size, longer pre-training time, and a larger corpus for training. These changes have resulted in RoBERTa outperforming BERT and other pre-trained models on various NLP benchmarks (e.g. A. Wang et al. (2018)), cementing the Transformer architecture's position as the current state-of-the-art in NLP. Given the significant performance of RoBERTa, we used it as a benchmark for our classifiers to aim towards (see Section 4.1).

2.4 Chapter Summary

In this chapter, we provided background on the assessment of textual proficiency levels and parsing methods, and frameworks used for discourse classification. We also reviewed work on corpus research, provided a linguistic perspective on the tools used in our work, and presented a deeper dive into the machine learning methods used in our work for classifying learner texts. In the next chapter, we will discuss the methodology and results of the discourse analysis of English learner texts.

Chapter 3

Discourse Analysis of English Learner Texts

This chapter provides an in-depth description of our methodology for the discourse analysis of essays, and the results. The findings of this chapter were accepted for publication at the forthcoming RANLP-2023 conference (Hanel & Kosseim, 2023).

Section 3.1 introduces the various datasets that were considered, including the International Corpus Network of Asian Learners and the Corpus and Repository of Writing.

In order to extract discourse relation and connective information from the dataset, we used two parsers from each discourse analysis framework. Section 3.2 will discuss the RST parsers, their agreement, and present the results of the discourse analysis for RST relations. Section 3.3 will discuss the PDTB parsers, their agreement, and present the results of the discourse analysis for level-1 PDTB relations and discourse connectives. Section 3.4 will describe the mapping between the two frameworks.

3.1 Datasets

In order to analyze discourse phenomena across learner texts, we needed corpora of essays pre-labelled with language proficiency levels. As Webber (2009) and Bachand et al. (2014) have shown, the genre of a text has an influence on the usage of discourse relations (see Section 2.2.4), thus it was important to find corpora of essays of the same genre. We used two datasets, the International Corpus Network of Asian Learners (see Section 3.1) and the Corpus & Repository of Writing (see Section 3.1).

The International Corpus Network of Asian Learners The first dataset we used was the International Corpus Network of Asian Learners (ICNALE) (Ishikawa, 2013). The ICNALE is an extension of the Corpus of English Essays Written by Asian University Students (CEEAUS), released in 2009. Compared to CEEAUS, ICNALE covers a greater diversity of writers from Asia, which makes it more reliable for international contrastive studies. In this dataset, writers presented their opinions about one of two statements: (a) *It is important for college students to have a part-time job.* and (b) *Smoking should be completely banned at all the restaurants in the country.* The ICNALE dataset used the ETS mapping (see Section 2.1.1) to convert TOEFL scores into CEFR scores. The dataset contains essays from 5 CEFR levels: A2, B1.1, B1.2, B2, and C2. In order to be compatible with the second dataset, we merged B1.1 and B1.2 instances to create a single B1 label.

	ICNALE Dataset					CROW Dataset				
	A2	**B1**	**B2**	**C2**	**All**	**A2**	**B1**	**B2**	**C2**	**All**
Essays	960	3976	464	400	5600	208	221	865	133	1429
Words per Essay	225	233	241	225	231	1207	846	905	2176	1057
Sentences per Essay	15	15	14	9	14	63	44	45	106	53

Table 3.1: Statistics of the ICNALE and CROW datasets.

Example 5 is a sample A2-level essay from an English learner in Korea, from the ICNALE dataset:[1]

(5) *I agree for this topic Because university life is based on social life I think gain knowledge of what is important but much experience more important I had a part time job help to me This work too hard, but useful thing and make some money I have work the part-time job for kindergarten assistance for children's festival. That part-time job is tired but I was happy and worth Part-time job are many useful species for university students. For example only vacation part-time job and special treatment for only university students You get part-time job of your major study and your interest's part-time*

[1] All grammar and punctuation errors in this text are verbatim from the original text.

job is help for your future. I will work part-time job for summer vacation Part time job is much advantage First, learn important manners are in society. Second, knowing the importance of money The reason money management for hard-earned money and understanding importance of money. Third hang out with friends. Moreover university students have much time especially vacation is best I think level up study for important but work part-time job important too. Part-time job to university students gain money and many useful experiences Therefore "i agree. It is important for college students to have a part-time job. "

The Corpus and Repository of Writing The Corpus and Repository of Writing (CROW) (Staples & Dilger, 2018) is a collection of written learner texts established in 2015 at Purdue University. The CROW dataset consists of essays split into three assignment groups: Argumentative Papers, Rhetorical Analysis, and Reflection. For the sake of consistency in genre, we only used the argumentative papers for comparison with the ICNALE dataset. These argumentative papers are comprised of essays from two different courses/institutions (ENGL 105 from Northern Arizona University and ENGL 106i from Purdue).

It is worth noting that CROW dataset texts are not labelled with a CEFR score, but rather with a TOEFL score (see Section 2.1.1). In 2010, the Educational Testing Service (ETS) proposed a metric for mapping TOEFL scores directly to CEFR levels[2]. The ICNALE dataset used this mapping (though the mapping has since been changed) to convert TOEFL scores to CEFR scores, so we used the same metric on the CROW dataset.

Example 6 is an excerpt from a B2-level essay from an English learner in China, from the CROW dataset:

(6) *There has been much debate recently about what should be done to improve the user-experience of smartphones. Some say that the manufactors should put a large battery into the phone in order to make the phone last longer while using. Others think that the designer should do harder work and make the smartphone thinner and lighter, in order to have a better hands-on experience. However, a larger battery means a*

[2] https://language.sakura.ne.jp/icnale/images/about/toefl_mapping.pdf

thicker and heavier phone. Therefore, the controversial topic kept existing among all the designers. A larger battery can provide a longer using time. Making the phone the width and the weigh of the phone can affect the user-experience by a lot. Too much improvement on either side can make the phone extremely expensive which can be the biggest killer of user-experience.

Table 3.1 shows statistics of both datasets. A2-B2 essays are from English learners, while C2 essays are from countries with English as an official language. As the table shows, ICNALE is significantly larger than CROW (5600 essays compared to 1429). However, the essays in CROW are longer with a word-per-essay average of 1057 words vs 231. In addition, as shown in Table 3.1, the datasets do not contain samples of A1 and C1 CEFR levels, and are not balanced across levels. For the discourse analysis, all essays were considered in order to maximize the size of the dataset. However, for the automatic classification (see Chapter 4), ICNALE's B1 was under-sampled to 1000 random essays in order to balance the dataset and reduce bias.

Other Datasets Considered Other datasets were considered, but were not appropriate for our work. Among these, we considered the International Corpus of Learner English (ICLE) (Granger, Dupont, Meunier, Naets, & Paquot, 2020), the EFCAMDAT (Geertzen et al., 2013) and the CLC-FCE (Yannakoudakis et al., 2011).

ICLE (Granger et al., 2020) is a research database of written and spoken English produced by non-native speakers. It was created in 1994 by Sylviane Granger, a Belgian linguist and researcher in the field of second language acquisition. The ICLE dataset was designed to provide a comprehensive and representative sample of the English language produced by learners from a wide range of first language backgrounds, including Chinese, German, French, Spanish, and Russian. The corpus is based on data collected from learner essays, exam scripts, and spoken interactions. This dataset was not used in our work because of its small amount of labelled data, as only 20 essays from each first-language corpus were analyzed by a manual annotator to determine the writer's CEFR score.

We had also considered the use of the EFCAMDAT (Geertzen et al., 2013) and CLC-FCE (Yannakoudakis et al., 2011) datasets discussed in Schmalz and Brutti (2021); however these datasets consist of short question-answer pairs rather than essays, and hence current discourse parsers (see

Section 2.2.1) are not capable of parsing these to extract relevant discourse information.

3.2 RST-DT Parsing

In order to identify patterns across CEFR levels, we analyzed the data to find the average frequency for essays of each CEFR label in the dataset: A2, B1, B2, and native speakers. To determine if the differences were statistically significant, we ran a two-tailed t-test with a significance level of 0.05 for each of these spans, comparing A2 against native speakers, B1 against native speakers, and B2 against native speakers.

3.2.1 RST-DT Parsing Methodology

Today, many pre-trained discourse parsers are publicly available. For RST, the two most recent publicly-available parsers are the Y. Wang et al. (2017) parser and the Heilman and Sagae (2015) parser. These parsers are directly compared in Y. Wang et al. (2017), and both parsers are publicly available to download and train on GitHub.[3] In addition, the Heilman and Sagae (2015) parser is highly efficient, allowing for quick discourse parsing without reliance on an external GPU. However, the parsers were not ready to use and had to be trained before use.

Setting up the Heilman and Sagae (2015) parser As recommended by the authors, we trained the Heilman and Sagae (2015) parser on the Penn Treebank[4] (Marcus, Santorini, & Marcinkiewicz, 1993) for syntactic information, and on the RST Discourse Treebank (RST-DT) (Carlson et al., 2002) for discourse information. The RST-DT consists of 385 Wall Street Journal articles annotated with discourse structure in the RST framework. To train the Heilman and Sagae (2015) parser, we first split the RST-DT using the standard split of 38 documents for the test set, 40 for the validation set, and the other 307 for training. The next step was to train a discourse segmentation model from the newly created training set and the development set.

[3] `https://github.com/EducationalTestingService/rstfinder` and `https://github.com/yizhongw/StageDP`

[4] The Penn Treebank (PTB), not to be confused with the Penn Discourse TreeBank (PDTB), is a corpus annotated with *syntactic* information, such as syntactic tree structures for sentences, part-of-speech tags, and phrase structure.

We trained the Conditional Random Field model, tasked with segmenting the text into Elementary Discourse Units (EDUs). During the training process, the regularization parameter, C, of the CRF model was optimized to maximize the likelihood of the observed data given the model. The goal of optimizing this hyperparameter was to find the optimal balance between fitting the data well and avoiding overfitting. This was achieved by training the CRF model multiple times with different values for the regularization hyperparameter (ranging from 0.0625 to 16.0), using a simple 1-dimensional grid search, and evaluating its performance on the validation set. As shown in Table 3.2, a C value of 16.0 was shown to produce the highest F1 score on the RST-DT validation set.

C	Precision	Recall	F1
0.015625	92.82%	81.17%	86.60%
0.03125	93.16%	84.66%	88.71%
0.0625	93.87%	86.21%	89.88%
0.125	93.87%	87.18%	90.40%
0.25	93.48%	88.22%	90.78%
0.5	93.31%	88.48%	90.83%
1	93.37%	88.48%	90.86%
2	93.57%	88.61%	91.02%
4	93.58%	88.74%	91.10%
8	93.46%	88.87%	91.11%
16	**93.47%**	**88.93%**	**91.14%**
32	93.34%	88.93%	91.08%
64	93.34%	88.87%	91.05%

Table 3.2: Fine-tuning of the C regularization value for the Conditional Random Field of the Heilman and Sagae (2015) parser for segmentation.

After using the CRF for segmentation, the Heilman and Sagae (2015) parser uses a linear regression model to select the most appropriate RST relation. We trained this model with a similar hyperparameter tasked with normalization. The linear regression model is aided with pre-generated features from the SciKit-Learn Laboratory (SKLL[5]). As shown in Table 3.3, a regularization parameter value of 0.5 was shown to produce the highest F1 score on the RST-DT validation set

After running on the ICNALE and CROW datasets, the Heilman and Sagae (2015) parser produced a JSON file with separate objects for each EDU, as well as a dependency tree. `D3.js` was

[5]https://skll.readthedocs.io/en/latest/

C	Precision	Recall	F1
0.0625	55.86%	56.57%	56.21%
0.125	58.27%	58.98%	58.61%
0.25	58.65%	59.37%	59.01%
0.5	**59.18%**	**59.63%**	**59.41%**
1	57.63%	58.10%	57.87%
2	56.05%	56.46%	56.26%
4	55.37%	55.72%	55.55%
8	55.32%	55.62%	55.47%
16	54.61%	54.89%	54.75%

Table 3.3: Fine-tuning of the C regularization value for the Conditional Random Field of the Heilman and Sagae (2015) parser for relation labelling.

used to visualize this tree in HTML. An example of this dependency tree is shown in Figure 3.1. Red circles symbolize nuclei, while white circles symbolize satellites (except for the furthest left circle, which is the root).

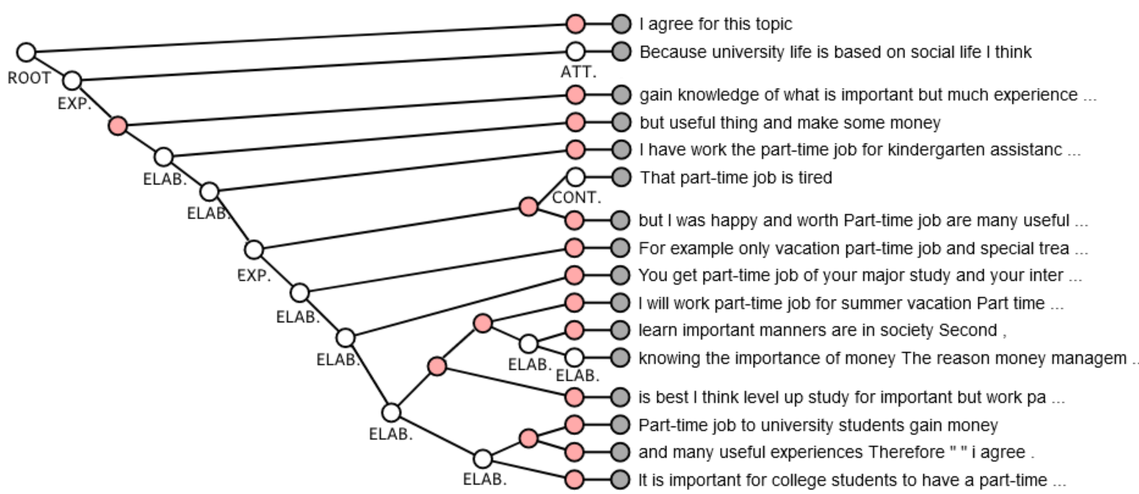

Figure 3.1: RST dependency tree visualized by D3.js using an A2-level essay from Ishikawa (2013), with satellites labelled with their RST relation.

Setting up the Y. Wang et al. (2017) parser The Y. Wang et al. (2017) parser was similarly trained on the RST-DT. For obtaining syntactic information, rather than training on the Penn Tree-Bank like the Heilman and Sagae (2015) parser, this parser used a CoreNLP server[6]. This server performed tokenization, part-of-speech tagging, sentence splitting, word lemmatization, and named

[6]https://stanfordnlp.github.io/CoreNLP/

entity recognition. Next, we extracted feature templates, action maps, and relation maps to convert the raw text data into a numerical representation that could be used as input to a machine learning model. Feature templates are predefined patterns that capture relevant linguistic aspects of a unit and represent them as feature vectors. Action maps are used for sequence labelling to map the current state and desired label to actions, and relation maps are used for dependency parsing to map the current word, context, and desired relation to actions. Example 7 shows a snippet of the output in the `.parse` file for the same A2-level essay from Korea used previously.

As this parser did not natively come with an EDU segmenter, we used an external tool[7].

Little preprocessing other than feature engineering had to be done on the raw text files, except for CROW documents, which were provided in a format which contained essay metadata within the `.txt` files, so the metadata needed to be removed prior to parsing.

(7)

```
(SN-Attribution
  (NS-Explanation (EDU _!I_agree_for_this_topic!_) (EDU _!
  Because_university_life_is_based_on_social_life_I_think!_))
  (NN-Contrast
    (EDU _!gain_knowledge_of_what_is_important_but_much
    _experience_more_important_I_had_a_part_time_job_help
    _to_me_This_work_too_hard_,!_)
    (NS-Elaboration
      (EDU _!but_useful_thing_and_make_some_money!_)
      (NS-Attribution
        (EDU _!I_have_work_the_part_-_time_job_for
        _kindergarten_assistance_for_children_'s_festival_)
        (SN-Contrast ...
```

[7] http://www.chokkan.org/software/crfsuite/

	Heilman and Sagae (2015)	Y. Wang et al. (2017)	Average
1. Enablement	58.00%	75.76%	66.88%
2. Attribution	88.84%	96.68%	92.76%
3. Elaboration	100.00%	99.93%	99.97%
4. Temporal	14.21%	35.37%	24.79%
5. Joint	84.68%	96.11%	90.40%
6. Contrast	64.93%	75.28%	70.11%
7. Explanation	77.96%	55.49%	66.73%
8. Cause	39.16%	25.14%	32.15%
9. Condition	63.57%	63.95%	63.76%
10. Background	56.55%	65.63%	61.09%
11. Manner-Means	20.57%	23.05%	21.81%
12. Comparison	9.73%	8.90%	9.32%
13. Summary	0.82%	0.05%	0.44%
14. Evaluation	6.73%	0.63%	3.68%
15. Topic-Comment	<0.01%	<0.01%	<0.01%
16. Topic-Change	<0.01%	<0.01%	<0.01%

Table 3.4: Percentage of essays in the ICNALE dataset containing at least 1 of the given RST relation.

Running the RST parsers After parsing the ICNALE and CROW datasets with both RST parsers, we used the output of each RST parser to find the frequency of each RST relation label per essay with respect to the total number of relation labels. This gave us the percentage of discourse relations for each essay.

Each RST parser used the same set of 16 labels. Statistics of the frequency of each relation were then collected. Table 3.4 shows the percentage of essays containing at least 1 of the given RST relation. As Table 3.4 shows, Evaluation, Topic-Comment, Topic-Change, and Summary were found on average in less than 5% of the documents. They were therefore considered too infrequent to evaluate further. In total, 12 RST relations were therefore considered for further analysis.

Both RST parsers were trained and parsed files on a desktop computer running Ubuntu 20.04.4 LTS with an Intel® Core™ i7-4770 CPU @ 3.40GHz with 4 cores. The total training time of each parser was below 5 minutes. The total parsing time on the ICNALE dataset of 5600 documents ranged from 8 to 12 hours on the Heilman and Sagae (2015)-based parser and 6 days on the Y. Wang et al. (2017)-based parser.

3.2.2 RST Parser Agreement

Given that each RST parser can make segmentation and labelling errors, we computed their agreement across the two datasets. Much research has addressed the alignment of RST and PDTB annotations (Demberg et al., 2017), but even between two RST parsers with the same labels, computing their agreement on the same dataset can be a difficult task, as the tree structures may not match. To align the annotations, we used the following method. Given 2 EDUs from each parser, $\mathrm{EDU}_{\mathrm{p1}}$ and $\mathrm{EDU}_{\mathrm{p2}}$:

Segment Alignment:

If $\mathrm{EDU}_{\mathrm{p1}}$ and $\mathrm{EDU}_{\mathrm{p2}}$ span the same text (sans punctuation), we align them and keep the pair ($\mathrm{EDU}_{\mathrm{p1}}$, $\mathrm{EDU}_{\mathrm{p2}}$) along with their associated discourse annotations for relation agreement. This case alone led to an inter-parser agreement of over 95%.

Relation Alignment:

(1) For each $\mathrm{EDU}_{\mathrm{p}i}$ in the aligned ($\mathrm{EDU}_{\mathrm{p1}}$, $\mathrm{EDU}_{\mathrm{p2}}$),

- If $\mathrm{EDU}_{\mathrm{p}i}$ was labelled as a satellite by parser pi, it is then labelled with its lowest-level discourse relation. In Figure 3.2, satellite A would be labelled ATTRIBUTION, while satellite C would be labelled EXPLANATION. Satellite B, as a nucleus, would not receive a label.
- If $\mathrm{EDU}_{\mathrm{p}i}$ was the second EDU of a multi-nucleic relation, it is labelled with the discourse relation that represents the pair (labelling only the second EDU in the pair prevents the relation from being double-counted).
- Otherwise, $\mathrm{EDU}_{\mathrm{p}i}$ is not assigned a relation.

(2) For each EDU:

- If BOTH parsers label the EDU as a satellite, and they have the same relation, mark them as an agreement.
- If BOTH parsers label the EDU as a satellite, and they have a different relation, mark them as a disagreement.

	Heilman and Sagae (2015)												
Y. Wang et al. (2017) parser	**Ena.**	**Att.**	**Ela.**	**Tem.**	**Joi.**	**Cont.**	**Exp.**	**M-M**	**Cau.**	**Cond.**	**Bac.**	**Com.**	**Total**
Enablement	**1236**	56	597	3	104	13	11	3	28	13	17	3	2084
Attribution	69	**9488**	488	8	281	81	43	10	66	132	108	10	10784
Elaboration	697	378	**10415**	60	628	114	88	69	124	105	336	34	13048
Temporal	2	44	50	**299**	26	43	8	1	7	6	131	2	619
Joint	15	46	187	10	**1732**	10	2	6	35	10	18	2	2073
Contrast	36	64	173	36	111	**951**	39	7	24	118	53	5	1617
Explanation	2	39	21	1	29	3	**503**	0	187	8	4	2	799
Manner-Means	1	14	9	0	7	0	1	**386**	0	2	38	2	460
Cause	9	9	15	2	13	4	99	2	**96**	5	13	3	270
Condition	21	125	106	13	44	11	17	12	13	**2594**	55	1	3012
Background	15	123	161	50	62	10	9	27	55	51	**1732**	68	2363
Comparison	1	7	3	0	6	0	0	0	1	2	19	**124**	163
Total	2104	10393	12225	482	3043	1240	820	523	636	3046	2524	256	37292

Table 3.5: RST Parser agreement between the Heilman and Sagae (2015) parser along the x-axis and the Y. Wang et al. (2017) parser on the y-axis, on the ICNALE dataset.

- Otherwise, if one or both parsers label the EDU as a nucleus, the EDU is ignored, since its relation has already been considered through its satellite.

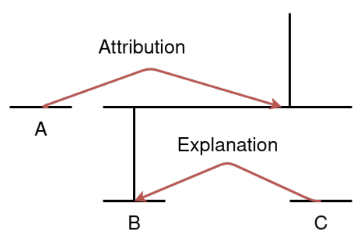

Figure 3.2: Example RST tree.

Using this method, we were able to verify the agreement between the two parsers on the 12 most frequent satellite-nucleus RST relations. The two parsers showed an agreement of 79.3% on relation tags on the ICNALE dataset, and an agreement of 80.3% on the CROW dataset, with the full results shown in Table 3.6 and Table 3.7. As the results show, the parsers disagreed most frequently on CAUSE relations, frequently mislabelling these relations as EXPLANATION. For the following analysis, only the EDUs with an agreed-upon relation between the two parsers were used. Therefore, only 29556 relation instances were kept out of 37292 from ICNALE and 31689 out of 39469 from CROW.

		Heilman and Sagae (2015)											
		Ena.	Att.	Ela.	Tem.	Joi.	Cont.	Exp.	M-M	Cau.	Cond.	Bac.	Com.
Y. Wang et al. (2017) parser	Enablement	**58.75%**	0.54%	4.88%	0.62%	3.42%	1.05%	1.34%	0.57%	4.40%	0.43%	0.67%	1.17%
	Attribution	3.28%	**91.29%**	3.99%	1.66%	9.23%	6.53%	5.24%	1.91%	10.38%	4.33%	4.28%	3.91%
	Elaboration	33.13%	3.64%	**85.19%**	12.45%	20.64%	9.19%	10.73%	13.19%	19.50%	3.45%	13.31%	13.28%
	Temporal	0.10%	0.42%	0.41%	**62.03%**	0.85%	3.47%	0.98%	0.19%	1.10%	0.20%	5.19%	0.78%
	Joint	0.71%	0.44%	1.53%	2.07%	**56.92%**	0.81%	0.24%	1.15%	5.50%	0.33%	0.71%	0.78%
	Contrast	1.71%	0.62%	1.42%	7.47%	3.65%	**76.69%**	4.76%	1.34%	3.77%	3.87%	2.10%	1.95%
	Explanation	0.10%	0.38%	0.17%	0.21%	0.95%	0.24%	**61.34%**	0.00%	29.40%	0.26%	0.16%	0.78%
	Manner-Means	0.05%	0.13%	0.07%	0.00%	0.23%	0.00%	0.12%	**73.80%**	0.00%	0.07%	1.51%	0.78%
	Cause	0.43%	0.09%	0.12%	0.41%	0.43%	0.32%	12.07%	0.38%	**15.09%**	0.16%	0.52%	1.17%
	Condition	1.00%	1.20%	0.87%	2.70%	1.45%	0.89%	2.07%	2.29%	2.04%	**85.16%**	2.18%	0.39%
	Background	0.71%	1.18%	1.32%	10.37%	2.04%	0.81%	1.10%	5.16%	8.65%	1.67%	**68.62%**	26.56%
	Comparison	0.05%	0.07%	0.02%	0.00%	0.20%	0.00%	0.00%	0.00%	0.16%	0.07%	0.75%	**48.44%**

Table 3.6: Percentages of RST Parser agreement between the Heilman and Sagae (2015) parser along the x-axis and the Y. Wang et al. (2017) parser on the y-axis, on the ICNALE dataset.

		Heilman and Sagae (2015)												
		Ena.	Att.	Ela.	Tem.	Joi.	Cont.	Exp.	M-M	Cau.	Cond.	Bac.	Com.	Total
Y. Wang et al. (2017) parser	Enablement	**1556**	35	746	4	151	15	11	4	40	16	20	3	2602
	Attribution	72	**9093**	553	9	137	110	57	11	77	130	144	12	10406
	Elaboration	738	203	**12845**	121	306	141	87	98	130	132	357	33	15193
	Temporal	3	9	62	**397**	26	54	10	1	7	6	172	3	750
	Joint	4	14	248	1	**714**	5	2	8	13	8	10	1	1030
	Contrast	48	59	167	38	117	**973**	49	9	33	144	67	6	1696
	Explanation	3	10	28	1	33	3	**578**	0	27	12	5	3	704
	Manner-Means	1	5	12	0	7	0	1	**553**	0	2	54	3	638
	Cause	10	8	242	3	12	3	125	3	**103**	6	14	3	533
	Condition	27	2	149	18	48	10	22	15	13	**2738**	67	1	3111
	Background	18	41	173	56	60	14	11	30	62	61	**1979**	80	2587
	Comparison	0	1	4	0	5	0	0	0	1	3	20	**174**	207
	Total	2480	9481	15230	649	1617	1315	955	733	507	3257	2909	322	39469

Table 3.7: RST Parser agreement between the Heilman and Sagae (2015) parser along the x-axis and the Y. Wang et al. (2017) parser on the y-axis, on the CROW dataset.

3.2.3 RST Relations Across CEFR Levels

While many RST relations showed some statistical differences between learner and native speaker essays, only two of the twelve showed the same patterns across the two datasets, in which at least two of the three t-tests showed a significant difference. For the relation of EXPLANATION, both parsers and both datasets showed a statistical difference in A2 vs C2 and B1 vs C2, but no statistical difference between B2 and C2. The data suggests a general downward trend in the usage of EXPLANATION relations, which flattens out as the learner reaches the B2 level. In Figure 3.3, "*" indicates a p-value less than 0.05 when comparing the marked data point against C2-level essays with a t-test. Intuitively, individuals with lower CEFR levels may have a more limited vocabulary and understanding of complex sentence structures, which can make it more difficult for them to express themselves in a clear and concise way. As a result, they may rely more heavily on the RST relation of EXPLANATION to clarify their meaning and provide additional detail to support their arguments or ideas.

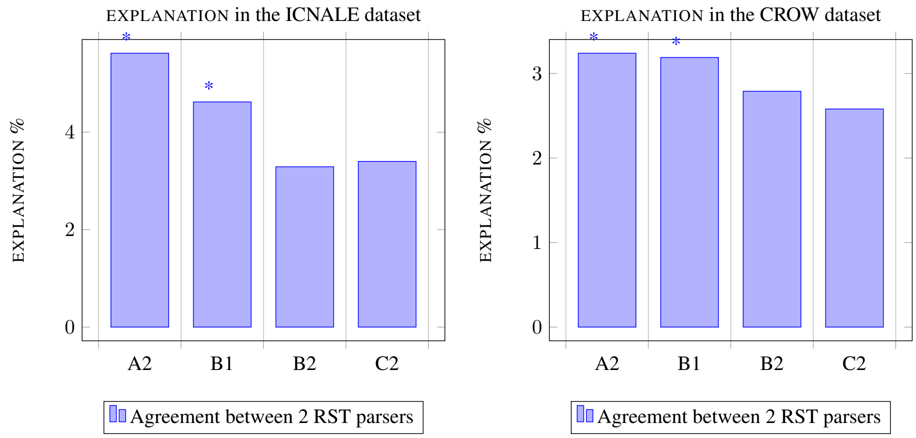

Figure 3.3: Percentage of RST relation of EXPLANATION in the CROW and ICNALE datasets, as parsed by both RST parsers.

For the RST relation of BACKGROUND, both parsers and both datasets show a statistical difference in B1 vs C2 and B2 vs C2, but no statistical difference between A2 and C2. This is shown in Figure 3.4, which suggests that newer learners use BACKGROUND relations at a similar rate to native English speakers (C2), whereas B-level learners show an increase in these relations. In this figure, "*" indicates a p-value less than 0.05 when comparing the marked data point against C2-level essays. The RST relation of BACKGROUND is used to provide information that is important to understanding the main idea or topic of a text. English language learners may rely more heavily on BACKGROUND to provide necessary context and establish the main topic or theme of their writing. However, A2 level English learners may not have the language skills necessary to effectively attribute a background to the points they are attempting to convey.

Example 8 shows an example of a BACKGROUND relation, from a B2-level essay, in the ICNALE dataset. The satellite's opening clause, "As we all know," is a strong indicator of this type of relation, as it often comes before factual (or presented by the writer as factual) background information relevant to understand the argument.

(8) [In this way, it is a quite reasonable rule that all the customers should obey. To ban smoking in restaurants can also help educate the teenagers.][1] [As we all know, the

		Elab.	Exp.	M-M	Att.	Joi.	Ena.	Back.	Comp.	Cont.	Cau.	Tem.	Cond.
ICNALE	**A2**	43.52	5.62	0.87	13.91	13.91	4.29	3.90	0.24	5.84	1.49	1.21	6.59
	B1	46.62	4.62	0.90	12.37	13.79	4.18	4.50	0.34	5.71	1.56	1.37	5.47
	B2	48.03	3.29	1.10	11.72	12.81	4.25	5.03	0.39	6.01	1.72	1.47	5.27
	C2	41.77	3.40	0.92	17.02	16.91	3.71	3.96	0.41	4.93	1.13	1.43	5.68
CROW	**A2**	65.75	3.24	3.16	6.14	8.32	1.89	2.34	0.38	3.16	1.05	0.46	1.23
	B1	63.47	3.19	2.93	6.56	9.72	2.01	2.78	0.30	2.93	1.02	0.53	1.47
	B2	64.62	2.79	2.77	6.06	9.00	2.01	2.75	0.38	2.77	1.02	0.50	1.12
	C2	63.97	2.58	2.58	5.70	11.22	1.62	2.21	0.21	2.58	0.86	0.43	1.26

Table 3.8: Frequencies of each RST relation by dataset and CEFR score.

number of teenagers who smoke is rising rapidly.][2]

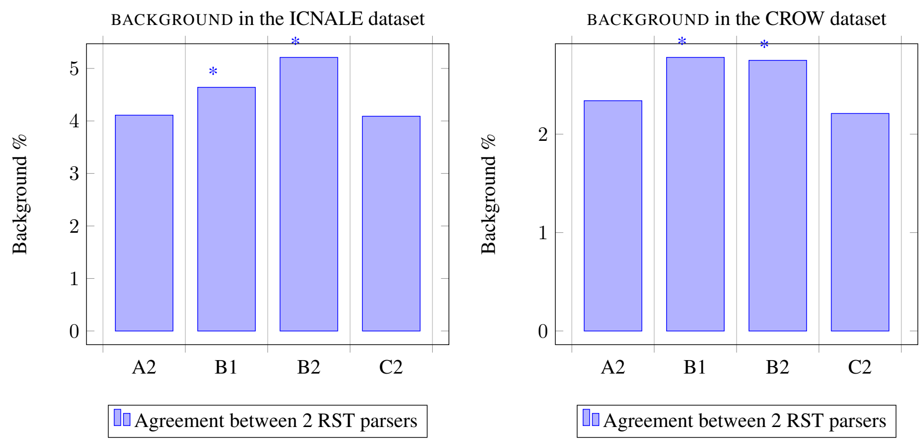

Figure 3.4: BACKGROUND mean in the CROW and ICNALE datasets, as parsed by 2 RST parsers.

Table 3.8 shows the frequencies of each of the twelve discourse relations. Highlighted scores have a p-value below 0.05 when compared to C2 in a two-tailed t-test. EXPLANATION and BACKGROUND have already been shown to have matching patterns of t-test results, but there are other notable relations as well. The RST relation of JOINT is used much more often among C2 writers than by English learners, but the t-test between CROW's B1-level and C2 is just above the 0.05 p threshold. While C2 speakers seem to use CAUSE relations less frequently than learners, the values don't vary enough to be statistically significant. CONTRAST relations are also used less frequently in C2-level essays, however the results are inconclusive for the difference among A- and B-level essays, as the frequency is highest in the B2 range in the ICNALE dataset, while it is highest in the A2 range in the CROW dataset.

3.3 PDTB Parsing

Two PDTB parsers were also used to compare results with the RST, including the popular PDTB parser presented by Lin et al. (2014) and the high scoring parser of J. Wang and Lan (2015) used at the CONLL-shared 2015 Shared Task (Xue et al., 2015).

Setting up the Lin et al. (2014) parser A Java version of the Lin et al. (2014) parser is available on GitHub[8]. It is pre-trained and ready to be used on raw text files out of the box. The parser, called the PDTB-Styled End-to-End Discourse Parser, uses a deep neural network architecture that takes raw text as input and outputs a discourse structure in the form of relations between discourse units. The network is trained on the PDTB training set (Prasad et al., 2008) to learn the patterns and features that are indicative of different types of discourse relations. During training, the parser learns to identify discourse connectives, such as *however* and *because*, and to use them to predict the type of relation that holds between two discourse units. The parser has an additional dependency on The Stanford CoreNLP Dataset (Manning et al., 2014) for reading and generating parse trees.

Once trained, the PDTB-Styled End-to-End Discourse Parser can be applied to new texts to produce a discourse analysis. The parser first segments the text into discourse units, then uses the learned patterns and features to predict the discourse relations between these units. The output of the parser is a set of discourse relations, each of which is characterized by the type of relation, the connective used (if any), and the arguments (i.e., the discourse segments) involved in the relation. Example 9 shows one line of a `.txt` file for the A2-level essay from ICNALE shown in Example 5. An explanation of each string in the output is laid out in Table 3.9.[9] Example 10 shows the full quote of the text being described.

[8]https://github.com/WING-NUS/pdtb-parser

[9]A full list of possible outputs is available at `https://github.com/WING-NUS/pdtb-parser/blob/master/README.md`, yet most are not relevant to our work.

Key	String
Relation type (Explicit/Implicit)	Explicit
Character location of connective	613...616
Connective (raw text)	and
Sentence number	1
Connective (in lowercase)	and
Relation	Expansion
Character location of Arg1	480..612
Arg1 text	For example only vacation part-time job and special treatment for only university students You get part-time job of your major study
Character location of Arg2	617..670
Arg2 text	your interest's part-time job is help for your future

Table 3.9: Explanation of the output of the Lin et al. (2014) parser.

(9)

```
Explicit|||613..616||and||1|and|||Expansion||||||||||
480..612||For example only vacation part-time job and
special treatment for only university students You get
part-time job of your major study||||||||617..670||your
interest's part-time job is help for your future||||||
||||||
```

(10) *For example only vacation part-time job and special treatment for only university students You get part-time job of your major study* <u>and</u> **your interest's part-time job is help for your future.**

Note that, in this example, the first argument is particularly long. This may be due a reliance on punctuation in PDTB segmentation, which poses a challenge on A-level essays such as this one in which the writer may not have a grasp on punctuation and writes long, run-on sentences.

Setting up the J. Wang and Lan (2015) parser For a point of comparison, the code[10] from J. Wang and Lan (2015) was also used to parse the data into PDTB format. The parser uses a neural

[10] https://github.com/lanmanok/conll2015_discourse

network architecture to generate discourse relation and connective predictions based on the representation of the input text. This representation is generated using a combination of word embeddings, contextual embeddings, and linguistic features. The main difference between the approaches of the Lin et al. (2014) and the J. Wang and Lan (2015) parsers lies in the way they generate the representations used to predict discourse relations. J. Wang and Lan (2015)'s approach generates the representations directly from the text, while Lin et al. (2014)'s approach generates the representations based on the PDTB-style discourse tree structure and then uses these representations to make the final relation predictions.

As the parser was originally run for the CONLL 2015 shared task (Xue et al., 2015), it required input of a JSON file containing the syntactic parse tree. To generate such files, we began by executing the Stanford parser's "lexparser" script on the input file. Then, we implemented a conversion function to transform the resulting output into JSON format. During the conversion process, we parsed the data line by line. We identified and stored the parse tree lines along with the corresponding dependencies for each sentence. We ensured that the parse tree lines were concatenated correctly. Finally, we constructed a JSON structure with the sentences, parse trees, and dependencies, and saved it to a JSON file with the same name as the input file. This resulted in a JSON representation of the input data that captured the sentence structure and linguistic dependencies. In a similar process with the Berkeley parser, we appended the output to the JSON, which provided tokenized words, including the character start and end points in the raw text file. Figure 3.5 is a snippet of example output from a sentence in an A2-level essay in the ICNALE dataset, with "dependencies" and "words" truncated to just one example.

Running the PDTB parsers Much like the RST parsers' outputs, we aligned the outputs of the two parsers, then extracted frequencies of PDTB level-1 discourse relations for each dataset. We used the output of each PDTB parser to find the frequency of every PDTB level 1 relation (see Figure 2.6) per essay with respect to the total number of relation labels.

Both PDTB parsers were run on the same desktop computer running Ubuntu 20.04.4 LTS with an Intel® Core™ i7-4770 CPU @ 3.40GHz with 4 cores. Total parsing time on the ICNALE dataset of 5600 documents ranged from 4 to 8 hours on both parsers.

```
{
    "parsetree": "(ROOT(S(ADVP (RB Nowadays))(, ,)(NP(NP (DT a)(JJ
    large) (NN number))(PP (IN of)(NP (NN college) (NNS students))))(VP
    (VBP are)(VP (VBG having)(NP (PRP it)))(. .)))",
    "dependencies": [
      [
        "advmod",
        "having-10",
        "Nowadays-1"
      ]
    ],
    "words": [
      [
        "Nowadays",
        {
          "CharacterOffsetBegin": 9,
          "CharacterOffsetEnd": 17,
          "Linkers": [
            "arg1_14890"
          ],
          "PartOfSpeech": "RB"
        }
      ]
    ]
}
```

Figure 3.5: Example snippet of `pdtb-parses.json` file necessary for running the J. Wang and Lan (2015) parser.

3.3.1 PDTB Parser Agreement

Given that each PDTB parser is also prone to segmentation and labelling errors, we computed their agreement across the two datasets. We have calculated parser agreement in a similar method to RST (see Section 3.2.2), with some slight differences:

(1) PDTB arguments are used instead of RST EDUs

(2) As these arguments are in list format, rather than tree format, the algorithm was simpler: For every (Arg1, Arg2) pair in parser *p1*, look for a corresponding exact match of (Arg1, Arg2) text in parser *p2*.

(3) As the previous step yielded a segmentation agreement of only 65.4% in ICNALE and 67.4% in CROW, we also looked for (Arg1, Arg2) text in parser *p2* which differed by 1 word. This boosted the agreement to 80.1% in ICNALE and 81.5% in CROW.

(4) As before, relations between every matching (Arg1, Arg2) in *p1* and (Arg1, Arg2) in *p2* are placed into a contingency table, shown in Table 3.11.

		Lin et al.				
		Contingency	**Expansion**	**Comparison**	**Temporal**	**Total**
Wang et al.	**Contingency**	**16876**	1322	284	663	19145
	Expansion	894	**18013**	231	544	19682
	Comparison	428	541	**9607**	320	10896
	Temporal	349	402	128	**7065**	7944
	Total	18547	20278	10250	8592	57667

Table 3.10: PDTB Parser agreement between the Lin et al. (2014) parser along the x-axis and the J. Wang and Lan (2015) parser on the y-axis, on the ICNALE dataset.

		Lin et al.				
		Contingency	**Expansion**	**Comparison**	**Temporal**	**Total**
Wang et al.	**Contingency**	**14509**	1290	225	672	16696
	Expansion	771	**15581**	179	469	17001
	Comparison	369	417	**8286**	243	9314
	Temporal	267	287	77	**6093**	6725
	Total	15916	17475	8767	7577	49735

Table 3.11: PDTB Parser agreement between the Lin et al. (2014) parser along the x-axis and the J. Wang and Lan (2015) parser on the y-axis, on the CROW dataset.

As expected, a 4-class classification performs a lot better than the classification for RST, with a 90.9% relation agreement for ICNALE and a 89.4% agreement for CROW. The remaining analysis was done only on the agreements, therefore 44469 instances of 49735 were used for CROW, and 51561 of 57667 were used for ICNALE.

3.3.2 PDTB Relations Across CEFR Levels

The results of the PDTB parser agreement were used to create an agreement between the two parsers. Table 3.12 shows the frequencies of each of the four top-level PDTB relations. Highlighted scores have a p-value below 0.05 when compared to C2 in a two-tailed t-test. As the table shows, no pattern of relation usage seems to exist across CEFR levels. Though all three categories of learner essays in EXPANSION are significantly different from C2 in both datasets, one is significantly higher, while the other is significantly lower.

		CONTINGENCY	t-test	EXPANSION	t-test	TEMPORAL	t-test	COMPARISON	t-test
ICNALE	A2	40.01	0.00	30.77	0.00	12.74	0.00	16.35	0.06
	B1	33.20	0.00	33.61	0.00	15.71	0.96	17.28	0.00
	B2	28.64	0.27	33.51	0.00	17.18	0.16	20.53	0.00
	C2	29.99	-	40.05	-	15.75	-	14.77	-
CROW	A2	25.51	0.15	36.63	0.00	15.52	0.00	20.93	0.89
	B1	27.25	0.01	35.47	0.01	16.93	0.00	19.95	0.31
	B2	26.28	0.04	35.01	0.01	16.99	0.00	20.69	0.69
	C2	23.68	-	31.81	-	21.94	-	21.07	-

Table 3.12: Frequencies of each top-level PDTB relation by dataset and CEFR score.

3.4 Inter-Framework Mapping

To compare the discourse relations across frameworks, we used the relation mapping proposed by Demberg et al. (2017). The mapping, shown in Table 3.13, does not account for all RST relations, as some relations, such as ATTRIBUTION, are generally not considered to be coherence discourse relations in other schemes such as PDTB.

PDTB level 1 relations	RST relations
TEMPORAL	TEMPORAL, BACKGROUND
CONTINGENCY	CAUSE, CONDITION, EXPLANATION
EXPANSION	ELABORATION, JOINT
COMPARISON	CONTRAST, COMPARISON

Table 3.13: Mapping of PDTB level 1 to RST relations proposed by Demberg et al. (2017).

A possible caveat to this mapping is that not all RST relations are covered by a PDTB 1st-level relation. A group of relations such as CAUSE + CONDITION + EXPLANATION represent a less complete relation group than their counterpart CONTINGENCY. We chose to do this mapping to see if there is an agreement between PDTB and RST parsers, but mapping between these two types of parser is currently imperfect.

We performed the inter-framework mapping discussed in Section 3.4 in order to directly compare RST and PDTB relations. As mentioned in Section 3.3.2, no PDTB top-level relation showed a pattern of usage consistent across datasets. However, for the ICNALE dataset, the PDTB relation of CONTINGENCY showed an interesting comparison with the RST relations of CAUSE + CONDITION + EXPLANATION. Though the CROW dataset did not match this pattern exactly, it also showed C2-level essays having the lowest frequency of CONTINGENCY relations.

Figure 3.6 compares the percentage of CONTINGENCY (the average of the two PDTB parsers) to the percentage of CAUSE + CONDITION + EXPLANATION (the average of the two RST parsers) on the ICNALE dataset. The left graph shows the frequency of the level 1 relation CONTINGENCY. The right graph shows the average frequency of CAUSE + CONDITION + EXPLANATION. "*" indicates a statistically significant difference with C2 essays. The mapping agrees with the pattern that emerges, in which A2 and B1-labelled texts show a statistically significant difference in frequency with C2 essays, whereas B2 essays do not.

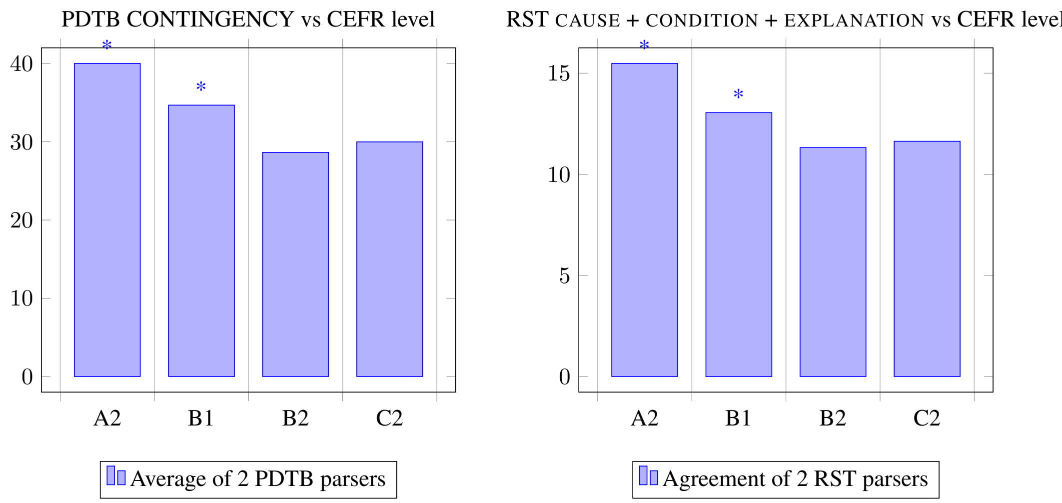

Figure 3.6: Percentage of CONTINGENCY across frameworks, in the ICNALE dataset.

3.5 Discourse Connective Analysis

The previous results compared only discourse relations. However, the PDTB parsers also provided other discourse information in their output, including the discourse connectives used with each relation, and the marking of each discourse connective as explicit or implicit.

For the sake of this analysis, we only looked at single discourse connectives connecting pairs of arguments (e.g. *but, however, in conclusion*), however in future analysis, paired discourse connectives (e.g. *on one hand... on the other hand*) could be considered as well.

To analyze the usage of discourse connectives and implicit vs explicit relations, every relation was put into a list in a .csv file, with the relation type, discourse connective, implicit/explicit marker, and CEFR level listed in each row. Using this, we were able to extract, for each PDTB relation, which discourse connectives are used most frequently by different levels of English learner. As for

Explicit vs Implicit relations, the PDTB parsers were not able to find implicit relations frequently enough to find useful data. Zhao and Webber (2022) discusses how parsers trained on PDTB-3.0 (Prasad et al., 2019) have improved in this regard, however those parsers are not currently easily accessible. The PDTB parsers showed an inability to classify enough implicit discourse connectives for an implicit vs explicit analysis to be significant. Lin et al. (2014) shows that the F1 score of implicit vs explicit classification is in the low 20s. The same researchers discuss in Lin, Kan, and Ng (2009) how this classification faces four main challenges: ambiguity of discourse relations, a parsers lack of ability to infer from a knowledge base, lack of context, and lack of world knowledge.

In addition to analyzing the discourse relations, we also looked at the distribution of discourse connectives.

For each level-2 discourse relation in the PDTB, we extracted all discourse connectives used to signal this relation, and calculated how frequently each connective was used by each CEFR level. A total of 51,561 discourse connective instances were found in the ICNALE dataset, and 44,469 were found in the CROW dataset. Details are shown in Appendix A.

The values in Table 3.14 show the frequency of the discourse connective used by the CEFR level to signal the discourse relation, divided by the total number of the discourse relation in the CEFR level. Each discourse connective in the table had a p-value less than 0.05 in at least two of the three t-tests: A2 vs C2, B1 vs C2, and B2 vs C2.

The ICNALE dataset, in general, showed a higher number of discourse connectives with statistical differences in frequency between learner essays and C2 essays. However, all 5 statistically different connectives in the CROW dataset were also statistically different in the ICNALE dataset. The smaller amount of statistically different discourse connectives in CROW dataset could be attributed to the smaller number of essays in the dataset.

3.6 Chapter Summary

In this chapter, we described our methodology to analyze discourse-level information across CEFR-level, as well as the results of this analysis. This included the presentation of the datasets

Relation	Connective	A2	B1	B2	C2
		ICNALE			
CONTINGENCY.cause	so	0.424	0.379	0.242	0.269
CONTINGENCY.cause	so that	0.021	0.045	0.036	0.105
COMPARISON.contrast	but	0.744	0.655	0.411	0.538
COMPARISON.contrast	though	0.026	0.046	0.051	0.122
COMPARISON.concession	nonetheless	0.013	0.018	0.026	0.132
TEMPORAL.asynchronous	then	0.361	0.298	0.192	0.190
TEMPORAL.asynchronous	after	0.210	0.225	0.180	0.150
TEMPORAL.synchrony	when	0.775	0.637	0.569	0.497
EXPANSION.restatement	overall	0.040	0.072	0.121	0.400
EXPANSION.restatement	rather	0.000	0.021	0.091	0.100
EXPANSION.restatement	in other words	0.480	0.338	0.364	0.100
		CROW			
CONTINGENCY.cause	so	0.389	0.316	0.209	0.173
COMPARISON.contrast	though	0.063	0.061	0.072	0.174
TEMPORAL.asynchronous	after	0.260	0.298	0.306	0.195
EXPANSION.restatement	rather	0.021	0.000	0.017	0.086
EXPANSION.restatement	in other words	0.426	0.422	0.506	0.286

Table 3.14: Discourse relations with a significant difference between frequencies in A2 vs C2-level essays.

in Section 3.1, the discourse parsing process and the results for the identification of discourse relations using RST (see Section 3.2) and PDTB (see Section 3.3), the inter-framework mapping (see Section 3.4), and the results for the identification of discourse connectives (see Section 3.5).

Overall, there seems to be a relation between learner CEFR level and the RST relations of EXPLANATION and BACKGROUND. Native speakers tend to have the lowest usage of CONTINGENCY relations in the PDTB. There seems to be an increase in the usage of the discourse connectives *though* and *rather* and a decrease in the usage of the discourse connectives *so*, *in other words*, and *after* as CEFR level increases.

In Chapter 4, we will explore how we used the results of this chapter in a machine learning process, and present the results of the classification.

Chapter 4

Machine Learning Classification

In this chapter, we will empirically validate our findings in Chapter 3 by measuring how discourse relations and connectives used as features can help to automatically assess the CEFR levels of essays. Section 4.1 will discuss our methodology, while Section 4.2 will show the results, with F1 scores of every model in the ablation study, and contingency tables of the best-performing models.

4.1 Classification Methodology

Previous work on the automatic assessment of CEFR-level shows that classification modelling yields better performance than regression modelling (Vajjala & Lõo, 2014). Given this, we decided to perform a classification task to empirically assess the CEFR level. As discussed in Section 3.1, ICNALE's B1 class was under-sampled to 1000 random essays.

We built upon the work of Montgomerie (2021) (see Section 2.1.3), who used standard machine learning with linguistic features for this task. Specifically, Montgomerie (2021) used syntactic information[1], part-of-speech tags, and readability level scores as features for SVM, Random Forest, and Logistic Regression models.

Montgomerie (2021) used a dataset which contained 1500 texts which were assigned a CEFR score by an automatic CEFR detector[2]. As both our datasets (ICNALE and CROW) contained

[1]Syntactic parse trees and part-of-speech information were found using spaCy linguistic tools. https://spacy.io/usage/linguistic-features

[2]https://textinspector.com/help/tu-lexical-profile/

	ICNALE	CROW
Gunning Fog	45.74%	38.25%
Flesch Kincaid Grade	45.00%	35.76%
Smog	43.53%	39.70%
Automated Readability	46.50%	37.03%
Text Standard	41.21%	38.53%
Dale Chall	44.22%	38.80%
Coleman Liau Index	40.79%	37.13%
Difficult Words	38.99%	42.74%
Linsear Write Formula	43.04%	38.50%
Random Baseline	25.00%	25.00%

Table 4.1: F1 score of ICNALE and CROW essays classified using only the readability scores as a baseline.

essays written by students who had taken a separate test to determine their TOEFL score (and thus CEFR level) we expect these datasets to pose a more difficult challenge to the models, as they were not labelled by a model that may have used similar features. In addition, the texts used by Montgomerie (2021) were written by native speakers for the purpose of readability by different language levels, rather than by language learners themselves, so though the methodology of this paper will be used in our work, a direct comparison of the results would not be useful.

Readability Scores, otherwise known as Text Complexity Metrics (Section 2.1.1) were used both as a baseline, and as features for the models. As a simple baseline, each text complexity metric for an N-class classifier had its output split into N equal sections. For example, if the lowest value on any essay for the Flesch Kincaid Grade Level was 2, the highest was 18, and N=4, the range of 16 values (2-18) would be split into fourths, with each ascending level outputting labels A2, B1, B2, and C2. This was shown to be particularly ineffective to classify learner texts, as outliers occurred in low-level essays with lack of punctuation. To remedy this, a robust scaler was used to ignore outliers, but overall the F1 score on this simple baseline was very low (See Table 4.1).

Three machine learning classifiers were chosen for their ease of use, usage in the previous work (Montgomerie, 2021), and availability with the scikit-learn package (Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E., 2011). The first classifier we chose is a Support Vector Classifier (SVC) since they are particularly good at handling high-dimensional data[3], and our methodology contains a high number of features. We used an RBF kernel to maximize its efficiency at the cost of processing time. As a second classifier, we chose a Random Forest Classifier, as the combination of multiple decision trees helps to reduce the risk of overfitting and improve the accuracy of the model. Random Forests handle both numerical and categorical features well, which fits with our choice of features. Though more common in binary classification tasks, we chose a Logistic Regression model as our third classifier, since the relationship between the progression of CEFR levels is linear.

Each model was trained on 5 types of features. These 5 feature types can be further classified into 2 groups: surface and syntactic-level features, and discourse-level features.

The full list of features includes:

(1) **Syn**: These include syntactic and part of speech features. We used the average per-sentence frequency of 18 part of speech tags[4], and one feature for average syntactic parse tree depth.[5]

(2) **Read**: Eight Readability scores (see Section 2.1.1) were used: We used the SMOG Index (McLaughlin, 1969), Flesch-Kincaid Index (Flesch, 1943), Coleman-Liau Index (Coleman & Liau, 1975), Automated Readability Index (Smith & Senter, 1967), Dale Chall Readability Score (Dale & Chall, 1948), Linsear Write Formula (O'Hayre, 1966), Gunning Fog Index (Gunning, 1952), and Textstat Text Standard (Bansal, 2016)[6].

As additional features, we incorporated the results of our discourse analysis (see Chapter 3) and created new features to model discourse relations and connectives. These new features were intended to capture the underlying discourse structure of the text.

(3) **RST**: For ICNALE, average per-sentence frequency of 8 RST relations with statistical difference in at least one t-test: ATTRIBUTION, BACKGROUND, CONTRAST, COMPARISON, CONDITION, ELABORATION, EXPLANATION, and JOINT. For CROW, average per-sentence

[3] `https://scikit-learn.org/stable/modules/svm.html`

[4] The POS tags include *Adjective, Adposition, Adverb, Auxiliary verb, Conjunction, Coordinating conjunction, Determiner, Interjection, Noun, Numeral, Particle, Pronoun, Proper noun, Punctuation, Subordinating conjunction, Symbol, Verb,* and *Space.*

[5] We used the Dependency Parser from spaCy `https://spacy.io/api/dependencyparser` to extract POS tags and syntactic parses.

[6] Readability scores were found using the Textstat library `https://pypi.org/project/textstat/`

frequency of 6 RST relations with statistical difference in at least one t-test: ATTRIBUTION, BACKGROUND, CONTRAST, COMPARISON, EXPLANATION, and JOINT.

(4) **PDTB**: Average per-sentence frequency of the 4 level-1 PDTB relations: CONTINGENCY, EXPANSION, COMPARISON, and TEMPORAL.

(5) **Con:** Discourse connectives Average per-sentence frequency of the 11 most discriminating discourse connectives for ICNALE and 5 discourse connectives for CROW (See Table 3.14).

To measure the contribution of each type of feature, each of the 5 categories was used individually, and features were added gradually until all were used to create a final model with relations and a final model with connectives. This process was repeated for ICNALE data and CROW data.

For feature engineering, some code design choices had to be made. As the syntactic-level features found by the SpaCy library (Honnibal & Montani, 2017) and Textstat[7] were very quick to calculate, they were recalculated for each essay dynamically while training and testing. Discourse-level features, on the contrary, take too long to extract dynamically, so a Python script was created to match each essay with its RST and PDTB parses to be placed into the same row in either the test or train files. The input feature array sent to the models contained the full text of the essay concatenated with its RST or PDTB parse, while the output array contained the CEFR label. In pre-processing, each input value was then split into its essay and its parse, for the essay to be sent to have surface and syntactic-level features extracted, while the RST or PDTB parses were sent to compute the frequency of every relation, simply by looking for instances of each relation in the string. This thread was also sent information from the essay containing the number of sentences in the essay, so that a mean could be calculated for each relation with respect to the number of sentences in the essay. The features were then combined to obtain an overall CEFR score prediction.

Two separate machines were used: A laptop computer running Windows 10 with an Intel® Core™ i5-9300H CPU @ 2.40GHz with 4 cores, and a desktop computer running Ubuntu 20.04.4 LTS with an Intel® Core™ i7-4770 CPU @ 3.40GHz with 4 cores. The training of each of the 3 models took between 5 and 6 minutes per document on the ICNALE dataset containing 1120 test documents and 4480 training documents. Models were saved in *.joblib* files, with SVC model

[7]https://pypi.org/project/textstat/

files ranging from 317KB - 1332KB, logistic regression model files ranging from 5KB-6KB, and Random Forest Classifier model files ranging from 3MB-22MB.

We used 5-fold cross-validation to obtain an average F1 score for each class.

For RST features, we had also experimented with using all of our 12 chosen RST relations as features, however we had found, in 5 out of 6 cases, as shown in Table 4.2, that using only the relations with statistical difference in at least one t-test showed an improved result in classification.

		Dataset	
Model	**Features**	**ICNALE**	**CROW**
SVM	Syn + Read + RST (All 12)	53.76 ± 1.54	56.05 ± 3.99
SVM	Syn + Read + RST (Only significant)	**53.95 ± 1.64**	**56.35 ± 3.01**
Logistic Regression	Syn + Read + RST (All 12)	53.04 ± 1.99	**57.53 ± 4.09**
Logistic Regression	Syn + Read + RST (Only significant)	**53.26 ± 1.73**	57.41 ± 3.59
Random Forest	Syn + Read + RST (All 12)	51.75 ± 0.98	59.09 ± 3.76
Random Forest	Syn + Read + RST (Only significant)	**53.14 ± 0.60**	**60.25 ± 4.47**

Table 4.2: Results of classification using syntactic features, readability score features, and RST features, comparing the use of all 12 relations versus the use of relations with at least some statistical difference between C2 levels and learners.

4.2 Classification Results

4.2.1 Overall Results

Table 4.3 shows the results of the classification for all 5 categories of features. Overall, the classification into 4 CEFR levels range between 39% to 61% F-measure. As Table 4.3 shows, for each model, either adding syntactic features to discourse relations or discourse connectives yielded the best performance, though most results are within the standard deviation. Between RST and PDTB features, PDTB features seem to provide a greater improvement, but the PDTB level 1 relations are more general than the more specific RST relations. This may be due to the more challenging discourse parsing into fine-grained relations or that, as Section 3.2 shows, apart from EXPLANATION and BACKGROUND, the usage of RST discourse relations do not vary greatly across CEFR levels.

As an additional point of comparison, we used RoBERTa from HuggingFace[8] as a standalone CEFR classifier. We trained a model for a language proficiency classification task using the RoBERTa transformer-based model. The model is fine-tuned on our datasets using the PyTorch deep learning framework. We loaded the dataset using pandas and preprocessed it with the LabelEncoder class from scikit-learn. Our training function trains the model using the Trainer class from the transformers library, which provides built-in capabilities for training and evaluation. We then used the resulting trained model to predict the language proficiency of texts in a test dataset.

RoBERTa was able to produce a significantly higher performance on both datasets, reaching F-measures of 59.69 and 67.56 without any feature engineering. This last experiment was performed as a comparison point, but without using explicit features, it is hard to measure of the contribution of discourse features to CEFR assessment.

4.2.2 Contingency Matrices of Highest-Performing Models

The contingency matrices for the six top-performing models (SVC, Logistic Regression, and Random Forest) on both ICNALE and CROW datasets are presented in the following tables.

Table 4.4 displays the results of the SVC model for both ICNALE and CROW datasets. In the ICNALE table, the classifier encounters challenges in distinguishing between A2 and B1 essays. On the other hand, the CROW table indicates that it may have been beneficial to under-sample the B2 essays in our classification. However, a majority of the essays classified as B2 are actually B1-level essays, differing only by one proficiency level. Table 4.5 for Logistic Regression and Table 4.6 for Random Forest follow a similar trend. In general, the models have the most difficulty distinguishing between B1 and B2 essays.

4.3 Chapter Summary

In this chapter, we set out to validate our previous findings through an empirical validation. Our goal was to automatically assess the CEFR levels of essays by using discourse relations and connectives as features. To achieve this, we performed a classification task and an ablation study.

[8] https://huggingface.co/docs/transformers/model_doc/roberta

In Section 4.1, we discussed the choice of classification modelling over regression modelling based on previous work in CEFR-related tasks. We built upon the work of Montgomerie (2021), who employed standard machine learning techniques with linguistic features for CEFR classification. Our methodology incorporated syntactic information, part-of-speech, readability scores, and discourse-level features.

We chose three machine learning classifiers: Support Vector Classifier (SVC), Random Forest Classifier, and Logistic Regression model. These models were trained on six different categories of features, which were further classified into syntactic-level features and discourse-level features. We used an ablation study of these features to automatically classify the texts.

In Section 4.2, we presented the performance of the models. We observed that the inclusion of discourse features, particularly PDTB features, led to a marginal increase in F-measure. However, the results were mostly within the standard deviation. We also compared our models with a standalone CEFR classifier based on the RoBERTa model, which yielded significantly high performance on both datasets.

Overall, our findings demonstrated that the inclusion of discourse features might have a mild improvement in classification performance. The mildness of the improvement relates to the findings in Chapter 3 that show most RST and PDTB relations did not show glaring differences among CEFR levels. Furthermore, the comparison with the RoBERTa model highlighted the superiority of large language models in CEFR classification tasks, even when additional features are incorporated.

		Dataset	
Model	**Features**	**ICNALE**	**CROW**
SVM	Syn	48.77 ± 2.08	49.19 ± 2.57
SVM	Read	49.01 ± 1.08	55.54 ± 3.75
SVM	Syn + Read	53.28 ± 1.97	55.57 ± 3.72
SVM	RST	40.06 ± 2.57	47.08 ± 3.59
SVM	PDTB	40.19 ± 1.98	46.00 ± 2.61
SVM	Conn	38.23 ± 1.57	46.15 ± 2.86
SVM	Syn + Read + RST	53.95 ± 1.64	56.35 ± 3.01
SVM	Syn + Read + PDTB	54.35 ± 2.52	56.86 ± 3.32
SVM	Syn + Read + RST + PDTB	55.01 ± 2.12	**57.10 ± 3.81**
SVM	Syn + Read + Conn	**55.70 ± 1.93**	55.16 ± 3.62
Logistic Regression	Syn	45.75 ± 2.78	49.98 ± 3.80
Logistic Regression	Read	46.43 ± 1.74	55.17 ± 4.10
Logistic Regression	Syn + Read	52.42 ± 2.63	56.85 ± 3.97
Logistic Regression	RST	39.54 ± 1.57	46.06 ± 2.84
Logistic Regression	PDTB	40.45 ± 1.44	45.69 ± 2.30
Logistic Regression	Conn	37.15 ± 1.69	46.11 ± 2.54
Logistic Regression	Syn + Read + RST	53.26 ± 1.73	57.41 ± 3.59
Logistic Regression	Syn + Read + PDTB	54.13 ± 2.61	58.55 ± 4.56
Logistic Regression	Syn + Read + RST + PDTB	54.81 ± 2.19	59.17 ± 3.94
Logistic Regression	Syn + Read + Conn	**55.90 ± 2.57**	**59.55 ± 4.26**
Random Forest	Syn	49.78 ± 0.84	53.37 ± 4.40
Random Forest	Read	47.80 ± 1.52	59.10 ± 3.14
Random Forest	Syn + Read	52.90 ± 1.39	60.49 ± 4.28
Random Forest	RST	39.02 ± 1.12	52.55 ± 2.23
Random Forest	PDTB	39.93 ± 2.03	48.77 ± 1.55
Random Forest	Conn	39.17 ± 1.86	55.94 ± 4.11
Random Forest	Syn + Read + RST	53.14 ± 0.60	60.25 ± 4.47
Random Forest	Syn + Read + PDTB	53.74 ± 1.86	61.15 ± 4.40
Random Forest	Syn + Read + RST + PDTB	**54.54 ± 2.60**	**62.19 ± 4.02**
Random Forest	Syn + Read + Conn	53.29 ± 2.75	61.28 ± 3.57
RoBERTa	N/A	59.69 ± 1.99	67.56 ± 2.11

Table 4.3: Results of the classification, showing average F1 on 5-fold cross-validation ± standard deviation.

		ICNALE - SVC			
		Predicted			
		A2	**B1**	**B2**	**C2**
Actual	**A2**	**69.06% ± 3.58%**	27.29% ± 5.03%	0.94% ± 0.93%	2.71% ± 1.13%
	B1	33.3% ± 3.11%	**58% ± 3.87%**	4.9% ± 1.78%	3.8% ± 0.45%
	B2	18.97% ± 6.12%	52.8% ± 7.88%	**20.91% ± 3.46%**	7.33% ± 1.77%
	C2	5.24% ± 2.7%	18.45% ± 3.57%	8.73% ± 4.23%	**67.58% ± 7.6%**

		CROW - SVC			
		Predicted			
		A2	**B1**	**B2**	**C2**
Actual	**A2**	**46.02% ± 3.03%**	2.21% ± 2.31%	30.77% ± 12.27%	20.99% ± 5.7%
	B1	0.4% ± 0.89%	**45.14% ± 3.02%**	42.11% ± 7.89%	12.35% ± 11.3%
	B2	7.46% ± 2.48%	9.27% ± 3.11%	**82.46% ± 2.94%**	0.81% ± 0.52%
	C2	7.52% ± 5.94%	12.03% ± 11.71%	15.04% ± 17.02%	**65.41% ± 12.02%**

Table 4.4: Contingency matrix with standard deviation for the 5-fold cross validation, for the two highest-performing SVM models: Syn + Read + Conn for ICNALE and Syn + Read + RST + PDTB for CROW.

		ICNALE - Logistic Regression			
		Predicted			
		A2	**B1**	**B2**	**C2**
Actual	**A2**	**65.42% ± 4.75%**	30.31% ± 7.29%	1.88% ± 1.08%	2.4% ± 0.47%
	B1	31.6% ± 3.73%	**55.5% ± 5.58%**	8.1% ± 3.56%	4.8% ± 0.57%
	B2	18.53% ± 7.36%	42.46% ± 9.29%	**30.82% ± 1.96%**	8.19% ± 1.63%
	C2	2.74% ± 1.63%	18.2% ± 5.03%	11.97% ± 5.48%	**67.08% ± 4.17%**

		CROW - Logistic Regression			
		Predicted			
		A2	**B1**	**B2**	**C2**
Actual	**A2**	**38.65% ± 2.15%**	1.44% ± 1.32%	51.25% ± 14.16%	8.65% ± 4.02%
	B1	0.9% ± 2.02%	**37.24% ± 4.05%**	60.95% ± 14.96%	0.9% ± 1.24%
	B2	8.97% ± 3.72%	9.5% ± 2.66%	**80.49% ± 2.28%**	1.04% ± 0.95%
	C2	8.27% ± 4.9%	0% ± 0%	21.8% ± 11.4%	**69.92% ± 24.45%**

Table 4.5: Contingency matrix with standard deviation for the 5-fold cross validation, for the two highest-performing Logistic Regression models: Syn + Read + Conn for ICNALE and CROW.

		ICNALE - Random Forest			
		Predicted			
		A2	B1	B2	C2
Actual	A2	**63.33% ± 3.35%**	30.94% ± 6.98%	2.81% ± 0.79%	2.92% ± 0.87%
	B1	32.3% ± 3.91%	**53% ± 2.87%**	8.3% ± 3.75%	6.4% ± 1.19%
	B2	18.32% ± 6.14%	45.47% ± 7.97%	**27.37% ± 2.91%**	8.84% ± 3.27%
	C2	5.49% ± 3.48%	16.96% ± 2.87%	9.73% ± 3.78%	**67.83% ± 5.82%**

		CROW - Random Forest			
		Predicted			
		A2	B1	B2	C2
Actual	A2	**36.8% ± 4.66%**	1.94% ± 2.03%	52.52% ± 18.88%	8.74% ± 4.41%
	B1	0.45% ± 1.01%	**47.19% ± 5.43%**	51.9% ± 8.96%	0.45% ± 1.01%
	B2	8.46% ± 0.48%	8.27% ± 1.11%	**82.46% ± 2.94%**	0.81% ± 0.52%
	C2	3.76% ± 4.6%	0% ± 0%	26.32% ± 10.3%	**69.92% ± 21.03%**

Table 4.6: Contingency matrix with standard deviation for the 5-fold cross validation, for the two highest-performing Random Forest models: Syn + Read + RST + PDTB for ICNALE and CROW.

Chapter 5

Conclusions and Future Work

In this thesis, we investigated the use of discourse information in essays across language proficiency levels. A corpus analysis with state-of-the-art RST and PDTB parsers showed a marginal relation between learner CEFR level and the RST relations of EXPLANATION and BACKGROUND. Using the mapping of PDTB and RST proposed by Demberg et al. (2017), we showed a decrease in use of CONTINGENCY relations as the CEFR level increases. When used as additional features to automatically determine the CEFR level of learner essays, features measuring the frequency of RST and PDTB relations, as well as discourse connectives, lead to a mild improvement in performance, though adding these features is not enough for traditional models to outperform large language models such as RoBERTa.

5.1 Contributions

This thesis makes several significant contributions to the field of Natural Language Processing (NLP) and language assessment for English learners.

Firstly, in Chapter 3, a comprehensive analysis of discourse structures in English learner essays is presented, focusing on the Rhetorical Structure Theory (RST) and Penn Discourse Treebank (PDTB) frameworks. By examining the usage of discourse relations and connectives across Common European Framework of Reference (CEFR) levels, distinct patterns and trends that indicate varying levels of language proficiency are identified.

Additionally, an investigation is conducted in Section 3.2.2 and Section 3.3.1 to assess the agreement between widely used PDTB and RST parsers. This analysis provides valuable insights for future users of these discourse parsers, aiding in their selection and implementation.

The analysis of RST and PDTB discourse relation patterns in Chapter 3 uncovers specific challenges that English learners face at different proficiency levels. These findings contribute to the development of targeted teaching materials and strategies, enabling educators to address these challenges and enhance language learning outcomes.

Furthermore, Chapter 4 explores the feasibility of using discourse relations and connectives as features for machine learning models to automatically assess the proficiency levels of English learner essays. Through experiments with Random Forest, Support Vector Machine, and Logistic Regression models, the potential of these features to enhance accurate and efficient language assessment is measured.

Overall, this thesis aims to drive progress in the field of NLP, facilitate language learning and assessment, and contribute to the development of more effective and intelligent language technologies.

5.2 Limitations

This section discusses the limitations of the research conducted in the thesis, highlighting potential challenges and biases that could affect the findings of the results.

The study employed two separate RST (Rhetorical Structure Theory) and PDTB (Penn Discourse Treebank) parsers to calculate the agreement for discourse information. However, it is important to acknowledge that these parsers were trained on corpora primarily composed of text written by fluent English speakers, such as Wall Street Journal articles. While every effort was made to ensure accurate parsing, it is essential to recognize that the parsers might not achieve perfect accuracy, as they may not fully capture the nuances and variations in English usage by non-native speakers or speakers from diverse linguistic backgrounds.

The research heavily relied on data from the ICNALE (International Corpus Network of Asian

Learners of English) A2-B2 essays, which exclusively represent English learners from Asian countries. Furthermore, the CROW (Corpus of Research Writing) dataset prominently featured learners from the five most represented native countries in Asia. This heavy bias towards Asian learners of English poses a potential limitation in terms of the broad applicability of the findings. It is important to acknowledge that learners from African, South American, or European countries may exhibit different patterns and behaviours when it comes to discourse information. Consequently, if the same experiment were to be conducted with learners from these regions, the results might differ significantly.

Although we employed a consistent mapping of TOEFL scores to CEFR levels across all datasets, there are certain concerns that need to be addressed. Firstly, the mapping used in ICNALE categorizes any essay written by a non-native speaker with a TOEFL score above 87 as B2. Consequently, we adopted this same classification for the CROW dataset. However, this approach introduces a potential issue where essays might be labelled as B2 even though they would technically fall under the C1 classification.

5.3 Future Work

In this section, we will discuss potential future applications of the analysis and data generated from this research, along with avenues for further exploration.

To enhance the broad applicability and accuracy of the results, future studies should consider incorporating data from a more diverse range of English learners, taking into account their varying linguistic backgrounds. This can be achieved by leveraging parsers trained on corpora that encompass a broader range of linguistic sources. By incorporating a more diverse set of learners and linguistic backgrounds, the findings can be more robust and representative of the broader English learner population.

Similarly, future work could explore replicating the methodology employed in this study for learners of languages other than English. However, it should be acknowledged that the availability of CEFR-labelled learner corpora in English, despite being the most over-represented language in NLP research (Søgaard, 2022), was already challenging. Therefore, the task of finding similar corpora in other languages might prove even more difficult. Nevertheless, investigating discourse analysis and language assessment methodologies in languages beyond English would contribute to a more comprehensive understanding of language learning and assessment across diverse linguistic contexts. It could provide valuable insights into the transferability of findings and the development of tailored language assessment models for learners of different languages. Efforts should be made to expand the resources and availability of CEFR-labelled corpora in various languages to facilitate such future research endeavours.

Moreover, it would be valuable for future work to investigate differences in discourse relations based on the first language of English learners while considering their CEFR levels. The corpora used in this study provide information on the native language or country of origin of the learners, offering an opportunity to explore the influence of first language on discourse patterns. Appendix B provides some preliminary work towards this topic. This line of inquiry aligns with the suggestions put forth by Perkins (2014), who proposed that first language identification could be useful in various contexts, including author identification in criminal investigations. Therefore, empirically validating the impact of first language on discourse relations in English learner essays would provide valuable insights into language transfer and interlanguage development.

Expanding on the analysis of discourse relations, future work could explore the usage of explicit versus implicit relations. Investigating these additional aspects of discourse information across different levels of language proficiency would provide a more comprehensive understanding of English learners' discourse abilities and enrich our insights into their language learning process.

Furthermore, the availability of PDTB-3.0 (Prasad et al., 2019) parsers opens up new avenues for analysis. The updated corpus, along with previous work mapping RST and PDTB-3.0 relations (Costa, Sheikh, & Kosseim, 2023), provides an opportunity to investigate discourse relations across various textual genres using the latest corpus resources. Incorporating PDTB-3.0 relations into the analysis can contribute to a more comprehensive understanding of discourse structures and their usage by English learners.

In addition, it would be interesting to explore the development of discourse parsers specifically trained on learner texts. While this thesis used RST and PDTB parsers trained on writings by fluent English speakers, such as Wall Street Journal articles, the creation of parsers tailored to effectively parse discourse information from learner texts could yield more accurate and insightful results. This specialized parsing approach would enhance our understanding of discourse usage by English learners and better capture the nuances of learner-specific language.

Future research can expand its focus beyond discourse analysis in argumentative texts and delve into discourse structures across various text genres, including narratives, academic papers, and conversational dialogues. Notably, recent work has explored this avenue in the realm of spontaneous spoken dialogue (López Cortez & Jacobs, 2023). By extending the examination of discourse relations and connectives to diverse genres, a more comprehensive understanding of language learning can be achieved, shedding light on genre-specific discourse patterns.

This work ties into current trends in NLP. While, in recent years, CEFR assessment has begun to be taken over by Large Language Models (LLMs), and has yielded effective results on automated essay scoring (Naismith et al., 2023), traditional machine learning models remain relevant for this task. Firstly, LLMs are computationally expensive and require significant computational resources to train and deploy. Thus, systems with limited computing power must rely on lower-cost models, such as traditional machine learning models, to complete tasks. Additionally, applications of traditional machine learning models can be used as baselines for the further study of LLMs.

Current CEFR assessment tools currently struggle with polysemy disambiguation– the disambiguation of words that have multiple meanings. Settles, LaFlair, and Hagiwara (2020) discussed the problem of accurately categorizing lexical items while developing Duolingo's CEFR-checker, and suggests using two regression models that are trained on lexical item representations made up of surface-based features. With the addition of discourse-based features, focusing on the context of sentences rather than surface-based features, future work could attempt to solve this polysemy disambiguation problem. Recent work has attempted to tackle this issue for usage in the creation of level-appropriate pedagogical content (Aleksandrova & Pouliot, 2023), by examining the most common usage of polysemic words in language learner essays using BERT, for the task of creating a CEFR-based classifier designed for English and French to assess the lexical complexity of single-word and multi-word expressions in context, as tool for the language learning application Mauril[1]. This work expanded upon recent previous work (Garí Soler & Apidianaki, 2021) which found that

[1]https://mauril.ca/

BERT was able to automatically create embeddings for polysemic words, especially the uncased model.

Recent previous work has explored numerous handcrafted features for examining texts at a discourse and surface level. Lee and Lee (2023) presents a list of 220 handcrafted features commonly used in computational linguistics. The list does contain the syntactic features we chose in Chapter 4, save for a few of the readability scores, however there are many more features that could be explored in combination with our discourse features.

Other LLMs, such as Google's T5[2] and OpenAI's GPT-2, can be fine-tuned for CEFR classification tasks (Roos & Sidorova, 2022). While our work focused on using discourse as features for traditional machine learning models, we could use these features to fine-tune LLMs as well. Although this work used OpenAI's GPT-2, more up-to-date versions of GPT require human fine-tuning as well (Young & Shishido, 2023). As for our own investigation into this topic, we tested ChatGPT's understanding to see if ChatGPT[3] was capable of automatic generation of English-learner texts. The response given in Figure 5.1, written in perfect fluent English, implies that ChatGPT doesn't fully understand the concept of an A2-level learner, and that fine-tuning of this model on CEFR-related tasks is necessary.

By considering these suggested avenues for future work, researchers can continue to advance the field of discourse analysis, expand the knowledge base on language learning, and contribute to the development of effective language assessment tools and instructional approaches.

[2] https://github.com/google-research/text-to-text-transfer-transformer
[3] https://openai.com/blog/chatgpt

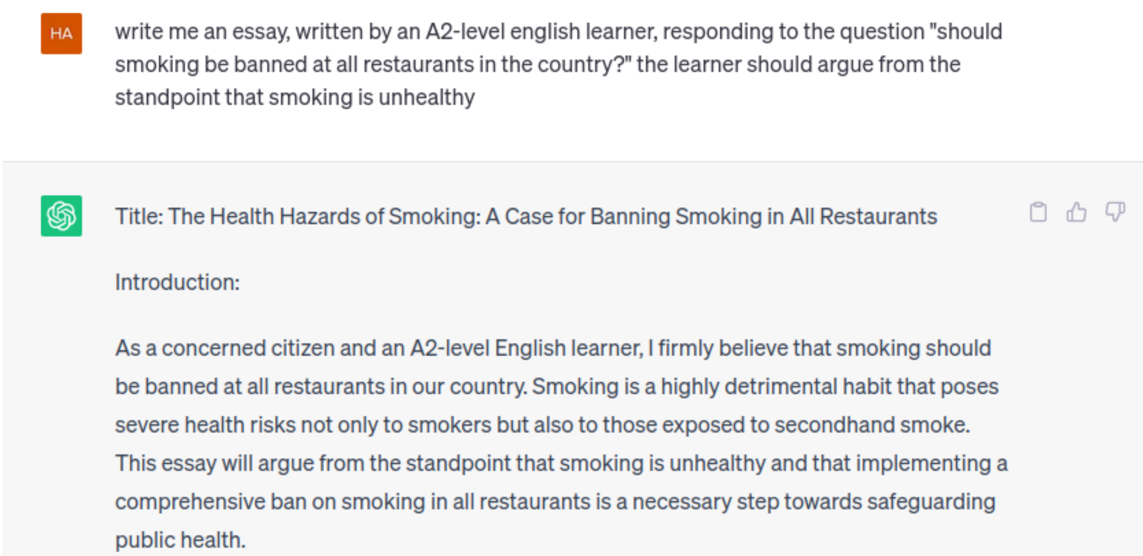

Figure 5.1: Example of an attempt at using ChatGPT (powered by GPT-3.5) to generate an essay from an A2-level English learner

References

Abdalla, M., Rudzicz, F., & Hirst, G. (2018). Rhetorical structure and Alzheimer's disease. *Aphasiology*, *32*(1), 41-60. Retrieved from https://doi.org/10.1080/02687038.2017.1355439

Aleksandrova, D., & Pouliot, V. (2023, July). CEFR-based Contextual Lexical Complexity Classifier in English and French. In *Proceedings of the 18th workshop on innovative use of nlp for building educational applications (bea 2023)* (pp. 518–527). Toronto, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.bea-1.43

Aoyama, T. (2022). Comparing native and learner englishes using a large pre-trained language model. In *Proceedings of the 11th Workshop on NLP for Computer Assisted Language Learning* (pp. 1–9). Louvain-la-Neuve, Belgium: LiU Electronic Press. Retrieved from https://aclanthology.org/2022.nlp4call-1.1

Bachand, F.-H., Davoodi, E., & Kosseim, L. (2014, April). An Investigation on the Influence of Genres and Textual Organisation on the Use of Discourse Relations. In *International Conference on Intelligent Text Processing and Computational Linguistics (CICLING-2014)* (p. 454-468). doi: 10.1007/978-3-642-54906-9_37

Bansal, S. (2016). *textstat: Statistical Text Analysis in Python.* https://pypi.org/project/textstat/. (Accessed: March 1, 2023)

Breiman, L. (2001). Random Forests. *Machine learning*, *45*(1), 5–32.

Browning, J. (2017, April). *Using Machine Learning Techniques to Identify the Native Language of an English User* (Tech. Rep.). University of Edinburgh School of Informatics. Retrieved from https://project-archive.inf.ed.ac.uk/ug4/20170963/ug4_proj.pdf

Bunt, H., & Prasad, R. (2016). Core Concepts for the Annotation of Discourse Relations. In *Proceedings 10th Joint ACL-ISO Workshop on Interoperable Semantic Annotation* (pp. 45–54). Portoroz, Slovenia. Retrieved from https://sigsem.uvt.nl/isa12/

Carlson, L., & Marcu, D. (2001, September 11). *Discourse Tagging Reference Manual* (Technical Report No. ISI-TR-2001-003). Information Sciences Institute, University of Southern California. Retrieved from https://www.isi.edu/~marcu/discourse/tagging-ref-manual.pdf

Carlson, L., Marcu, D., & Okurowski, M. E. (2002). *RST Discourse Treebank.* Borealis. Retrieved from https://doi.org/10.5683/SP2/TJQZHH doi: 10.5683/SP2/TJQZHH

Chiarcos, C. (2014, May). Towards interoperable discourse annotation. Discourse Features in the Ontologies of Linguistic Annotation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 4569–4577). Reykjavik, Iceland: European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2014/pdf/893_Paper.pdf

Coleman, J. P., & Liau, T. L. (1975). A Computer Program for Analyzing Readability. *Journal of Applied Psychology*, *60*(2), 283-284. Retrieved from https://psycnet.apa.org/record/1975-20529-001 doi: 10.1037/h0076860

Cortes, C., & Vapnik, V. (1995, September). Support-Vector Networks. *Mach. Learn.*, *20*(3), 273–297. Retrieved from https://doi.org/10.1023/A:1022627411411

Costa, N. F., Cheng, Y., Muermans, T. C., Hanel, B., & Kosseim, L. (2023, February). Automatic Identification of Chinese Paired Discourse Connectives. In *Proceedings of the 17th International Conference on Semantic Computing.* Laguna Hills, California, USA.

Costa, N. F., Sheikh, N., & Kosseim, L. (2023, September). Mapping explicit and implicit discourse relations between the RST-DT and the PDTB 3.0. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2023).* Varna, Bulgaria.

Dale, E., & Chall, J. S. (1948). A Formula for Predicting Readability. *Educational Research Bulletin*, *27*(1), 11–28. Retrieved 2023-01-11, from http://www.jstor.org/stable/1473169

Davoodi, E. (2017). *Computational Discourse Analysis Across Complexity Levels* (Doctoral dissertation, Concordia University Department of Computer Science and Software Engineering). Retrieved from https://spectrum.library.concordia.ca/id/eprint/982967/

Demberg, V., Asr, F. T., & Scholman, M. C. J. (2017). How consistent are our discourse annotations? Insights from mapping RST-DT and PDTB annotations. *CoRR*, *abs/1704.08893*. Retrieved from http://arxiv.org/abs/1704.08893

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, *abs/1810.04805*. Retrieved from http://arxiv.org/abs/1810.04805

Feng, V. W., & Hirst, G. (2014). A Linear-Time Bottom-Up Discourse Parser with Constraints and Post-Editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL-2014)* (pp. 511–521). Baltimore, Maryland. Retrieved from https://aclanthology.org/P14-1048 doi: 10.3115/v1/P14-1048

Flesch, R. (1943). *Marks of Readable Style: A Study in Adult Education*. Teachers College, Columbia University. Retrieved from https://books.google.ca/books?id=JSIWAAAAIAAJ

Garí Soler, A., & Apidianaki, M. (2021). Let's Play Mono-Poly: BERT Can Reveal Words' Polysemy Level and Partitionability into Senses. *Transactions of the Association for Computational Linguistics*, *9*, 825–844. Retrieved from https://aclanthology.org/2021.tacl-1.50 doi: 10.1162/tacl_a_00400

Geertzen, J., Alexopoulou, T., Korhonen, A., et al. (2013). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). In *Proceedings of the 31st Second Language Research Forum (SLRF-2013)* (p. 240-254). Somerville, MA: Cascadilla Proceedings Project.

Granger, S. (1998). The computer learner corpus: A versatile new source of data for SLA research. In *Learner English on Computer* (p. 3-18). Addison Wesley Longman: London & New York.

Granger, S., Dupont, M., Meunier, F., Naets, H., & Paquot, M. (2020). *International Corpus of Learner English. Version 3*. Published by Presses Universitaires de Louvain.

Gunning, R. (1952). *The Technique of Clear Writing*. McGraw-Hill. Retrieved from https://books.google.ca/books?id=Zl5bAAAAMAAJ

Hanel, B., & Kosseim, L. (2023, September). Discourse Analysis of Argumentative Essays of English Learners based on their CEFR Level. In *Proceedings of Recent Advances in Natural Language Processing*. Varna, Bulgaria.

Hastie, Trevor and Tibshirani, Robert and Friedman, Jerome. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. Retrieved from https://hastie.su.domains/Papers/ESLII.pdf

Heilman, M., & Sagae, K. (2015). Fast Rhetorical Structure Theory Discourse Parsing. *Computing Research Repository*, *abs/1505.02425*. Retrieved from http://arxiv.org/abs/1505.02425

Hernault, H., Prendinger, H., duVerle, D., & Ishizuka, M. (2010, December). HILDA: A Discourse Parser Using Support Vector Machine Classification. *Dialogue & Discourse; Vol 1, No 3 (2010)*. doi: 10.5087/dad.2010.003

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.* Published by Sentometrics Research.

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. In *Learner corpus studies in Asia and the world* (Vol. 1, p. 91-118). Retrieved from https://language.sakura.ne.jp/icnale/

Jasinskaja, K., & Karagjosova, E. (2020, November). Rhetorical relations. In *The Wiley Blackwell Companion to Semantics* (p. 1-29). doi: 10.1002/9781118788516.sem061

Ji, Y., & Eisenstein, J. (2014). Representation Learning for Text-level Discourse Parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL-2014)* (pp. 13–24). Baltimore, Maryland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P14-1002 doi: 10.3115/v1/P14-1002

Joty, S., Carenini, G., Ng, R., & Mehdad, Y. (2013, August). Combining Intra- and Multi-sentential Rhetorical Parsing for Document-level Discourse Analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL-2013* (pp. 486–496). Sofia, Bulgaria: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P13-1048

Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. USA: Prentice Hall PTR. Retrieved from https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf (Accessed February 13, 2023)

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. *Institute for Simulation and Training, University of Central Florida*. Retrieved from https://stars.library.ucf.edu/istlibrary/56/

Kornai, A., & Tuza, Z. (1992). Narrowness, Path-width, and their Application in Natural Language Processing. In *Discrete applied mathematics* (Vol. 36, p. 87-92).

Lacelle-Peterson, M., & Rivera, C. (1994, February). Is It Real for All Kids? A Framework for Equitable Assessment Policies for English Language Learners. *Harvard Educational Review*, *64*(1), 55-76. Retrieved from https://doi.org/10.17763/haer.64.1.k3387733755817j7 doi: 10.17763/haer.64.1.k3387733755817j7

Lee, B. W., & Lee, J. (2023, July). LFTK: Handcrafted Features in Computational Linguistics. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 1–19). Toronto, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.bea-1.1

Li, Q., Li, T., & Chang, B. (2016). Discourse Parsing with Attention-based Hierarchical Neural

Networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP-2016)* (pp. 362–371). Austin, Texas: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D16-1035 doi: 10.18653/v1/D16-1035

Liaw, A., & Wiener, M. (2002). Classification and Regression by Random Forest. *R News*, *2*(3), 18–22.

Lin, Z., Kan, M.-Y., & Ng, H. T. (2009). Recognizing Implicit Discourse Relations in the Penn Discourse Treebank. In *Conference on Empirical Methods in Natural Language Processing(EMNLP-2009).*

Lin, Z., Ng, H. T., & Kan, M. (2014). A PDTB-Styled End-to-End Discourse Parser. *Natural Language Engineering*, *20*, 151 - 184. Retrieved from http://arxiv.org/abs/1011.0835

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, *abs/1907.11692*. Retrieved from http://arxiv.org/abs/1907.11692

López Cortez, S. M., & Jacobs, C. L. (2023, July). The distribution of discourse relations within and across turns in spontaneous conversation. In *Proceedings of the 4th workshop on computational approaches to discourse (codi 2023)* (pp. 156–162). Toronto, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.codi-1.21

MacWhinney, B., Fromm, D., Forbes, M., & Holland, A. (2011). Aphasiabank: Methods for studying discourse. *Aphasiology*, *25*(11), 1286-1307. Retrieved from https://doi.org/10.1080/02687038.2011.589893 (PMID: 22923879)

Mann, W. C., & Taboada, M. (2005). *Rhetorical Structure Theory*. https://www.sfu.ca/rst/01intro/intro.html. (Accessed February 11, 2023)

Mann, W. C., & Thompson, S. A. (1988, January). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, *8*, 243-281.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics (ACL-2014): System Demonstrations* (pp. 55–60). Baltimore, Maryland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P14-5010 doi: 10.3115/v1/P14-5010

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330. Retrieved from https://aclanthology.org/J93-2004

McLaughlin, H. G. (1969, May). SMOG grading - a new readability formula. *Journal of Reading*, 639-646.

Mieskes, M., & Padó, U. (2018, November). Work smart - reducing effort in short-answer grading. In *Proceedings of the 7th workshop on NLP for computer assisted language learning* (pp. 57–68). Stockholm, Sweden: LiU Electronic Press. Retrieved from https://aclanthology.org/W18-7107

Montgomerie, A. (2021, March). *Attempting to Predict the CEFR Level of English Texts.* https://amontgomerie.github.io/2021/03/14/cefr-level-prediction.html. (Accessed December 2, 2022)

Naismith, B., Mulcaire, P., & Burstein, J. (2023, July). Automated evaluation of written discourse coherence using GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 394–403). Toronto, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.bea-1.32

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, *135*(3), 370–384. Retrieved 2023-07-20, from http://www.jstor.org/stable/2344614

O'Hayre, J. (1966). *Gobbledygook Has Gotta Go*. Washington, U.S. Dept. of the Interior, Bureau of Land Management. Retrieved from https://eric.ed.gov/?id=ED144073

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011, October). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Perkins, R. (2014). *Linguistic identifiers of L1 Persian speakers writing in English: NLID for authorship analysis* (Unpublished doctoral dissertation). Aston University, Birmingham, United Kingdom.

Pope, C., & Davis, B. H. (2011). Finding a balance: The Carolinas Conversation Collection. , *7*(1), 143–161. Retrieved 2023-02-13, from https://doi.org/10.1515/cllt.2011.007

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., & Webber, B. (2008, May). The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08).* Marrakech, Morocco: European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2008/pdf/754_paper.pdf

Prasad, R., Forbes-Riley, K., & Lee, A. (2017, August). Towards full text shallow discourse relation annotation: Experiments with cross-paragraph implicit relations in the pdtb. In *Sigdial conference.*

Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., & Webber, B. (2006, March). The Penn Discourse TreeBank 1.0 Annotation Manual.

Prasad, R., Webber, B., Lee, A., & Joshi, A. (2019). Penn Discourse Treebank Version 3.0. Philadelphia: Linguistic Data Consortium. Retrieved from https://hdl.handle.net/11272.1/AB2/SUU9CB doi: 11272.1/AB2/SUU9CB

Rimmer, W. (2006, October). Beyond the Sentence: Introducing Discourse Analysis Grammar. *ELT Journal*, *60*(4), 392-394. Retrieved from https://doi.org/10.1093/elt/ccl033 doi: 10.1093/elt/ccl033

Roos, Q., & Sidorova, J. (2022). *Fine-Tuning Pre-Trained Language Models for CEFR-Level and Keyword Conditioned Text Generation: A Comparison between Google's T5 and OpenAI's GPT-2* (Master's thesis, KTH Royal Institute of Technology). Retrieved from https://www.diva-portal.org/smash/get/diva2:1708538/FULLTEXT01.pdf

Rysová, K., Rysová, M., & Mírovský, J. (2016, October). Automatic evaluation of surface coherence in L2 texts in Czech. In *Proceedings of the 28th conference on computational linguistics and speech processing (ROCLING 2016)* (pp. 214–228). Tainan, Taiwan: The Association for Computational Linguistics and Chinese Language Processing (ACLCLP). Retrieved from https://aclanthology.org/O16-1021

Sanders, T., Demberg, V., Hoek, J., Scholman, M., Asr, F., Zufferey, S., & Evers-Vermeul, J. (2018, May). Unifying dimensions in coherence relations: How various annotation frameworks are related. *Corpus Linguistics and Linguistic Theory*, *17*. doi: 10.1515/cllt-2016-0078

Schmalz, V. J., & Brutti, A. (2021, December). Automatic Assessment of English CEFR Levels Using BERT Embeddings. In *Proceedings of 2021 Italian Conference on Computational Linguistics 2021 (CLiC-2021)* (Vol. 3033).

Schölkopf, B., & Smola, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.

Settles, B., LaFlair, G. T., & Hagiwara, M. (2020). Machine Learning Driven Language Assessment. *Transactions of the Association for Computational Linguistics*, *8*, 247–263. Retrieved from https://aclanthology.org/2020.tacl-1.17 doi: 10.1162/tacl_a_00310

Smith, E., & Senter, R. (1967). *Automated Readability Index*. Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command. Retrieved from https://books.google.ca/books?id=HejUGwAACAAJ (Accessed December 2, 2022)

Søgaard, A. (2022, December). Should We Ban English NLP for a Year? In *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 5254–5260). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.emnlp-main.351

Staples, S., & Dilger, B. (2018). *Corpus and repository of writing [Learner corpus articulated with repository]*. Retrieved from https://writecrow.org/

Texas Department of Insurance. (1992, June 15). *Public Claims Adjuster FAQ*. https://www.tdi.texas.gov/pubs/pc/pccpfaq.html.

Tyen, G., Brenchley, M., Caines, A., & Buttery, P. (2022, July). Towards an Open-Domain Chatbot for Language Practice. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)* (pp. 234–249). Seattle, Washington: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.bea-1.28 doi: 10.18653/v1/2022.bea-1.28

Vajjala, S., & Lõo, K. (2014, November). Automatic CEFR level prediction for Estonian learner text. In *Proceedings of the third workshop on NLP for computer-assisted language learning* (pp. 113–127). Uppsala, Sweden: LiU Electronic Press. Retrieved from https://aclanthology.org/W14-3509

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention Is All You Need. *CoRR*, *abs/1706.03762*. Retrieved from http://arxiv.org/abs/1706.03762

Vijay-Shanker, K. (1992). Using Descriptions of Trees in a Tree Adjoining Grammar. *Computational Linguistics*, *18*(4), 481–518. Retrieved from https://aclanthology.org/J92-4004

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018, November). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP* (pp. 353–355). Brussels, Belgium: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W18-5446 doi: 10.18653/v1/W18-5446

Wang, J., & Lan, M. (2015). A Refined End-to-End Discourse Parser. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning (CoNLL-2015)* (pp. 17–24). Beijing, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/K15-2002 doi: 10.18653/v1/K15-2002

Wang, Y., Li, S., & Wang, H. (2017, July). A Two-Stage Parsing Method for Text-Level Discourse Analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL-2017)* (Vol. 2, pp. 184–188). Vancouver, Canada. Retrieved from https://aclanthology.org/P17-2029 doi: 10.18653/v1/P17-2029

Webber, B. (2004). D-LTAG: Extending Lexicalized Tag to Discourse. *Cognitive Science*, *28*(5), 751–779. doi: 10.1207/s15516709cog2805_6

Webber, B. (2009, August). Genre distinctions for discourse in the Penn TreeBank. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics (ACL-2009) and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (AFNLP-2009)* (pp. 674–682). Suntec, Singapore. Retrieved from https://aclanthology.org/P09-1076

XTAG Research Group. (1998). A Lexicalized Tree Adjoining Grammar for English. *CoRR*, *cs.CL/9809024*. Retrieved from https://arxiv.org/abs/cs/9809024

Xue, N., Ng, H. T., Pradhan, S., Prasad, R., Bryant, C., & Rutherford, A. (2015). The CoNLL-2015 shared task on shallow discourse parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning (CoNLL-2015)* (pp. 1–16). Beijing, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/K15-2001 doi: 10.18653/v1/K15-2001

Yannakoudakis, H., Briscoe, T., & Medlock, B. (2011). A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 180–189). Portland, Oregon, USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P11-1019

Young, J. C., & Shishido, M. (2023). Investigating OpenAI's ChatGPT Potentials in Generating Chatbot's Dialogue for English as a Foreign Language Learning. *International Journal of Advanced Computer Science and Applications*, *14*(6). Retrieved from http://dx.doi.org/10.14569/IJACSA.2023.0140607 doi: 10.14569/IJACSA.2023.0140607

Zhao, Z., & Webber, B. (2022). *Revisiting Shallow Discourse Parsing in the PDTB-3: Handling Intra-sentential Implicits.* arXiv. Retrieved from https://arxiv.org/abs/2204.00350 doi: 10.48550/ARXIV.2204.00350

Appendix A

Detailed Results of Discourse Connectives

The following shows each PDTB level-2 relation, and the ratios of discourse connectives used to signal them, as well as the number of instances of discourse connectives for each CEFR level. Connectives with a p-value less than 0.05 in at least 2 of 3 t-tests (A2 vs C2, B1 vs C2, and B2 vs C2) are highlighted.

CONTINGENCY.Cause	**A2**	**B1**	**B2**	**C2**
so	0.424	0.379	0.242	0.269
because	0.428	0.390	0.289	0.431
so that	0.021	0.045	0.036	0.105
since	0.007	0.022	0.053	0.039
therefore	0.078	0.084	0.189	0.102
Total (Count)	**2371**	**7987**	**748**	**636**

Table A.1: Ratios of each discourse connective used to signal a CONTINGENCY.CAUSE relation in ICNALE.

COMPARISON.Contrast	A2	B1	B2	C2
but	0.744	0.655	0.411	0.538
however	0.148	0.185	0.372	0.191
on the other hand	0.023	0.039	0.024	0.020
nevertheless	0.004	0.008	0.005	0.002
while	0.029	0.040	0.100	0.077
though	0.026	0.046	0.051	0.122
Total (Count)	**1430**	**5313**	**760**	**493**

Table A.2: Ratios of each discourse connective used to signal a COMPARISON.CONTRAST relation in ICNALE.

Comparison.Concession	A2	B1	B2	C2
nonetheless	0.013	0.018	0.026	0.132
although	0.658	0.686	0.744	0.763
still	0.076	0.029	0.064	0.000
nevertheless	0.013	0.014	0.000	0.000
as if	0.051	0.026	0.013	0.000
though	0.165	0.221	0.154	0.053
but	0.013	0.004	0.000	0.000
in the end	0.013	0.000	0.000	0.000
Total (Count)	**79**	**510**	**78**	**38**

Table A.3: Ratios of each discourse connective used to signal a COMPARISON.CONCESSION relation in ICNALE.

Contingency.Condition	A2	B1	B2	C2
if	0.987	0.986	0.990	0.984
as long as	0.000	0.001	0.002	0.002
once	0.000	0.003	0.004	0.000
when	0.012	0.008	0.002	0.007
lest	0.000	0.000	0.002	0.000
when and if	0.000	0.000	0.000	0.005
Total (Count)	**1315**	**4100**	**486**	**426**

Table A.4: Ratios of each discourse connective used to signal a CONTINGENCY.CONDITION relation in ICNALE.

Expansion.Alternative	A2	B1	B2	C2
or	0.681	0.658	0.391	0.582
otherwise	0.071	0.092	0.072	0.036
until	0.000	0.005	0.000	0.000
unless	0.142	0.099	0.116	0.236
rather	0.000	0.007	0.014	0.018
instead	0.097	0.137	0.333	0.073
alternatively	0.009	0.000	0.072	0.055
Total (Count)	**113**	**437**	**69**	**55**

Table A.5: Ratios of each discourse connective used to signal a EXPANSION.ALTERNATIVE relation in ICNALE.

Temporal.Asynchronous	A2	B1	B2	C2
before	0.182	0.203	0.224	0.210
then	0.361	0.298	0.192	0.190
until	0.050	0.035	0.061	0.080
as soon as	0.007	0.009	0.012	0.035
ultimately	0.000	0.003	0.016	0.010
in turn	0.002	0.005	0.004	0.010
after	0.210	0.225	0.180	0.150
once	0.026	0.050	0.065	0.060
since	0.059	0.076	0.143	0.105
when	0.038	0.039	0.020	0.015
later	0.038	0.024	0.033	0.020
Total (Count)	**424**	**1812**	**245**	**200**

Table A.6: Ratios of each discourse connective used to signal a TEMPORAL.ASYNCHRONOUS relation in ICNALE.

Temporal.Synchrony	A2	B1	B2	C2
when	0.775	0.637	0.563	0.497
as	0.102	0.219	0.273	0.263
while	0.112	0.122	0.116	0.199
meanwhile	0.005	0.006	0.014	0.000
simultaneously	0.000	0.001	0.000	0.000
once	0.000	0.000	0.002	0.000
Total (Count)	**785**	**3447**	**490**	**376**

Table A.7: Ratios of each discourse connective used to signal a TEMPORAL.SYNCHRONY relation in ICNALE.

Expansion.Restatement	**A2**	**B1**	**B2**	**C2**
overall	0.040	0.072	0.121	0.400
much as	0.040	0.007	0.030	0.000
in particular	0.160	0.029	0.121	0.000
in other words	0.480	0.338	0.364	0.100
in the end	0.000	0.022	0.000	0.100
indeed	0.000	0.007	0.061	0.000
in short	0.160	0.323	0.061	0.300
in sum	0.040	0.050	0.061	0.000
rather	0.000	0.022	0.091	0.100
Total (Count)	**25**	**139**	**33**	**10**

Table A.8: Ratios of each discourse connective used to signal a EXPANSION.RESTATEMENT relation in ICNALE.

Expansion.Instantiation	**A2**	**B1**	**B2**	**C2**
for example	0.829	0.834	0.880	0.900
for instance	0.167	0.164	0.120	0.100
in particular	0.005	0.003	0.000	0.000
Total (Count)	**210**	**652**	**100**	**40**

Table A.9: Ratios of each discourse connective used to signal a EXPANSION.INSTANTIATION relation in ICNALE.

Expansion.Conjunction	**A2**	**B1**	**B2**	**C2**
and	0.578	0.501	0.406	0.622
in addition	0.034	0.023	0.047	0.009
moreover	0.023	0.032	0.040	0.001
in fact	0.017	0.016	0.014	0.005
finally	0.025	0.019	0.015	0.021
also	0.283	0.342	0.371	0.228
besides	0.016	0.022	0.030	0.002
Total (Count)	**2463**	**10092**	**1214**	**1304**

Table A.10: Ratios of each discourse connective used to signal a EXPANSION.CONJUNCTION relation in ICNALE.

Contingency.Cause	**A2**	**B1**	**B2**	**C2**
so	0.389	0.216	0.209	0.173
because	0.245	0.372	0.388	0.578
so that	0.048	0.057	0.040	0.033
since	0.049	0.046	0.072	0.067
therefore	0.117	0.133	0.134	0.074
Total (Count)	**1464**	**1274**	**4789**	**973**

Table A.11: Ratios of each discourse connective used to signal a CONTINGENCY.CAUSE relation in CROW.

Comparison.Contrast	A2	B1	B2	C2
but	0.397	0.343	0.366	0.356
however	0.338	0.411	0.366	0.273
on the other hand	0.038	0.039	0.045	0.009
nevertheless	0.012	0.019	0.019	0.003
while	0.075	0.055	0.062	0.112
though	0.063	0.061	0.072	0.174
Total (Count)	**1286**	**1006**	**4256**	**1199**

Table A.12: Ratios of each discourse connective used to signal a COMPARISON.CONTRAST relation in CROW.

Comparison.Concession	A2	B1	B2	C2
nonetheless	0.128	0.055	0.050	0.006
although	0.705	0.843	0.798	0.710
still	0.040	0.031	0.029	0.024
nevertheless	0.013	0.008	0.005	0.000
as if	0.047	0.000	0.010	0.083
though	0.054	0.063	0.104	0.160
in the end	0.007	0.000	0.000	0.000
Total (Count)	**149**	**127**	**584**	**169**

Table A.13: Ratios of each discourse connective used to signal a COMPARISON.CONCESSION relation in CROW.

Contingency.Condition	A2	B1	B2	C2
if	0.979	0.979	0.971	0.979
as long as	0.006	0.002	0.003	0.000
once	0.008	0.006	0.013	0.006
when	0.006	0.011	0.013	0.011
Total (Count)	**517**	**466**	**1570**	**621**

Table A.14: Ratios of each discourse connective used to signal a CONTINGENCY.CONDITION relation in CROW.

Expansion.Alternative	A2	B1	B2	C2
or	0.489	0.536	0.473	0.475
otherwise	0.098	0.179	0.083	0.049
until	0.000	0.000	0.000	0.008
unless	0.163	0.071	0.087	0.115
rather	0.022	0.000	0.033	0.057
instead	0.217	0.214	0.324	0.287
alternatively	0.011	0.000	0.000	0.008
Total (Count)	**92**	**56**	**241**	**122**

Table A.15: Ratios of each discourse connective used to signal a EXPANSION.ALTERNATIVE relation in CROW.

Temporal.Asynchronous	**A2**	**B1**	**B2**	**C2**
before	0.100	0.174	0.154	0.154
then	0.281	0.223	0.196	0.280
until	0.049	0.032	0.024	0.069
as soon as	0.014	0.002	0.005	0.010
ultimately	0.023	0.002	0.007	0.003
in turn	0.012	0.016	0.010	0.039
after	0.260	0.298	0.306	0.195
once	0.037	0.068	0.066	0.066
since	0.128	0.120	0.161	0.071
when	0.044	0.018	0.027	0.049
later	0.033	0.023	0.022	0.046
Total (Count)	**430**	**443**	**1476**	**590**

Table A.16: Ratios of each discourse connective used to signal a TEMPORAL.ASYNCHRONOUS relation in CROW.

Temporal.Synchrony	**A2**	**B1**	**B2**	**C2**
when	0.588	0.660	0.543	0.649
as	0.302	0.218	0.313	0.216
while	0.072	0.076	0.096	0.113
meanwhile	0.017	0.027	0.020	0.001
simultaneously	0.001	0.000	0.001	0.002
Total (Count)	**709**	**633**	**2653**	**924**

Table A.17: Ratios of each discourse connective used to signal a TEMPORAL.SYNCHRONY relation in CROW.

Expansion.Restatement	**A2**	**B1**	**B2**	**C2**
overall	0.319	0.156	0.136	0.400
much as	0.000	0.000	0.023	0.000
in particular	0.043	0.022	0.023	0.000
in other words	0.426	0.422	0.506	0.286
in the end	0.021	0.000	0.000	0.000
indeed	0.000	0.000	0.011	0.000
in short	0.043	0.267	0.125	0.086
in sum	0.000	0.044	0.017	0.029
rather	0.021	0.000	0.017	0.086
Total (Count)	**47**	**45**	**176**	**35**

Table A.18: Ratios of each discourse connective used to signal a EXPANSION.RESTATEMENT relation in CROW.

Expansion.Instantiation	**A2**	**B1**	**B2**	**C2**
for example	0.809	0.725	0.768	0.723
for instance	0.186	0.275	0.224	0.277
in particular	0.005	0.000	0.008	0.000
Total (Count)	**215**	**167**	**655**	**65**

Table A.19: Ratios of each discourse connective used to signal a EXPANSION.INSTANTIATION relation in CROW.

Expansion.Conjunction	**A2**	**B1**	**B2**	**C2**
and	0.494	0.451	0.456	0.556
in addition	0.030	0.032	0.030	0.012
moreover	0.036	0.040	0.039	0.003
in fact	0.022	0.016	0.018	0.019
finally	0.008	0.013	0.013	0.004
also	0.331	0.367	0.359	0.327
besides	0.013	0.021	0.017	0.002
Total (Count)	**2475**	**1932**	**7352**	**1857**

Table A.20: Ratios of each discourse connective used to signal a EXPANSION.CONJUNCTION relation in CROW.

Appendix B

Native Language Analysis

Each essay in each of the datasets was marked with its CEFR score, but it was also marked with the L1[1] language or country of origin of the writer. Because of this, we were able to extract discourse relation frequency regarding native language or country of origin. Though we had not gone into as deep a dive into the machine learning aspects and statistical significance tests for this secondary question, the data could be used in future work on discourse differences among ELL writers from various L1 backgrounds.

	Ela.	Att.	Joi.	M-M	Ena.	Bac.	Com.	Cont.	Temp.	Cond.	Cau.	Exp.
China	62.89%	6.31%	9.51%	0.66%	2.12%	2.28%	0.34%	2.96%	0.49%	1.25%	1.08%	3.89%
India	68.95%	5.75%	8.35%	0.85%	1.69%	1.71%	0.32%	2.38%	0.39%	0.85%	0.79%	2.92%
Korea	64.39%	5.70%	8.44%	0.81%	1.92%	2.32%	0.33%	3.19%	0.38%	1.18%	1.12%	4.06%
Malaysia	67.06%	5.98%	6.14%	0.82%	1.94%	2.58%	0.64%	2.76%	0.38%	1.10%	0.64%	3.30%
Taiwan	62.57%	5.93%	9.50%	0.69%	1.79%	2.50%	0.36%	2.50%	0.43%	1.07%	1.50%	4.14%
USA	63.97%	5.70%	11.22%	0.93%	1.62%	1.95%	0.21%	2.58%	0.43%	1.26%	0.86%	2.94%

Table B.1: Percentage analysis of 12 RST relations, in texts from the 6 most represented countries of origin in the CROW dataset, using parser info from the Heilman and Sagae (2015) parser.

Data for both RST parsers in the ICNALE (Ishikawa, 2013) dataset comes from the entirety of the output of each parser, prior to filtering for only spans which agree on the relation (see Section 3.2.2).

The International Corpus of Learner English (ICLE) (Granger et al., 2020), as discussed in Section 3.1, did not contain enough CEFR-labelled data to be used for the main task of the thesis.

[1]"L1" in ICNALE consists of essays from six English-speaking countries: United States, Canada, United Kingdom, Australia, New Zealand, and Nigeria

	Elab.	Att.	Joi.	Ena.	Bac.	Comp.	Cont.	M-M	Tem.	Cond.	Cau.	Exp.
China	52.51%	7.80%	12.21%	3.50%	4.01%	0.40%	4.67%	0.66%	0.52%	3.23%	1.39%	4.32%
Hong Kong	56.36%	7.49%	10.86%	3.55%	4.05%	0.58%	3.72%	0.92%	0.50%	3.38%	1.56%	3.85%
Indonesia	52.70%	6.04%	12.46%	2.56%	2.38%	0.34%	4.03%	0.81%	0.70%	3.63%	2.58%	7.32%
Japan	50.72%	10.09%	11.51%	2.99%	2.90%	0.24%	4.79%	1.03%	0.52%	4.16%	1.45%	5.73%
Korea	52.57%	7.72%	12.44%	2.56%	2.55%	0.32%	4.67%	0.82%	0.47%	3.85%	1.85%	5.96%
Pakistan	59.69%	4.13%	15.20%	2.86%	1.51%	0.28%	3.71%	0.81%	0.63%	1.67%	2.07%	5.06%
Phillippines	57.14%	6.72%	11.54%	1.97%	2.40%	0.28%	4.20%	0.75%	0.71%	3.06%	1.75%	5.17%
Singapore	54.87%	7.31%	12.52%	2.90%	3.37%	0.65%	4.12%	1.27%	0.70%	3.12%	1.51%	3.29%
Taiwan	51.12%	9.10%	12.76%	3.09%	3.10%	0.21%	4.85%	0.65%	0.62%	3.71%	1.34%	4.43%
Thailand	52.05%	7.42%	14.87%	2.64%	2.31%	0.21%	3.57%	0.52%	0.56%	3.98%	1.86%	6.19%
L1	48.16%	11.10%	16.30%	2.22%	2.42%	0.41%	3.71%	0.93%	0.79%	3.53%	1.48%	4.03%

Table B.2: Percentage analysis of 12 RST relations, in texts from all countries of origin in the ICNALE dataset, using parser info from the Heilman and Sagae (2015) parser.

	Elab.	Att.	Joi.	Ena.	Bac.	Comp.	Cont.	M-M	Tem.	Cond.	Cau.	Exp.
China	39.94%	17.09%	14.69%	5.38%	8.19%	0.32%	7.26%	1.04%	2.26%	7.01%	1.46%	3.14%
Hong Kong	40.66%	16.42%	14.27%	5.40%	7.93%	0.35%	5.81%	1.22%	3.05%	6.90%	2.28%	3.38%
Indonesia	39.00%	14.54%	14.78%	5.45%	4.84%	0.17%	7.17%	1.22%	1.87%	8.57%	1.37%	3.18%
Japan	34.77%	22.41%	13.25%	5.63%	5.42%	0.23%	8.09%	1.05%	2.20%	9.63%	0.81%	3.25%
Korea	36.23%	18.64%	14.17%	5.59%	5.53%	0.20%	7.66%	1.07%	2.09%	9.78%	1.56%	3.09%
Pakistan	46.06%	13.15%	17.70%	5.40%	5.01%	0.19%	6.87%	0.89%	1.21%	4.58%	1.21%	3.18%
Phillippines	42.81%	15.37%	13.31%	5.97%	4.70%	0.31%	7.80%	0.93%	2.09%	6.55%	2.01%	3.06%
Singapore	43.38%	15.68%	15.05%	5.45%	7.08%	0.32%	7.01%	0.89%	2.50%	5.60%	1.89%	3.11%
Taiwan	36.13%	20.33%	14.15%	5.38%	6.85%	0.22%	8.15%	1.03%	2.14%	7.82%	1.55%	3.04%
Thailand	36.76%	17.57%	15.80%	5.81%	4.44%	0.32%	5.97%	0.96%	2.03%	9.31%	1.14%	3.13%
L1	35.36%	23.35%	17.39%	5.24%	4.73%	0.24%	6.15%	0.90%	2.06%	7.39%	0.82%	3.15%

Table B.3: Percentage analysis of 12 RST relations, in texts from all countries of origin in the ICNALE dataset, using parser info from the Y. Wang et al. (2017) parser.

However, this data consisted of written texts from 27 different language backgrounds. Thus, this data could be used for analysis based on native language. Texts are sorted by the label "first language at home" from the ICLE dataset.[2]. This data is shown in Table B.4.

This data was not verified to the extent that the CEFR-labelled data was, given that the ICLE data was only run on one parser, however based on this data, we can make a few observations that could be further explored in future research. For these observations, we are comparing ICNALE's "Country" label with ICLE's "first language at home" label, specifically China, Taiwan, and Hong Kong with Chinese, Japan with Japanese, Republic of Korea with Korean, and Pakistan with Urdu and Punjabi. While it is not always the case that speakers of these languages are from these countries and vice versa, the data can give a general idea of trends in discourse usage.

- Pakistan is well ahead of all other listed countries in the use of Elaboration with both parsers

[2]The full manual can be downloaded from https://dial.uclouvain.be/pr/boreal/object/boreal:229877

	Elab.	Att.	Joi.	Ena.	Back.	Comp.	Cont.	Temp.	Cond.	Cau.	Exp.
Bulgarian	54.90%	6.56%	13.72%	2.14%	1.94%	0.33%	4.52%	0.43%	1.95%	1.38%	5.33%
Chinese-Cantonese	56.29%	9.79%	9.94%	2.96%	2.87%	0.34%	3.33%	0.50%	2.72%	1.50%	4.81%
Chinese-Mandarin	53.68%	7.00%	13.13%	2.65%	3.38%	0.26%	4.60%	0.57%	2.46%	1.22%	5.65%
Czech	55.99%	7.35%	14.81%	1.70%	1.92%	0.39%	4.17%	0.41%	1.79%	1.35%	5.57%
Dutch	54.81%	6.88%	12.44%	2.02%	2.22%	0.40%	4.92%	0.44%	2.05%	1.50%	6.10%
Finnish	55.22%	7.00%	12.94%	1.88%	2.27%	0.46%	4.94%	0.42%	2.55%	1.33%	5.65%
French	56.81%	7.10%	11.03%	2.65%	2.16%	0.31%	4.36%	0.36%	2.22%	1.30%	4.83%
German	55.38%	6.09%	13.69%	1.86%	2.76%	0.34%	4.73%	0.63%	2.17%	1.13%	4.79%
Greek	54.83%	6.50%	14.14%	2.67%	1.93%	0.35%	3.94%	0.30%	1.47%	1.40%	5.66%
Hungarian	57.70%	6.23%	9.34%	2.38%	2.87%	0.43%	4.29%	0.36%	2.45%	1.77%	5.60%
Italian	54.62%	7.39%	11.30%	1.71%	1.86%	0.30%	5.64%	0.23%	2.03%	1.23%	5.22%
Japanese	49.50%	9.06%	14.11%	2.66%	2.99%	0.25%	4.63%	0.47%	3.11%	1.37%	6.63%
Korean	52.60%	6.95%	12.57%	2.48%	2.90%	0.34%	4.91%	0.43%	2.83%	1.81%	6.64%
Lithuanian	54.83%	7.24%	12.34%	2.32%	2.37%	0.42%	4.54%	0.46%	1.72%	1.38%	5.69%
Macedonian	52.50%	6.42%	15.61%	2.29%	2.40%	0.33%	4.43%	0.38%	1.97%	1.44%	5.32%
Norwegian	51.86%	7.69%	15.54%	2.10%	2.27%	0.52%	4.66%	0.55%	2.15%	1.10%	5.41%
Persian	53.10%	6.63%	15.58%	2.55%	2.10%	0.21%	3.90%	0.61%	1.92%	1.63%	5.38%
Portuguese	55.25%	6.45%	12.96%	2.45%	2.26%	0.35%	4.05%	0.37%	2.00%	1.43%	5.13%
Polish	59.29%	5.77%	11.68%	2.12%	2.25%	0.45%	3.54%	0.37%	1.60%	1.23%	5.24%
Punjabi	59.47%	5.14%	13.72%	1.82%	1.79%	0.40%	3.51%	0.44%	1.86%	1.29%	5.90%
Russian	54.36%	7.43%	13.49%	2.22%	1.97%	0.34%	4.70%	0.37%	1.75%	1.46%	5.31%
Serbian	51.83%	6.12%	15.41%	1.90%	2.72%	0.52%	5.19%	0.64%	2.42%	1.37%	5.19%
Spanish	54.88%	6.89%	11.32%	1.93%	2.23%	0.31%	4.97%	0.41%	2.01%	1.52%	5.24%
Swedish	53.28%	7.63%	13.61%	2.14%	2.62%	0.50%	4.73%	0.42%	2.38%	1.06%	5.38%
Tswana	56.34%	7.37%	12.59%	2.40%	1.99%	0.25%	2.05%	0.44%	2.42%	2.42%	6.69%
Turkish	52.77%	6.94%	12.78%	2.44%	2.74%	0.35%	3.99%	0.61%	2.75%	1.66%	7.51%
Urdu	60.58%	4.72%	13.37%	2.23%	1.63%	0.40%	3.02%	0.55%	1.77%	1.77%	6.08%

Table B.4: Frequency analysis of 12 RST relations, in texts from all 27 ”first language at home” labels in the ICLE dataset, using parser info from the Heilman and Sagae (2015) parser.

for ICNALE, while Urdu and Punjabi are 1st and 2nd highest in this category for ICLE.

- Japan/Japanese always scores high in Attribution relations, trailing only L1 learners in the ICNALE dataset and Cantonese in the ICLE dataset.

- Hong Kong and China are 1st and 2nd highest in the use of Background with both parsers for ICNALE, while Chinese-Mandarin and Chinese-Cantonese are 1st and 2nd highest in this category for ICLE.

- Conversely, Pakistan has the lowest score for the relation of BACKGROUND for one parser in the ICNALE dataset (and the 4th-lowest in the other), while Urdu and Punjabi are 1st- and 2nd-lowest in this category for ICLE.

Appendix C

RST-DT Labels

The following are examples of the 12 different RST relations that are discussed in this thesis. The list of all 16 RST-DT relation classes are located in the Discourse Tagging Reference Manual (Carlson & Marcu, 2001)[1]. This manual was used as a guide to create these example sentences. All explanations of these relations come from this guide as well.

Note that the manual gives a total of 78 different relations, and the 16 RST-DT relations are the 16 classes they are sorted into. Thus, this list is not exhaustive, which is the reason, for example, some CONTRAST relations are a nucleus-satellite pair, despite the provided example being multi-nucleic.

In the following examples, EDUs are listed in [brackets], while nuclei are in *italics*.

1. ATTRIBUTION Instances of reported speech, including both direct and indirect forms, necessitate the identification of the rhetorical relation called ATTRIBUTION. The satellite component of this relation comprises the source of the attribution, which can be a clause containing a reporting verb or a phrase introduced by *according to*. On the other hand, the nucleus represents the actual content of the reported message, which is typically presented in a separate clause. Note that the ATTRIBUTION relation is not limited to speech alone but can also be employed with cognitive predicates, encompassing emotions, thoughts, hopes, and similar expressions.

[1] https://www.isi.edu/~marcu/discourse/tagging-ref-manual.pdf
The full list of classes is located on page 32, examples of each relation begin on page 45.

(11) [*According to Toad,*] [the princess is in another castle.]

2. BACKGROUND In a BACKGROUND relation, the satellite component serves to establish the context or basis for interpreting the nucleus. Comprehending the satellite element aids the reader in understanding the nucleus. It is important to note that the satellite does not represent the cause, reason, or motivation behind the situation presented in the nucleus. Furthermore, the intentions of the reader or writer do not play a role in determining the existence of this relation.

(12) [Mario wears overalls.] [*Overalls are trousers with an extra piece of cloth covering the chest, held in place by a strap over each shoulder.*]

3. CAUSE The situation portrayed in the nucleus gives rise to the situation depicted in the satellite. The nucleus, acting as the cause, holds paramount importance. Meanwhile, the satellite illustrates the outcome or consequence of the action. The writer's intention is to underscore and highlight the cause.

(13) [This year, Princess Peach enacted a new charter,] [*expanding the Mushroom Kingdom beyond its original borders.*]

4. COMPARISON In a COMPARISON relation, two textual spans are examined and evaluated based on a specific dimension, which can be abstract in nature. Such relationships can convey similarities, differences, greater than, less than, and other comparative aspects concerning abstract entities associated with the comparison relation. It is important to note that in the context of a comparison relation, the spans, entities, and other elements being compared are not presented contrastingly.

(14) [The Mushroom Kingdom ball expects 1,000 Mushroom People as guests,] [*compared with 900 guests in the previous year.*]

5. CONDITION In a CONDITION relation, the validity of the proposition connected to the nucleus depends on the fulfillment of the condition stated in the satellite. The satellite presents a hypothetical situation that has not actually occurred. This relation can also encompass negative conditions, where the non-fulfillment of a particular condition leads to a specific outcome.

(15) [Mario is going to win this race,] [*unless Yoshi or Birdo have a blue shell.*]

6. CONTRAST In a CONTRAST relation, two or more nuclei are juxtaposed and contrasted with each other based on a specific dimension. This contrast may occur in only a few respects, while other aspects remain unchanged. Typically, a CONTRAST relation involves the use of a contrastive discourse cue (referred to as a discourse connective in PDTB), such as *but, however, while*, whereas a COMPARISON relation does not require such cues.

(16) [But from the beginning of the game, Yoshi's star count grew,] [while Luigi's never did.]

7. ELABORATION "The satellite provides specific information to help define a very general concept introduced in the nucleus."

(17) [Bowser's army is big.] [*It consists of 1 million Goombas.*]

8. ENABLEMENT In an ENABLEMENT relation, the scenario described in the nucleus is not yet actualized. The action presented in the satellite, however, enhances the likelihood of the situation described in the nucleus becoming a reality.

(18) [*Mario should jump over the barrels.*] [Doing this will increase the likelihood he is able to rescue the Princess from Donkey Kong.]

9. EXPLANATION The EXPLANATION relation "provides a factual explanation for the situation presented in the nucleus." Relations that fall within the category of EXPLANATION may or may not be attempting to convince the reader of a point.

(19) [Ultimately, Yoshi is doing well in this game.] [*He's been able to collect a star every few turns.*]

10. JOINT There are two categories of JOINT relations: a list and a disjunction. A list is a multinuclear relation where the elements are presented sequentially without being compared, contrasted, or engaged in a stronger form of multinuclear relation. Typically, a list relation demonstrates a parallel structure among the units involved. On the other hand, a disjunction is a multinuclear relation where the elements are listed as alternative options, either positive or negative in nature.

(20) [Mario, sooner or later, has to stomp a Goomba,] [and he has to get a power-up.]

11. MANNER-MEANS The MANNER-MEANS relation can be further categorized into a *manner* relation and a *means* relation. In a *manner* relation, the satellite elucidates the manner or way in which something is done. It may also involve expressing similarity or making comparisons. The satellite answers questions such as ”in what manner?” or ”in what way?”

On the other hand, in a *means* relation, the satellite specifies the method, mechanism, instrument, channel, or conduit used to achieve a particular goal. It provides information on how something was or is to be accomplished. In other words, the satellite answers questions like ”by which means?” or ”how?” and is often indicated by the preposition *by*.

(21) [Birdo could still win this race] [*by using the shortcut on Rainbow Road.*]

12. TEMPORAL In a TEMPORAL relation, the situation presented in the nucleus occurs before, at the same time, or after the situation presented in the satellite. As the name suggests, the two spans are related by time.

(22) [*Following its massive success in North America,*] [the arcade game was ported to the Game & Watch in 1982.]

Appendix D

Tools

This work would not be possible without the tools developed over the years and open sourced by their creators. All the work to run the RST parsers, preprocess the text, index and rank the documents was run inside the GNU/Linux[1] environment, or Windows Subsystem for Linux [2].

Python[3], Pytorch[4], Scikit-Learn[5] and Transformers[6] from Huggingface[7] provided us with the libraries required to implement RoBERTa.

SpaCy linguistic tools `https://spacy.io/usage/linguistic-features` were used to collect linguistic features such as part-of-speech tags and syntactic parse tree depth for use in machine learning classification. This classification also relied on the Joblib[8], Scikit-Learn, pandas[9], and NumPy[10] libraries.

Additional support for running the RST and PDTB parsers was provided by CoreNLP[11], which performed useful syntactic tasks such as tokenization and lemmatization.

[1] https://www.gnu.org/gnu/linux-and-gnu.en.html
[2] https://learn.microsoft.com/en-us/windows/wsl/install
[3] https://www.python.org/
[4] https://pytorch.org/
[5] https://scikit-learn.org
[6] https://huggingface.co/docs/transformers/index
[7] https://huggingface.co/
[8] https://joblib.readthedocs.io/en/stable/
[9] https://pandas.pydata.org/
[10] https://numpy.org/
[11] https://stanfordnlp.github.io/CoreNLP/

The Stanford[12] and Berkeley[13] parsers, as well as implementation code from University of Trento[14] were used to convert raw text to the format necessary to run the J. Wang and Lan (2015) parser.

ZPar Parser[15] was used to generate constituency parsing for use by the Heilman and Sagae (2015) parser, while the CRF was trained with the aid of binaries from CRFPP[16].

CRFSuite[17] was used to aid in EDU segmentation for the Y. Wang et al. (2017) parser.

Many LateX tables in this thesis were generated using Tables Generator[18].

Several figures in this thesis were created with the help of Photopea[19].

Additional support provided by the work of Richard A[20].

[12] https://nlp.stanford.edu/software/lex-parser.shtml

[13] https://github.com/slavpetrov/berkeleyparser

[14] https://github.com/esrel/DP

[15] https://github.com/frcchang/zpar

[16] https://github.com/taku910/crfpp

[17] http://www.chokkan.org/software/crfsuite/

[18] https://www.tablesgenerator.com/

[19] https://www.photopea.com/

[20] https://www.youtube.com/watch?v=dQw4w9WgXcQ