Discourse Tree Structure and Dependency Distance in EFL Writing

Jingting Yuan, Qiuhan Lin, John S. Y. Lee
Department of Linguistics and Translation
City University of Hong Kong
Hong Kong SAR, China
{jingtyuan2-c,qiuhanlin2-c}@my.cityu.edu.hk
jsylee@cityu.edu.hk

Abstract

Quantitative research on learner writing has traditionally focused on lexical and syntactic features, but there has been increasing interest in incorporating discourse-level properties. This paper evaluates discourse complexity measures on learner texts in the framework of Rhetorical Structure Theory (RST). Specifically, we investigate whether discourse dependency distance and embedded structures in RST trees are correlated to learner proficiency level. In an analysis of manually annotated English essays, we found that more proficient learners tend to use longer dependency distance and more embedded structures. Further, an evaluation based on automatic discourse parsing suggests that dependency distance can potentially contribute to automatic assessment of learner texts.

1 Introduction

Text complexity depends on linguistic characteristics of the text at various levels, including lexical, syntactic and discourse features. Earlier research on text complexity mostly focused on surface features such as word length and sentence length (Kincaid et al., 1975) and n-grams (Schwarm and Ostendorf, 2005). Corpus development for discourse structure (Carlson et al., 2002; Prasad et al., 2008) has facilitated investigation into a variety of features related to discourse organization, including text cohesion, coherence and distribution of discourse relations (Lee et al., 2006; Pitler and Nenkova, 2008; Sun and Xiong, 2019). Some studies have found coherence features to be more highly correlated to text complexity than other types of features (Davoodi and Kosseim, 2016).

Text complexity is also closely related to research on language acquisition and interlanguage. Lexical and syntactic features in learner writing have been extensively studied (Jiang et al., 2019; Lu, 2011). Overuse and underuse of rhetorical relations in learner texts, and distinctive discourse structures, have also been identified (Skoufaki, 2009; Brown, 2019). Further, text complexity models can offer feedback for essay revision (Burstein et al., 2003) and support automatic assessment in language learning (Lyashevskaya et al., 2021). For example, discourse connectives and relations can help predict the level of learner proficiency (Rysová et al., 2016), and RST tree patterns can improve the performance of a speech scoring system for proficiency assessment (Wang et al., 2019).

This paper studies two types of discourse features derived from RST dependency trees. The first, *dependency distance*, refers to the linear distance between a discourse unit and its head in the dependency tree (Sun and Xiong, 2019). Second, we examine the usage of an *embedded structure* in which an elementary discourse unit (EDU) governs both its left and right neighbors, and also serves as the dependent of another discourse unit. To the best of our knowledge, these features have not yet been evaluated on their correlation to learner proficiency level. Using a corpus of texts written by learners of English as a Foreign Language (EFL), we investigate whether texts written by more proficient learners exhibit longer discourse dependency distance, and contain more embedded structures.

The rest of the paper is organized as follows. After summarizing previous work (Section 2), we define the proposed discourse complexity measures (Section 3), followed by a description of our dataset and its conversion to dependency trees (Section 4). We then present the results and analyze the correlation between these discourse complexity measures and learner proficiency level (Section 5). Finally, we conclude with a summary of our findings and suggestions for future work (Section 6).

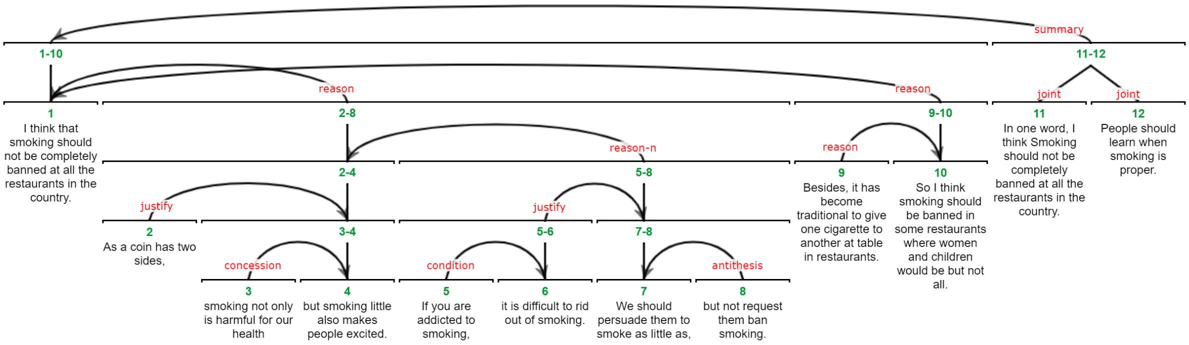

summary
1-10
11-12
reason
reason
joint
joint
1
I think that smoking should not be completely banned at all the restaurants in the country.
2-8
9-10
11
In one word, I think Smoking should not be completely banned at all the restaurants in the country.
12
People should learn when smoking is proper.
reason-n
reason
2-4
5-8
9
Besides, it has become traditional to give one cigarette to another at table in restaurants.
10
So I think smoking should be banned in some restaurants where women and children would be but not all.
justify
justify
2
As a coin has two sides,
3-4
5-6
7-8
concession
condition
antithesis
3
smoking not only is harmful for our health
4
but smoking little also makes people excited.
5
If you are addicted to smoking,
6
it is difficult to rid out of smoking.
7
We should persuade them to smoke as little as,
8
but not request them ban smoking.

Figure 1: A manually annotated RST tree from our dataset (Section 4.2)

2 Previous work

All parts of a coherent text should be held together with appropriate discourse relations. Judicious use of discourse connectives, including their frequency and diversity, as well as the distribution of discourse relations, have been found useful in predicting essay quality, the proficiency of language learners, and language development of native speakers (Crossley et al., 2016; Rysová et al., 2016; Weiss and Meurers, 2019). Davoodi and Kosseim (2016) applied features based on discourse relations and their realizations to assess the text complexity in datasets derived from the Penn Discourse Treebank (PDTB) (Prasad et al., 2008) and Simple English Wikipedia. In both datasets, coherence features were shown to be more correlated to text complexity than lexical, syntactic and other surface features.

While the PDTB focuses on text spans that explicitly or implicitly serve as arguments of a discourse connective, Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) models the full hierarchical structure of a text. According to RST, the discourse organization of a text can be represented by a constituent tree whose leaves are the elementary discourse units (EDUs). Two adjacent EDUs or text spans can be combined into a longer span with a rhetorical relation. Each EDU or span is either assigned to be the nucleus, which contains the more essential information; or the satellite, which provides supporting information. Figure 1 shows an example tree. The RST Discourse Treebank (RST-DT) (Carlson et al., 2002), which consists of 385 Wall Street Journal articles, has formed the basis of quantitative research in this framework.

Texts written by native and non-native speakers have been compared in terms of the use of RST discourse relations. In a corpus of English essays, relations concerned with ideation and/or content, called the "subject matter" relations (Brown, 2019), have been found to be used less frequently by native speakers of Japanese compared to native speakers of English. Since RST-style discourse parsing can better capture long-distance discourse dependencies, it has been found to outperform PDTB-style parsing on coherence assessment tasks such as essay scoring (Feng et al., 2014). RST tree-based features, such as tree depth, the ratio between the depth and the number of EDUs, and the frequency of different types of rhetorical relations, have been applied to proficiency assessment in the context of a speech scoring system (Wang et al., 2017; Wang et al., 2019).

To facilitate natural language processing tasks, algorithms have been proposed to convert RST constituent trees to dependency trees (Li et al., 2014). Adopting this methodology, Sun and Xiong (2019) constructed a dependency treebank from the RST-DT, examined text complexity in terms of dependency distance, and analyzed differences in discourse distance among various rhetorical relations.

ID	Elementary discourse unit (EDU)	Head	Relation	Dist.
1	We can't only focus on the working experience of a student	2	`antithesis`	1
2	Instead, the knowledge in professional field is more important	4	`reason`	2
3	when we are still students.	2	`circumstance`	1
4	So we should try to pay more attention to our study now.	n/a	`root`	0

Table 1: Dependency distance of the discourse relations in an extract from a learner essay

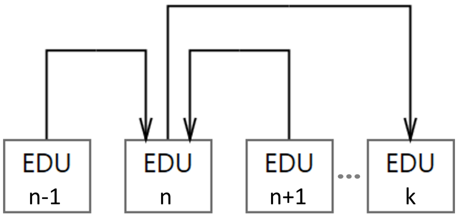


Figure 2: An embedded structure in an RST dependency tree, as defined in Section 3.2

3 Discourse complexity measures

3.1 Dependency distance

A syntactic dependency tree specifies grammatical relations between the words in a sentence. Each relation in the tree involves a governor (head) and a dependent, which modifies the head. In this context, *dependency distance* is defined as the number of words between the dependent and its head. Dependency distance can serve as a metric for syntactic complexity and language comprehension difficulty (Gibson, 1998; Liu, 2008), given its relation to cognitive load in linguistic processing (Fedorenko et al., 2013).

Following Sun and Xiong (2019), we apply the concept of dependency distance on RST trees. We define the *dependency distance* of an RST relation to be the number of EDUs between the dependent EDU and its head. Table 1 shows the dependency distance for each discourse relation in a sample text. The mean dependency distance (MDD) in an RST dependency tree can be calculated as:

$$MDD = \frac{1}{n-s}\sum_{i=1}^{n}|DD_i| \tag{1}$$

where n is the total number of EDUs; s is the total number of texts; and DD_i is the dependency distance of the i^{th} dependency link of the text.[1] We will henceforth use the term MDD in the context of the discourse dependency tree, rather than its syntactic counterpart.

We hypothesize that MDD is indicative of learner proficiency. As will be shown in our manually annotated dataset, almost 60% of the discourse relations have a dependency distance of 1, and 15% have a distance of 2 (Figure 7). We will therefore compute the proportion of relations whose dependency distance is at least 3 (to be referred to as "**length-3 relations**") and at least 4 ("**length-4 relations**"), in addition to MDD.

3.2 Embedded structures

Two major types of dependency patterns have been identified in the PDTB: a pair of discourse relations may be "independent relations", or have "full embeddings" (Lee et al., 2006). A "fully embedded" structure consists of a discourse relation that is entirely realized as an argument of another discourse connective.

Similar structural patterns in RST dependency trees could potentially help characterize learner writing. As a preliminary investigation, we examine the subtree pattern illustrated in Figure 2.[2] The pattern contains a sequence of three EDUs where the middle unit (EDU n) governs both its left and right neighbor

[1] The MDD of the sample text in Table 1 is calculated as $(|2-1|+|4-2|+|2-3|)/(4-1) = 1.3$.

[2] Dependency tree diagrams in this paper are produced with Dependency Viewer developed by the Natural Language Processing Group at Nanjing University, China (http://nlp.nju.edu.cn/tanggc/tools/DependencyViewer_en.html).

Proficiency	Essay Topic		# tokens	# tokens	# EDUs	# EDUs
level	“Part-time job”	“Smoking”		per text (SD)		per text (SD)
Native	9	9	4,045	224.7 (16.5)	352	19.6 (3.5)
B2+	9	9	4,221	234.5 (34.1)	368	20.4 (2.0)
B1	9	9	3,991	221.7 (23.2)	371	20.6 (3.4)
A2	9	9	3,820	212.2 (8.4)	388	21.6 (2.6)
Total	72		16,077	223.3 (23.5)	1,479	20.5 (3.0)

Table 2: Our dataset contains a total of 72 essays written by learners at three different proficiency levels and by native speakers.

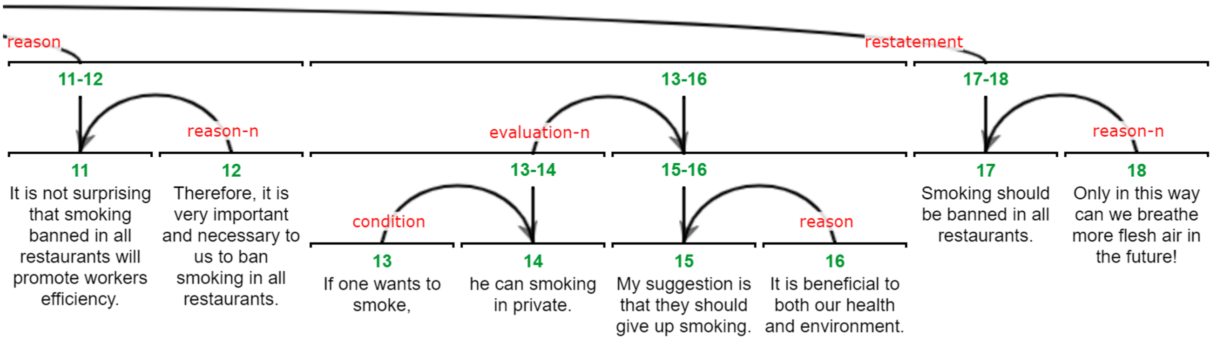

reason
restatement
11-12
13-16
17-18
reason-n
evaluation-n
reason-n
11
It is not surprising that smoking banned in all restaurants will promote workers efficiency.
12
Therefore, it is very important and necessary to us to ban smoking in all restaurants.
13-14
15-16
condition
reason
13
If one wants to smoke,
14
he can smoking in private.
15
My suggestion is that they should give up smoking.
16
It is beneficial to both our health and environment.
17
Smoking should be banned in all restaurants.
18
Only in this way can we breathe more flesh air in the future!

Figure 3: Text span 13-16 forms a dangling structure in the RST tree (Section 4.2)

(EDU $n - 1$ and EDU $n + 1$); in addition, the middle unit itself serves as a dependent of another unit (EDU k) that may precede or follow it. The text extract in Table 1 is an instance of such a structure. EDU 2 governs both EDU 1 and EDU 3, with the relations antithesis and circumstance, respectively. Further, EDU 2 is the dependent of EDU 4 in the reason relation, contributing to the writer’s opinion expressed therein.

We will henceforth use the term “embedded structure” to refer to the specific pattern in Figure 2. We hypothesize that the complex discourse organization in this structure is indicative of learner proficiency, and will calculate the proportion of EDUs in a text that exhibit this embedded structure.

4 Data

We first present the textual material of our dataset (Section 4.1). We then describe the annotation guidelines (Section 4.2) and report inter-annotator agreement (Section 4.3).

4.1 Textual material

Our dataset consists of written essays drawn from the *International Corpus Network of Asian Learners of English* (ICNALE) (Ishikawa, 2013).[3] The corpus identifies the proficiency level of each writer according to the *Common European Framework of Reference for Languages* (CEFR, 2001). We randomly selected nine writers from the subcorpus of Chinese EFL learners at the A2, B1 and B2+ levels.[4] We also randomly selected nine native speakers of English to serve as control.

To facilitate a direct comparison, we selected one essay by each writer on the topic “Part-time job”, and one essay on “Smoking”, such that all essays had similar length. The final dataset contains 72 essays from 36 writers spanning four proficiency levels, with a total of 16,077 tokens (Table 2).

4.2 Annotation guidelines

We annotated the RST structure of each essay in our dataset according to the guidelines from Stede et al. (2017), which have been adopted by a variety of corpora (Das and Stede, 2018; Musi et al., 2018). The

[3] Downloaded from http://language.sakura.ne.jp/icnale/

[4] B2+ is defined as level B2 or above. The A1 level is not available.

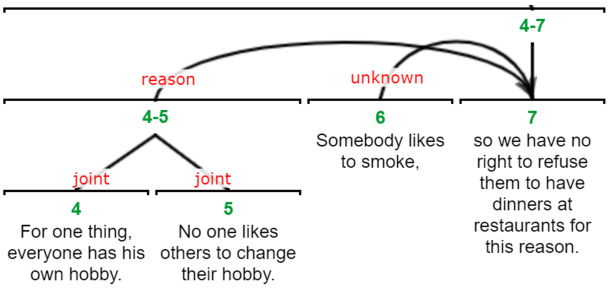
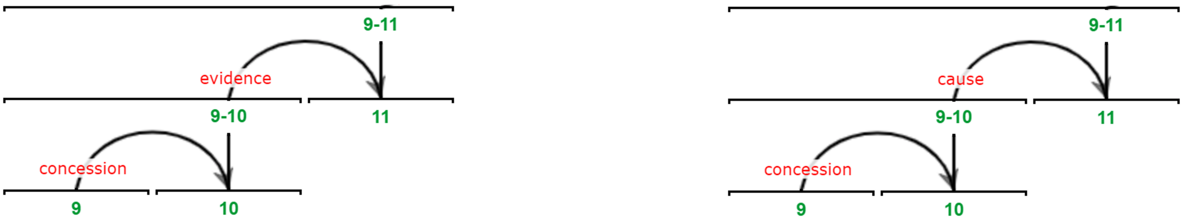

4-7
reason
unknown
4-5
6
Somebody likes to smoke,
7
so we have no right to refuse them to have dinners at restaurants for this reason.
joint
joint
4
For one thing, everyone has his own hobby.
5
No one likes others to change their hobby.

Figure 4: Example learner text annotated with an unknown relation



(a) Annotation with the evidence relation

(b) Annotation with the cause relation

Figure 5: Two interpretations for the discourse relation between text spans 9-10 and 11 in the following text: [No matter how hard they try,]^{9} [computer science textbook publishers cannot keep up with the rapid development of new computer hardware, software, and programming languages.]^{10} [C.S. student who wish to maintain a competitive edge in the job market should find part-time positions where they can learn state-of-the-art techniques.]^{11}

guidelines include 31 rhetorical relations, belonging to the pragmatic, semantic, textual and multinuclear categories. Figure 1 shows an annotated RST tree from our dataset.[5]

Since learner texts are not always grammatical, logical and well-organized, they may contain irrelevant text spans that form "dangling structures", without any link to the rest of the text (Skoufaki, 2009). Figure 3 provides such an example with the discourse span 13-16, which gives reasons for quitting smoking. The neighboring discourse spans 11-12 and 17-18, which argue why restaurants should ban smoking, have no apparent relation to this span.

The relation between two text spans is labeled as unknown when it is unclear or cannot be readily understood. Consider the text spans 6 and 7 in Figure 4. Although the connective "so" in span 7 signals a causality between these two spans, the sentence "Somebody likes to smoke" does not appear to be directly relevant to the opinion expressed in span 7, hence the unknown label.

4.3 Inter-annotator agreement

Two annotators, both with academic background in linguistics, performed the annotation using RSTTool version 3.0 (O'Donnell, 2000). To measure inter-annotator agreement, they double-annotated five texts written by learners and five written by native speakers in our dataset. Since paragraph boundaries are not marked in the raw text, the main topical units of a text can often be open to multiple interpretations. As an initial step in the annotation of each essay, the two annotators discussed its segmentation into elementary discourse units (EDUs) and larger text spans. After reaching an agreement, the annotators independently labeled the nuclearity of each span or text span, and assigned the rhetorical relations. There were a total of 94 EDUs in the native texts and 112 EDUs in the learner texts.

The two annotators achieved a Cohen's Kappa of 0.95 for nuclearity (i.e., assignment of nucleus vs. satellite) and 0.92 for relation (i.e., assignment of the rhetorical relation) on the native texts. The Kappa was slightly lower for the learner texts, at 0.94 for nuclearity and 0.86 for relation, likely reflecting the more ambiguous language therein. For relation assignment, a majority of the discrepancies (69%)

[5] RST tree diagrams in this paper are produced with rstWeb (Zeldes, 2016).

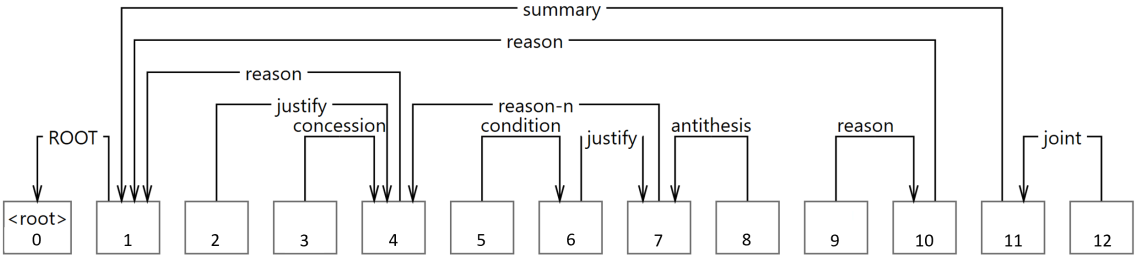


Figure 6: RST dependency tree automatically derived from the RST constituent tree in Figure 1

Complexity measure	Proficiency level			
	A2	B1	B2+	Native
Mean dependency distance	2.23	2.66*	2.97*	**3.12**
% length-3 relations	22.8	26.9	**27.8**	27.3
% length-4 relations	12.5	17.9*	20.1	**22.2**
% embedded structures	8.2	7.5	8.1	**9.5**

Table 3: Discourse complexity measures on the manually annotated dataset at different levels of language proficiency. An asterisk means the difference with the figure to its left is statistically significant.

between the two annotators occurred in the non-terminal text spans of the tree. Figure 5 shows two possible interpretations of a text. Both annotators agreed on text span 11 as nucleus and text span 9-10 as a satellite. However, one annotator regarded span 9-10 as providing `evidence` to the subjective claim in span 11 (Figure 5a), while the other saw span 9-10 as describing the `cause` of the objective state in span 11 (Figure 5b).

The RST trees for the remaining essays in our corpus were derived by one of the annotators. After manual annotation, we automatically converted the RST constituent trees into dependency trees (Li et al., 2014).[6] The RST constituent tree in Figure 1, for example, was converted to the dependency tree in Figure 6.

5 Analysis

We analyze the extent to which learner proficiency is correlated to the discourse complexity measures proposed in Section 3. Tables 3 and 4 show the results for the manually annotated and automatically parsed datasets, respectively, with a breakdown into different proficiency levels.

5.1 Manually annotated dataset

Mean Dependency Distance (MDD). MDD appears correlated to proficiency: it increases from 2.23 for A2 writers to 2.66 for B1 writers[7], then 2.97 for B2+ writers[8], then finally reaches the highest value with native speakers, at 3.12 (Table 3). A possible explanation is that, since less proficient writers need to dedicate more cognitive resources for linguistic processing, including retrieval of unfamiliar words and grammar checking, they lack the resources to produce more complex discourse (Yan and Li, 2019). The native speakers' MDD in our corpus is slightly shorter than the 3.18 observed in RST-DT (Sun and Xiong, 2019), perhaps reflecting the more formal register of news material.

Long-distance discourse relations. Texts produced by more proficient learners tend to contain more long-distance relations. In terms of the proportion of length-4 relations (see definition in Section 3.1), there is an upward trend from the A2 writers (12.5%), B1 writers (17.9%), B2+ writers (20.1%)[9], and to

[6] We used the conversion tool provided at https://github.com/amir-zeldes/rst2dep

[7] The difference between B1 and A2 is statistically significant at $p = 0.011$ by t-test.

[8] The difference between B2 and B1 is statistically significant at $p = 0.043$ by t-test.

[9] The difference between B2+ and B1 is not statistically significant ($p = 0.481$ by chi-squared test), but it is statistically significant between B2+ and A2 ($p = 0.006$).

Complexity measure	Proficiency level			
	A2	B1	B2+	Native
Mean dependency distance	1.49	1.59*	1.58	**1.74***
% length-3 relations	13.2	14.0	15.7	**18.7**
% length-4 relations	4.1	5.8	6.9	**11.5***
% embedded structures	7.5	**9.5**	8.4	9.1

Table 4: Discourse complexity measures on the automatically annotated dataset at different levels of language proficiency. An asterisk means the difference with the figure to its left is statistically significant.

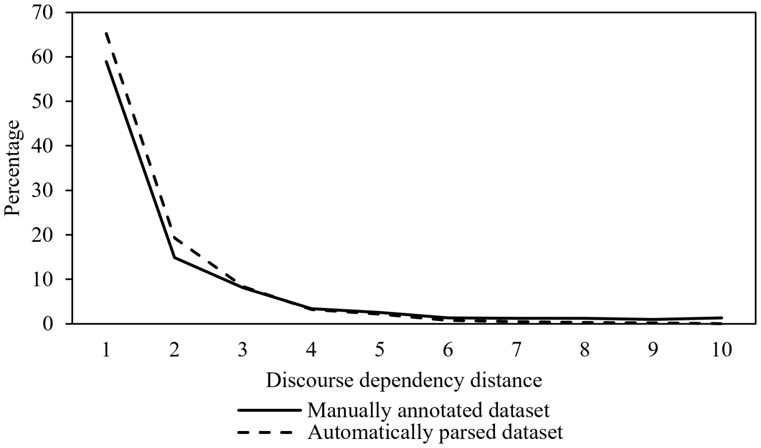


Figure 7: Distribution of discourse dependency distance in our dataset

the native speakers (22.2%) (Table 3).[10] A similar correlation is observed among the learners in terms of length-3 relations, whose proportion progresses from the A2 writers (22.8%), B1 writers (26.9%) to the B2 writers (27.8%). The differences between these groups are not statistically significant, however, perhaps in part due to the small size of the dataset. The native speakers use a higher proportion of length-3 relations (27.3%) than A2 and B1 but, contrary to expectation, their proportion is slightly lower than B2.

Embedded structures. We now turn to the embedded structures in RST dependency trees as defined in Section 3.2. The proportion of these structures shows an upward trend among the intermediate learners from B1 (7.5%) to B2+ (8.1%), and attains even higher frequency among the native speakers (9.5%). However, unexpectedly, these structures are more common in the A2 texts (8.2%) than their B1 and B2+ counterparts in our dataset. A larger dataset would be needed to clarify the correlation between the usage of this structure to language proficiency among learners.

5.2 Automatically parsed dataset

To gauge the effectiveness of the proposed features in text complexity assessment, we automatically parsed all essays in our dataset with the RST Extractor (Koto et al., 2019)[11], which implemented a state-of-the-art neural discourse parsing algorithm (Yu et al., 2018). We then converted the parser output to the dependency format, using the same procedure for the manually annotated dataset. Figure 7 compares the distribution of dependency distance between the dependency trees derived from the manually annotated and automatically parsed constituent trees.

Mean dependency distance (MDD). Across all proficiency levels, the MDD is lower in the automatically parsed dataset (Table 4) than in the manually annotated dataset (Table 3), suggesting difficulties for the automatic parser to identify long-distance relations. Table 5 shows the MDD of various discourse relations. The parser identified much fewer `evidence` and `summary` relations, for example, both of

[10] The difference between Native and A2 is statistically significant ($p = 0.001$ by chi-squared test).
[11] Downloaded from https://github.com/fajri91/RSTExtractor

Discourse Relation	Overall	Proficiency level				Frequency	
		A2	B1	B2+	Native	Auto	Manual
`means`	1.07	1.00	**1.25**	1.00	1.00	14	15
`condition`	1.08	1.00	1.00	1.05	**1.19**	65	79
`cause`	1.38	1.27	1.20	**1.85**	1.19	10	50
`evaluation`	2.00	1.50	2.00	**2.17**	1.83	13	36
`contrast`	2.82	1.83	3.33	**4.38**	2.11	19	29
`evidence`	3.34	2.78	3.12	2.37	**5.85**	4	67
`summary`	13.26	12.56	11.54	13.08	**15.83**	0	46

Table 5: Mean dependency distance (MDD) of selected RST discourse relations in the manually annotated corpus; and the frequency of these relations in the manually annotated ('Manual') and automatically produced ('Auto') datasets

which have relatively long MDD in manual annotation. The `summary` relation exhibits by far the longest MDD, at 13.26, which is comparable with its MDD of 9.34 in the RST-DT (Sun and Xiong, 2019).

MDD statistics in the automatically parsed dataset still largely demonstrate correlation to the proficiency level of EFL learners. MDD is shortest for the A2 writers (1.49), increases with the B-level writers (1.59 B1, 1.58 B2+) and reaches the largest value for native speakers (1.74).[12] However, the MDD of B1 and B2+ are indistinguishable. Hence, with automatic discourse parsing, MDD appears robust in discriminating between the beginner, intermediate and native texts, but not between subcategories at the intermediate level.

Long-distance discourse relations. Given the lower MDD in the automatically parsed dataset, the proportion of long-distance discourse relations is also lower. The correlation to the proficiency level continues to hold: the A2 writers used only 4.1% of length-4 relations; the B1 writers, 5.8%; and the B2+ writers, 6.9%[13]; and the native speakers exceeded all learners by far, at 11.5%.[14] Length-3 relations are less effective in distinguishing between the proficiency levels. While the native speakers still yielded a higher proportion (18.7%) than the learners[15], there is no statistically significant difference between A2 (13.2%), B1 (14.0%) and B2+ (15.7%).

Embedded structures. The proportion of embedded structures in the automatically parsed dataset is comparable to the manual version. These structures appear at a rate of 9.1% in native texts, more frequent than in A2 (7.5%) and B2+ (8.4%) texts. Surprisingly, however, the proportion of embedded structures is even higher at the B1 level (9.5%). This unexpected result may reflect the sensitivity of this measure to parser errors. A larger dataset would help determine if discourse parsers are sufficiently robust in detecting embedded structures, and whether the correlation holds among proficiency level and the frequency of these structures.

6 Conclusion

This paper analyzed learner texts according to discourse-level features based on RST dependency trees. Specifically, we investigated whether the dependency distance of discourse relations and the frequency of an embedded structure are correlated to learner proficiency level. Our dataset consists of English essays on similar topics written by native speakers of English, and by native speakers of Chinese at three proficiency levels.

In an analysis of the manually annotated dataset, we found mean dependency distance (MDD) to be significantly higher in texts written by more proficient learners than less proficient ones. Our results also suggested correlation between proficiency level and the proportion of discourse relation of at least

[12] The difference between B1 and A2 is statistically significant ($p = 0.046$ by t-test); so is the difference between B2+ and A2 ($p = 0.033$) and between Native and B2+ ($p = 0.002$).

[13] The difference between B2+ and A2 is statistically significant ($p = 0.043$ by chi-squared test).

[14] The difference between Native and B2+ is statistically significant ($p = 0.010$ by chi-squared test).

[15] The difference is statistically significant between Native and A2 only ($p = 0.021$ by chi-squared test).

length 4, which is significantly higher among native speakers than beginners. Further, native speakers utilized embedded structures more frequently, although the difference with learners did not reach statistical significance. In the automatic setting, despite parsing errors, texts written by beginners, intermediate learners and native speakers could still be differentiated in terms of MDD, suggesting its potential use in automatic text assessment.

In future work, we plan to expand our dataset to verify the effectiveness of the proposed complexity measures. We would also like to explore a wider range of embedded structures, as well as other discourse-level features such as coreference (Kunz et al., 2016).

Acknowledgements

We thank the anonymous reviewers for their insightful comments.

References

Jonathan D. Brown. 2019. *Using Rhetorical Structure Theory for contrastive analysis at the micro and macro levels of discourse: An investigation of Japanese EFL learners' and native-English speakers' writing*. PhD Dissertation, Leiden University.

Jill Burstein, Daniel Marcu, and Kevin Knight. 2003. Finding the WRITE stuff: automatic identification of discourse structure in student essays. *IEEE Intelligent Systems*, 18(1):32–39.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2002. *RST Discourse Treebank*. Linguistic Data Consortium.

CEFR. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, Cambridge.

Scott A. Crossley, Kristopher Kyle, and Danielle S. McNamara. 2016. The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing*, 32:1–16.

Debopam Das and Manfred Stede. 2018. Developing the Bangla RST Discourse Treebank. In *Proc. 11th International Conference on Language Resources and Evaluation (LREC)*.

Elnaz Davoodi and Leila Kosseim. 2016. On the Contribution of Discourse Structure on Text Complexity Assessment. In *Proc. 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*.

Evelina Fedorenko, Rebecca Woodburym, and Edward Gibson. 2013. Direct evidence of memory retrieval as a source of difficulty in non-local dependencies in language. *Cognitive Science*, 37(2):378–394.

Vanessa Wei Feng, Ziheng Lin, and Graeme Hirst. 2014. The Impact of Deep Hierarchical Discourse Structures in the Evaluation of Text Coherence. In *Proc. 25th International Conference on Computational Linguistics (COLING)*.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Shin'ichiro Ishikawa. 2013. The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. *Learner Corpus Studies in Asia and the World*, 1:91–118.

Jingyang Jiang, Peng Bi, and Haitao Liu. 2019. Syntactic complexity development in the writings of EFL learners: Insights from a dependency syntactically-annotated corpus. *Journal of Second Language Writing*, 46:100666.

Peter J. Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas for Navy enlisted personnel. In *Research Branch Report 8–75*. Chief of Naval Technical Training: Naval Air Station Memphis.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2019. Improved Document Modelling with a Neural Discourse Parser. In *Proc. 17th Annual Workshop of the Australasian Language Technology Association (ALTA)*.

Kerstin Kunz, Ekaterina Lapshinova-Koltunski, and Jose Manuel Martínez. 2016. Beyond Identity Coreference: Contrasting Indicators of Textual Coherence in English and German. In *Proc. Workshop on Coreference Resolution Beyond OntoNotes (CORBON)*.

Alan Lee, Rashmi Prasad, Aravind Joshi, Nikhil Dinesh, and Bonnie Webber. 2006. Complexity of Dependencies in Discourse: Are Dependencies in Discourse More Complex than in Syntax? In *Proc. 5th International Workshop on Treebanks and Linguistic Theories (TLT).*

Sujian Li, Liang Wang, Ziqiang Cao, and Wenjie Li. 2014. Text-level Discourse Dependency Parsing. In *Proc. 52nd Annual Meeting of the Association for Computational Linguistics (ACL).*

Haitao Liu. 2008. Dependency Distance as a Metric of Language Comprehension Difficulty. *Journal of Cognitive Science*, 9(2):159–191.

Xiaofei Lu. 2011. A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, 45(1):36–62.

Olga Lyashevskaya, Irina Panteleeva, and Olga Vinogradova. 2021. Automated assessment of learner text complexity. *Assessing Writing*, 49:100529.

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(3):243–281.

Elena Musi, Tariq Alhindi, Manfred Stede, Leonard Kriese, Smaranda Muresan, and Andrea Rocci. 2018. A Multi-layer Annotated Corpus of Argumentative Text: From Argument Schemes to Discourse Relations. In *Proc. 11th International Conference on Language Resources and Evaluation (LREC).*

Michael O'Donnell. 2000. RSTTOOL 2.4-A Markup Tool for Rhetorical Structure Theory. In *Proceedings of the First International Conference on Natural Language Generation (INLG).*

Emily Pitler and Ani Nenkova. 2008. Revisiting Readability: a Unified Framework for Predicting Text Quality. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In *Proc. 6th International Conference on Language Resources and Evaluation (LREC).*

Kateřina Rysová, Magdaléna Rysová, and Jiří Mírovský. 2016. Automatic Evaluation of Surface Coherence in L2 texts in Czech. In *Proc. Conference on Computational Linguistics and Speech Processing (ROCLING).*

Sarah E. Schwarm and Mari Ostendorf. 2005. Reading Level Assessment Using Support Vector Machines and Statistical Language Models. In *Proc. 43rd Annual Meeting of the Association for Computational Linguistics (ACL).*

Sophia Skoufaki. 2009. An exploratory application of rhetorical structure theory to detect coherence errors in L2 English writing: Possible implications for automated writing evaluation software. *International Journal of Computational Linguistics and Chinese Language Processing: Special Issue in Computer Assisted Language Learning*, 14(2):181–203.

Manfred Stede, Maite Taboada, and Debopam Das. 2017. *Annotation Guidelines for Rhetorical Structure*. Manuscript. University of Potsdam and Simon Fraser University.

Kun Sun and Wenxin Xiong. 2019. A computational model for measuring discourse complexity. *Discourse Studies*, 21(6):690–712.

Xinhao Wang, James V. Bruno, Hillary R. Molloy, Keelan Evanini, and Klaus Zechner. 2017. Discourse Annotation of Non-native Spontaneous Spoken Responses Using the Rhetorical Structure Theory Framework. In *Proc. 55th Annual Meeting of the Association for Computational Linguistics (ACL).*

Xinhao Wang, Binod Gyawali, James V. Bruno, Hillary R. Molloy, Keelan Evanini, and Klaus Zechner. 2019. Using Rhetorical Structure Theory to Assess Discourse Coherence for Non-native Spontaneous Speech. In *Proceedings of Discourse Relation Parsing and Treebanking (DISRPT2019).*

Zarah Weiss and Detmar Meurers. 2019. Analyzing Linguistic Complexity and Accuracy in Academic Language Development of German across Elementary and Secondary School. In *Proc. 14th Workshop on Innovative Use of NLP for Building Educational Applications.*

Hengbin Yan and Yinghui Li. 2019. Beyond Length: Investigating Dependency Distance Across L2 Modalities and Proficiency Levels. *Open Linguistics*, 5:601–614.

Nan Yu, Meishan Zhang, and Guohong Fu. 2018. Transition-based neural rst parsing with implicit syntax features. In *Proc. COLING.*

Amir Zeldes. 2016. rstWeb - A Browser-based Annotation Interface for Rhetorical Structure Theory and Discourse Relations. In *Proc. NAACL-HLT 2016 System Demonstrations*.