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

Blaise Hanel and **Leila Kosseim**
Computational Linguistics at Concordia (CLaC) Laboratory
Department of Computer Science and Software Engineering
Concordia University, Montréal, Québec, Canada
blaise.hanel@concordia.ca, leila.kosseim@concordia.ca

Abstract

In this paper, we investigate the relationship between the use of discourse relations and the CEFR-level of argumentative English learner essays. Using both the Rhetorical Structure Theory (RST) and the Penn Discourse TreeBank (PDTB) frameworks, we analyze essays from The International Corpus Network of Asian Learners (ICNALE), and the Corpus and Repository of Writing (CROW). Results show that the use of the RST relations of EXPLANATION and BACKGROUND, as well as the first-level PDTB sense of CONTINGENCY, are influenced by the English proficiency level of the writer.

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 surface linguistic features and language learning, such as Webber (2009), Bachand et al. (2014), and Abdalla et al. (2018) have shown significant difference in discourse usage across textual genre and simplicity level. To our knowledge, very few studies have investigated the relationship between discourse structures and language learning.

Corpus research on the use of discourse structures among different CEFR levels can provide valuable insights into how well language learners are able to organize and convey their ideas in written or spoken language. Such an analysis can also identify common patterns of language use that are particularly challenging for learners at different CEFR levels, leading to the development of more effective teaching materials and strategies that target learners' specific needs (Aoyama, 2022), while simultaneously reducing the workload of human graders (Mieskes and 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.

In this paper, we investigate the usage of discourse relations using the Rhetorical Structure Theory (Mann and Thompson, 1988) and the Penn Discourse TreeBank (Prasad et al., 2008) frameworks to discover trends in their usage in argumentative English learners across various proficiency levels. Results show that the RST relations of EXPLANATION and BACKGROUND are statistically used more often by writers with a lower CEFR language level, and the use of the PDTB relation of CONTINGENCY decreases as CEFR level increases.

2 Background

2.1 Discourse Analysis Frameworks

In order to analyze the discourse structure of texts computationally, two main frameworks have been developed: Rhetorical Structure Theory (RST), proposed by Mann and Thompson (1988) and Discourse Lexicalized Tree-Adjoining Grammar (Webber and Joshi, 1998), the basis for the Penn Discourse TreeBank (PDTB) (Prasad et al., 2008).

RST describes a text in terms of a tree structure, where leafs are textual units, known as *Elementary Discourse Units* (EDUs). EDUs are the minimal unit of discourse, and are linked to one another to form nodes corresponding to contiguous text spans. The tree describes how each node is related to another via a discourse rela-

Proceedings of Recent Advances in Natural Language Processing, pages 468–474
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_052

tion. Several RST parsers have been developed (e.g. (Heilman and Sagae, 2015) and (Wang et al., 2017)) using the annotated RST-DT dataset (Carlson et al., 2001). The RST-DT uses an inventory of 78 relations organized into 16 major relation groups, namely ATTRIBUTION, BACKGROUND, CAUSE, COMPARISON, CONDITION, CONTRAST, ELABORATION, EVALUATION, ENABLEMENT, EXPLANATION, JOINT, MANNER-MEANS, SUMMARY, TOPIC-COMMENT, TOPIC-CHANGE, and TEMPORAL.

The other main discourse framework is the Penn Discourse TreeBank (PDTB). Three versions of the PDTB have been developed: PDTB-1.0 (Prasad et al., 2006), PDTB-2.0 (Prasad et al., 2008), and PDTB-3.0 (Prasad et al., 2019). We used the PDTB-2.0, as most work has been done with this version and several freely available parsers have been developed (e.g. (Lin et al., 2014; Wang and Lan, 2015)). Unlike RST, the PDTB-2.0 organizes discourse relations (called *senses*) into a 3-tier hierarchy. Four top-level discourse relations (CONTINGENCY[1], EXPANSION, COMPARISON, and TEMPORAL) are further split into second-level and third-level relations.

An important difference between the RST and PDTB frameworks is that RST segments are non-overlapping and cover the entire text as a tree-structure, with every pair of segments assigned an RST relation. On the other hand, PDTB parsing forms a flat structure that links adjacent texts segments (called *arguments*) which may contain segments that overlap. Though the frameworks differ in their structure and inventory of relations, works such as (Demberg et al., 2017) have provided guidelines to compare them.

2.2 Language Proficiency Levels

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) proposed a metric[2] for mapping TOEFL scores directly to CEFR levels.

3 Previous Work

3.1 Discourse Features Across Texts

Differences of 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) evaluated the usefulness of both RST and PDTB relations as features to measure text complexity, and explore 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*).

Abdalla et al. (2018) identified changes in the usage of discourse relations among patients with Alzheimer's disease. They used the RST parser

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

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

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

Table 1: Statistics of the ICNALE and CROW datasets. A2-B2 essays are all from English learners, while C2 essays are from countries with English as an official language.

of Feng and Hirst (2014) to analyze written material by patients with Alzheimer's, from the DementiaBank (MacWhinney et al., 2011) and CCC (Pope and 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. To our knowledge, our work is the first to analyze differences in discourse structures across language proficiency levels.

4 Datasets

In order to analyze discourse structures across CEFR levels, we aimed for texts long enough to have rich discourse structures. We used two datasets of argumentative essays: ICNALE (Ishikawa, 2013) and CROW (Staples and Dilger, 2018). We did not use the datasets of Schmalz and Brutti (December 2021) (see Section 3) as these largely consist of short 2-3 sentence texts.

The first dataset we used was the International Corpus Network of Asian Learners (ICNALE) (Ishikawa, 2013). The ICNALE dataset used the ETS mapping (see Section 2.2) 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.

The second dataset we used was the Corpus and Repository of Writing (CROW) (Staples and Dilger, 2018). For the sake of consistency in genre, we only used the argumentative papers from this dataset for comparison with the ICNALE dataset. The CROW dataset is not labelled with CEFR scores, but rather with TOEFL scores. For comparative purposes, we used the ETS mapping on the CROW dataset to determine the CEFR score.

Table 1 shows statistics of both datasets. 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 1, the datasets do not contain samples of A1 and C1 CEFR levels, and are not balanced across levels.

5 Discourse Analysis

In order to extract reliable discourse information from the datasets, we used two publicly-available discourse parsers from each framework. For RST, we used the Wang et al. (2017) and the Heilman and Sagae (2015) parsers. We chose these parsers because they use the same set of RST relations, and they achieve high performance for relation tagging. Heilman and Sagae (2015) achieves an F-score of 57.4% on the RST-DT test set, Wang et al. (2017) achieves 59.7%, while human performance is 65.8% (Wang et al., 2017). For PDTB parsing, we used the (Lin et al., 2014) and the (Wang and Lan, 2015) parsers due to their high performance and availability.

We parsed the ICNALE and the CROW datasets (see Section 4) with all four parsers using all 16 RST relations and the 4 level-1 PDTB relations. The outputs of both RST parsers and both PDTB parsers were then compared. In order to have significant statistics, we ignored any discourse relation that appeared in less than 10% of the documents. These included the RST relations of EVALUATION, SUMMARY, TOPIC-COMMENT and TOPIC-CHANGE. This left us with the 12 most frequent relations: ATTRIBUTION, BACKGROUND, CONTRAST, CAUSE, COMPARISON, CONDITION, ELABORATION, ENABLEMENT, EXPLANATION, JOINT, MANNER-MEANS, and TEMPORAL. All PDTB level 1 relations appeared in more than 10% of the documents, hence all were considered.

We computed the average frequency of each RST and level 1 PDTB relation for each CEFR label in the dataset: A2, B1, B2, and C2. To determine if there was a statistical difference in the usage of these relations across CEFR levels, we ran a two-tailed t-test with a p-value of 0.95, comparing A2 against C2, B1 against C2, and B2 against C2.

5.1 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, EDU_{p1} and EDU_{p2}:

Segment Alignment:

If EDU_{p1} and EDU_{p2} span the same text (sans punctuation), we align them and keep the pair (EDU_{p1}, EDU_{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 $\text{EDU}_{\text{p}i}$ in the aligned (EDU_{p1}, EDU_{p2}),
    - If $\text{EDU}_{\text{p}i}$ was labelled as a satellite by parser pi, or as the second half of a multi-nucleic relation, it is then labelled with its lowest-level discourse relation (see EDUs A and C in Figure 1).
    - Otherwise, if $\text{EDU}_{\text{p}i}$ was labelled as a nucleus by parser pi, it is not assigned a relation.

   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.
    - 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.

Using this method, we were able to verify the agreement between the two parsers on the 11 satellite-nucleus RST relations. The RST relation of JOINT is multi-nucleic, and not considered in the approach. The two parsers on the ICNALE dataset showed an agreement of 80.10% on relation tags, with the full results shown in Table 2.

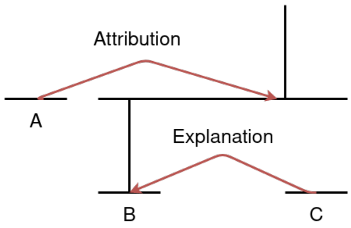

Figure 1: Example RST tree. In our method, satellite A would be labelled ATTRIBUTION, while satellite C would be labelled ELABORATION. Satellite B, as a nucleus, would not receive a label.

As the results show, the parsers disagreed most frequently on CAUSE relations, frequently mislabelling these relations as EXPLANATION. ENABLEMENT relations were also frequently mislabelled as ELABORATION by both parsers. For the following analysis, only the EDUs with an agreed-upon relation between the two parsers were used.

5.2 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. 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, shown in Table 3, suggests a general downward trend in the usage of EXPLANATION relations, which flattens out as the learner reaches the B2 level. 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, or to explain concepts they can not recall the terms for.

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. Table 3 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. The RST relation of BACKGROUND is used to provide information that is important to understanding the main idea or topic of a text. En-

		(Heilman and Sagae, 2015)												
		Ena.	**Att.**	**Ela.**	**Tem.**	**Joi.**	**Cont.**	**Exp.**	**M-M**	**Cau.**	**Cond.**	**Bac.**	**Com.**	**Total**
(Wang et al., 2017) parser	**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 2: RST Parser agreement between the Heilman and Sagae (2015) parser along the x-axis and the Wang et al. (2017) parser on the y-axis, on the ICNALE dataset.

glish 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.

Table 3 shows that JOINT relations have an increased usage at the C2 level, while CONTRAST relations have a decreased usage at the C2 level. However, for these relations, the trend in usage among language learners varies.

5.3 PDTB Relations Across CEFR Levels

The relation frequencies of the Lin et al. (2014) and the Wang and Lan (2015) parsers were averaged together. As shown in Table 4, none of the level-1 PDTB relations showed a statistically different usage across CEFR levels that agreed across both datasets. C2-level users use the relation of CONTINGENCY less frequently than lower-level learners, but the trends among learners are not consistent.

5.4 Cross-Framework Results

To compare the usage of discourse relations across frameworks, we used the relation mapping proposed by Demberg et al. (2017). The mapping, shown in Table 5, proposes to map PDTB level 1 relations to RST relations.

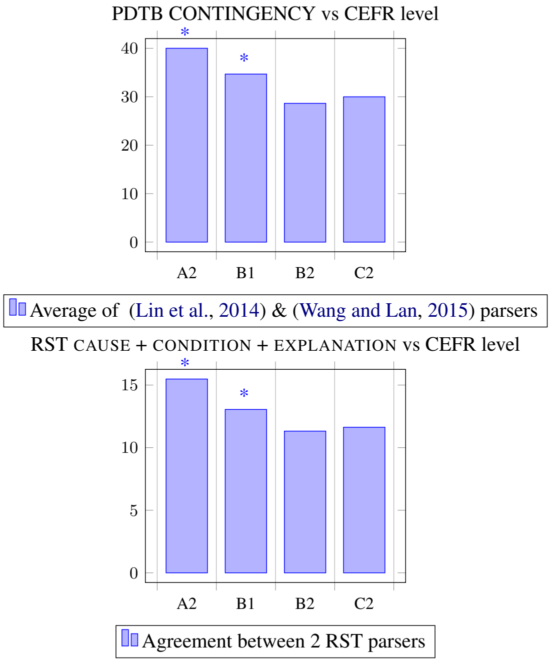

Figure 2: Percentage of CONTINGENCY across frameworks in the ICNALE dataset. The top graph shows the frequency of the level 1 relation CONTINGENCY. The bottom graph shows the average frequency of CAUSE+CONDITION+EXPLANATION. "*" indicates a statistically significant difference with C2 essays.

Using the Demberg et al. (2017) cross-framework mapping, the PDTB relation of CONTINGENCY showed an interesting comparison with the RST relations of CAUSE+CONDITION+EXPLANATION. Figure 2 compares the percentage of CONTINGENCY (the average of the two PDTB parsers) to the percentage of CAUSE+CONDITION+EXPLANATION (the agreement of the 2 RST parsers) on the ICNALE dataset. 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.

		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: Percentage of each RST relation by dataset and CEFR score.

		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 4: Percentage of each top-level PDTB relation by dataset and CEFR score.

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

Table 5: Mapping of PDTB level 1 to RST relations proposed by Demberg et al. (2017).

6 Conclusion and Future Work

In this paper, 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 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 in one dataset at the C2 level.

While discourse relations frequency would not be the sole factor for automatic CEFR assessment tools, the findings of this analysis could serve as a feature for improving the accuracy of these classifications.

Future work could look for differences in discourse relations based on the first language of the English learner, while accounting for the learner's CEFR level. The corpora used in this study provide the native language or country of origin of the learner. Previous work has begun mapping PDTB-3.0 (Prasad et al., 2019) relations to RST relations, such as (Costa et al., 2023), so future work could use inter-framework mapping with the updated PDTB. Finally, future research could 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 and 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.

Reproducibility

Work for this research included the usage of Python and a CoreNLP[3] server. Our code and a detailed description can be found on GitHub[4].

Acknowledgements

The authors would like to thank the anonymous reviewers for their feedback on the previous version of this paper. This work was financially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

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

[4] https://github.com/CLaC-Lab/discourse-parsers-and-agreement

References

Mohamed Abdalla, Frank Rudzicz, and Graeme Hirst. 2018. Rhetorical structure and Alzheimer's disease. *Aphasiology*, 32(1):41–60.

Tatsuya Aoyama. 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*, pages 1–9, Louvain-la-Neuve, Belgium. LiU Electronic Press.

Félix-Hervé Bachand, Elnaz Davoodi, and Leila Kosseim. 2014. 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)*, pages 454–468.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of Second SIGdial Workshop on Discourse and Dialogue (SIGDIAL 2001)*, pages 1–10, Aalborg, Denmark.

Nelson Filipe Costa, Nadia Sheikh, and Leila Kosseim. 2023. 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.

Elnaz Davoodi. 2017. *Computational Discourse Analysis Across Complexity Levels*. Ph.D. thesis, Concordia University Department of Computer Science and Software Engineering.

Vera Demberg, Fatemeh Torabi Asr, and Merel C. J. Scholman. 2017. How consistent are our discourse annotations? Insights from mapping RST-DT and PDTB annotations. *CoRR*, abs/1704.08893.

Vanessa Wei Feng and Graeme Hirst. 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)*, pages 511–521, Baltimore.

Michael Heilman and Kenji Sagae. 2015. Fast Rhetorical Structure Theory Discourse Parsing. *Computing Research Repository*, abs/1505.02425.

Shin Ishikawa. 2013. The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. In *Learner corpus studies in Asia and the world*, volume 1, pages 91–118.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A PDTB-Styled End-to-End Discourse Parser. *Natural Language Engineering*, 20:151 – 184.

S. Magalí López Cortez and Cassandra L. Jacobs. 2023. 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)*, pages 156–162, Toronto, Canada.

Brian MacWhinney, Davida Fromm, Margaret Forbes, and Audrey Holland. 2011. Aphasiabank: Methods for studying discourse. *Aphasiology*, 25(11):1286–1307. PMID: 22923879.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8:243–281.

Margot Mieskes and Ulrike Padó. 2018. Work smart - reducing effort in short-answer grading. In *Proceedings of the 7th workshop on NLP for Computer Assisted Language Learning*, pages 57–68, Stockholm, Sweden. LiU Electronic Press.

Charlene Pope and Boyd H. Davis. 2011. Finding a balance: The carolinas conversation collection. 7(1):143–161.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, and Eleni Miltsakaki. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, and Alan Lee. 2006. The Penn Discourse TreeBank 1.0 Annotation Manual.

Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. 2019. Penn Discourse Treebank Version 3.0. Linguistic Data Consortium.

Veronica Juliana Schmalz and Alessio Brutti. December 2021. Automatic Assessment of English CEFR Levels Using BERT Embeddings. In *Proceedings of 2021 Italian Conference on Computational Linguistics 2021 (CLiC)*, volume 3033.

Shelley Staples and Bradley Dilger. 2018. Corpus and Repository Of Writing [Learner corpus articulated with repository].

Jianxiang Wang and Man Lan. 2015. A Refined End-to-End Discourse Parser. In *Proceedings of the 19^{th} Conference on Computational Natural Language Learning (CoNLL-2015)*, pages 17–24, Beijing.

Yizhong Wang, Sujian Li, and Houfeng Wang. 2017. A Two-Stage Parsing Method for Text-Level Discourse Analysis. pages 184–188, Vancouver.

Bonnie Webber. 2009. Genre distinctions for discourse in the Penn TreeBank. In *Proceedings of the Joint Conference of the 47th Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL/AFNLP-2009)*, pages 674–682, Suntec, Singapore.

Bonnie L. Webber and Aravind K. Joshi. 1998. Anchoring a lexicalized tree-adjoining grammar for discourse. *CoRR*, cmp-lg/9806017.