Article

Kun Sun*, Rong Wang and Wenxin Xiong

Investigating genre distinctions through discourse distance and discourse network

https://doi.org/10.1515/cllt-2020-0064
Received February 12, 2020; accepted February 5, 2021;
published online February 24, 2021

Abstract: The notion of genre has been widely explored using quantitative methods from both lexical and syntactical perspectives. However, discourse structure has rarely been used to examine genre. Mostly concerned with the interrelation of discourse units, discourse structure can play a crucial role in genre analysis. Nevertheless, few quantitative studies have explored genre distinctions from a discourse structure perspective. Here, we use two English discourse corpora (RST-DT and GUM) to investigate discourse structure from a novel viewpoint. The RST-DT is divided into four small subcorpora distinguished according to genre, and another corpus (GUM) containing seven genres are used for cross-verification. An RST (rhetorical structure theory) tree is converted into dependency representations by taking information from RST annotations to calculate the *discourse distance* through a process similar to that used to calculate syntactic dependency distance. Moreover, the data on dependency representations deriving from the two corpora are readily convertible into network data. Afterwards, we examine different genres in the two corpora by combining discourse distance and discourse network. The two methods are mutually complementary in comprehensively revealing the distinctiveness of various genres. Accordingly, we propose an effective quantitative method for assessing genre differences using discourse distance and discourse network. This quantitative study can help us better understand the nature of genre.

Keywords: dependency representations; discourse network; genre differences; linear distance; RST relation

***Corresponding author: Kun Sun**, Department of Linguistics, University of Tübingen, Tübingen, Germany, E-mail: kun.sun@uni-tuebigen.de. https://orcid.org/0000-0001-9766-269X
Rong Wang, Institute of Computational Linguistics, University of Stuttgart, Stuttgart, Germany; and School of Foreign Languages, Hangzhou Dianzi University, Hangzhou, China, E-mail: rong4ivy@163.com
Wenxin Xiong, School of International Chinese Studies, Beijing Foreign Studies University, Beijing, China, E-mail: xiongwenxin@bfsu.edu.cn

1 Introduction

Human language is characterized by variability. People use different linguistic forms in different situations and for different purposes. Language users make pronunciation, morphology, diction, grammar and discourse organization choices based on a multiplicity of factors. The terms "genre", "register", and "stylistics" were created to describe text (discourse) varieties, and Biber and Conrad (2019: 16) noted the differences among these terms. Although these terms have different focuses, their shared purpose is to help understand text (discourse) variability (Lee 2001). Genre studies are based on analyzing complete texts from the variety. We easily see singular genre-specific conventions, such as, some genres surrounding the text body like "greeting". However, many genre-specific features (e.g. genre-specific moves [Swales 1990; Upton 2002]) can occur repeatedly and either their occurrence or their frequencies can be genre-specific, which makes quantitative investigations into genre analysis possible. Furthermore, genre can be treated as a discourse form or type of communication with social conventions — or as a kind of broad rhetorical model that writers/users easily take advantage of when encountering familiar contexts (Hyland 2012). The method adopted in this study aims to quantitatively analyze complete texts from the perspective of discourse structure, and our purpose is to use such a method to examine text variability. Considering these, we use the term "genre" to refer to text variability in the present study.

Generally a variety of linguistic characteristics in texts are considered for genre analysis. However, the linguistic characteristics in genre studies should not be restrained in lexical and grammatical aspects. Although genre can be treated as a discourse form with social conventions, most studies on genre analysis, including qualitative and quantitative methods, have focused on lexical and grammatical aspects (Eder et al. 2016; Lee 2001; Wang and Liu 2017). In fact, very few quantitative studies (Webber 2009) have explored genre from the perspective of discourse (or discourse structure) itself. Although there is widespread awareness that a discourse (text) approach should play a crucial role in genre analysis, past studies have tended to limit themselves to conducting conceptual discussions or providing examples (Bax 2010; Fludernik 2000; Gruber and Muntigl 2005). Clearly, the frequencies of discourse structure features can be taken as genre-specific characteristics and they can be used to perform quantitative analysis in genre studies. Despite this, few studies have provided effective algorithms for quantitative investigations. Some quantitative studies claiming to conduct discourse analysis have even had to turn to lexical or syntactic devices instead of discourse (structure) itself. Thus, it is desirable that effective algorithms for use in discourse structure be developed to help in investigating distinctions among various genres.

Before proposing new algorithms, we first need to know something about discourse structure. Discourse structure mostly concerns how discourse units (or elementary discourse units, EDUs) are interrelated. For this reason, discourse relations are a core matter of concern in previous studies on discourse structure. Saussure once treated linearity as one of the defining features of human language. A sentence is produced and received linearly. Similarly, discourse is also produced and received linearly. Due to this linearity, discourse structure can be analyzed hierarchically and relationally in most cases (Sanders and van Wijk 1996; Barabási 2016; Csardi and Nepusz 2006). The hierarchical aspect focuses on the dominance of one discourse unit over another as well as the distance between connected discourse units, while the relational aspect concerns the semantic or logical meaning of connections between discourse units (i.e., discourse relation). Rhetorical structure theory (RST) (Mann and Thompson 1988) addresses both aspects and concerns the rhetorical organization of texts. The theory can therefore be used to examine discourse structure for a complete text (rather than an excerpt of a text) in genre analysis.

While the bulk of quantitative RST studies focus on rhetorical relations (Beliankou et al. 2012; Zhang and Liu 2016), very few use both hierarchical and relational dimensions together by looking at how RST relations are unequally distributed. Instead, they still focus on discourse relations. Carlson and Marcu (2001), Williams and Reiter (2003) argue that certain rhetorical relations are likely to be found with greater frequency in the higher layers in an RST tree. However, some relations appear in lower layers. This suggests that when the discourse structure was quantitatively investigated in previous studies (and even in those published in recent years), the studies confined themselves to depicting the frequency of discourse relations (Das and Taboada 2018; Iruskieta et al. 2015; Sun and Zhang 2018). However, if we are to consider both hierarchical and relational dimensions, new algorithms need to be developed for this purpose.

RST is a type of constituency-based theory in which discourse units are connected to discourse relations to build up recursively larger units up to a global unit (forming a tree). This suggests that the parsing of discourse dependency is analogous to syntactic dependency analysis (Hudson 2007; Liu et al. 2017). In simple terms, RST structures can be depicted in a simplified form using dependency structures (Li et al. 2014; Morey et al. 2018). In a dependency relation, the linear distance between a governor and dependent can potentially be utilized to provide a measure for assessing the depth of human beings' processing of sentences (Hudson 2007; Liu 2008; Liu et al. 2017). This means that dependency parsing and dependency distance algorithms are very helpful in quantitatively investigating the connection between discourse relations and discourse units. As an RST tree is converted into a syntactic dependency tree, dependency parsing can be used to analyze RST discourse relations. The syntactic dependency distance (Liu 2008) can

be applied to compute the dependency distance between each EDU in discourse, so we call it "discourse distance" which concerns linear features of discourse.

In recent years, there has been an increase in the use of networks to analyze human languages. Network theory is useful in integrating language studies, that is, in making it more internally coherent and further connecting it to external disciplines (Cong and Liu 2014; Mehler et al. 2016). As we have already seen, the focus on the frequency of discourse relations in most quantitative analyses of discourse structure has hitherto meant the comparative neglect of connections between the units and relations of discourse and of discourse units themselves. A network approach also offers, beyond the parsing of discourse, a new way of quantitatively examining how discourse units are organized and connected with each other when discourse relations are investigated, which is termed discourse network. This allows us to make sense of the different types of discourse units and determine which units can be centralized and clustered. It also allows us to see how textual coherence involves discourse units from the perspective of topological features of discourse structure. Finally, the approach used here allows us to investigate genre differences and determine whether networks in different genres vary with respect to one another. Overall, discourse distance and discourse network could be treated as new effective algorithms to examine discourse structure quantitatively, which is better than the frequency of discourse relations.

As mentioned above, previous studies on genre distinctions mostly focused on lexical or grammar features, but very few of them examined this phenomenon from a discourse structure perspective in quantitative way (Berzlánovich and Redeker 2012). RST can provide a framework for analyzing discourse structure hierarchically and relationally. The other advantage of RST is that a bulk of RST corpora have been built. Quantitative analysis will become easy if algorithms are developed to extract the data that can be genre-specific. However, some quantitative studies using RST to analyze genre still have limitations. For instance, the size of the corpus (14 texts with 4231 Dutch words) and the number of genres (two types) in Berzlánovich and Redeker (2012) seem slightly too small to use the RST corpus to distinguish genres. Another problem is that effective algorithms and statistical analyses have seldom been applied to quantitatively investigate genre distinctions. To overcome these limitations, the present study proposes two algorithms (discourse distance and discourse network) and tests them with two large-scale corpora. The approach used to merge the two algorithms may provide a unique perspective on the differences among genres. Ultimately, we will address the following two questions in this study:

1. How do the given number of genres differ from each other from the perspective of discourse distance and discourse network?
2. To what extent does the method of merging discourse distance and discourse network help draw genre distinctions?

2 Related work

2.1 RST and genre distinctions

Textual organization is one approach to understanding discourse structure. The perspective of discourse structure is therefore an effective approach for analyzing the organization of texts and is thus helpful in understanding genre. For this reason, discourse structure is likely to become a useful device in genre analysis. As mentioned in the Introduction, the data on discourse structure features can be taken as genre-specific features and such data can help us perform quantitative analysis in genre studies (more details are provided in Section 1 of the Supplementary Material, abbreviated as SM).

As discussed previously, RST is able to examine two aspects of discourse structure. Although RST has previously been associated with genre distinctions (Gruber and Muntigl 2005; Taboada and Lavid 2003), few quantitative studies regarding this association have been performed. Fortunately, as computational algorithms have been developed and an increasing number of corpora are available, the quantitative exploration of genre distinction from the RST perspective becomes possible.

According to the RST theory, a hierarchically connected structure supports each text. Each component therein interacts with the other textual elements (Carlson and Marcu 2001; Mann and Thompson 1988). In the framework of RST (Mann and Thompson 1988), the discourse structure of a text can be depicted as a tree with three fundamental components: (i) the discourse units (i.e., EDUs) are the leaves; (ii) the chief characteristic of each node is its nuclearity; and (iii) a rhetorical relation between two or more text spans is also a distinguishing characteristic. RST has been extensively used in theoretical, experimental and computational investigations of the structure of discourse (Taboada and Mann 2006). RST corpora, which were established in many languages following based on this theory (such as the RST discourse treebank [RST-DT], Carlson and Marcu 2001; Carlson et al. 2002), are of enormous aid in the quantitative analysis of discourse structure and the automatic processing of texts.[1]

Figure 1 is the RST tree of an example from rstWeb (Zeldes 2016) and illustrates how RST works. An RST tree consists of diverse relations (e.g., background, contrast, elaboration). Each relation assigns a different status (*nucleus*, *satellite*) to

1 SDRT (segmented discourse representation theory) (Asher and Lascarides 2003) and CCR (cognitive approach to coherence relations) (Sanders et al. 2018) also incorporate hierarchical analyses. However, The number of corpora annotated using the two theories are not as large as those annotated using RST.

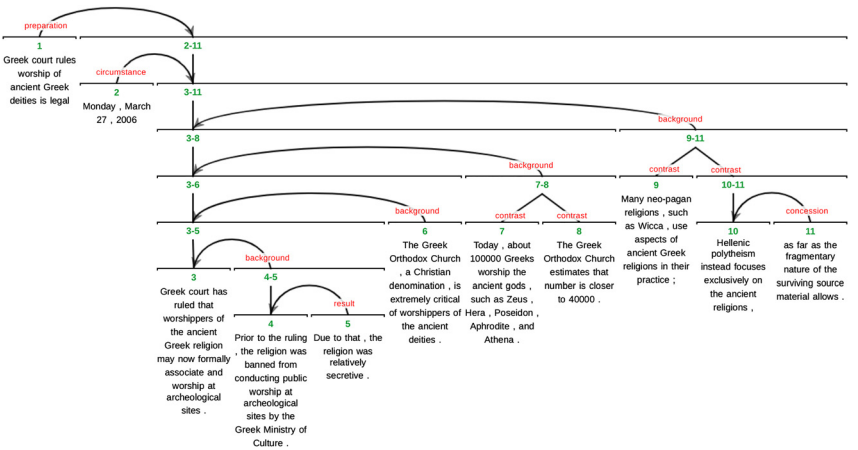

Figure 1: RST-tree of an example. The tree is a demo in rstWeb.

a node that is probably composed of one or more EDUs. There are two cases for nodes: nonterminal nodes and terminal nodes. A nonterminal node is usually composed of several EDUs. For example, node ("3–6") is a nonterminal node. Nodes "1" and "2" are terminal nodes. The nucleus is the head of an RST relation (the node where the arrow point is located in Figure 1), like the governor in a (syntactic) dependency relation. For instance, node ("3–6") has a "background" relation with node ("7–8"). Of the two nodes, the first ("3–6") is the nucleus, and the second ("7–8") is the satellite. Both nodes consist of more than one EDU (i.e., nonterminal nodes). At this stage, each node contains a nucleus EDU (leaf). We also describe terminal nodes (e.g., "7", "8") as child nodes; i.e., terminal node "7" or "8" is also called a child node (they are also leaves in an RST tree), but nonterminal node "7–8" is also called a parent node for child node "7" or "8". Note that trees in the RST follow the adjacency principle, whereby only neighboring units can be related. This implies that annotations in the RST are continuous constituent trees.

However, sometimes, several nuclei make up an RST relation. Carlson and Marcu (2001) claim that in the structure of discourse, multinuclear relations are constitutive of two or more spans of equal weight. In this sense, rhetorical relations must be either mononuclear or multinuclear. A mononuclear relation holds between two units with a nucleus and satellite, whereas a multinuclear relation holds among two or more units with a nucleus. For instance, both node "7" and node "8" are nuclei, so they form a multinuclear relation. Terminal node "9" and nonterminal node "10–11" also establish a multinuclear relation.

As mentioned in the Introduction section, as RST is an instance of constituency-based theory, the parsing of discourse dependency functions as an analogy to syntactic dependency analysis (Hudson 2007; Liu et al. 2017). In simple terms, RST structures can be depicted in a simplified form using dependency structures (Hayashi et al. 2016; Li et al. 2014; Morey et al. 2018). Hirao et al. (2013) and Li et al. (2014) defined the discourse dependency structure and determined an algorithm for transforming constituency trees in the RST annotations into dependency trees. Binary discourse relations are represented from dominant EDU (called "head/governor") to subordinate EDU (called "dependent"), which makes a nonprojective structure possible.

Stede et al. (2016) adopted the dependency tree format to compare the RST structure and SDRT structure of a corpus of short texts. The Georgetown University Multilayer Corpus (GUM) (Zeldes 2017, 2018) also annotated the same texts using both the RST-DT format and the dependency tree format. As an RST tree is convertible into a syntactic dependency tree, dependency parsing can be used to analyze RST discourse relations. There is a linear distance that runs between any given "head" node and any given "dependent" node. For a text, this is defined as the RST's "discourse distance", which uses the algorithm of *dependency distance* in dependency grammar (Liu 2008). By drawing upon the discourse distance algorithm (Sun and Xiong 2019), we can establish the average value of discourse distances for a number of texts. This will, of course, differ depending on the genre being examined for the simple reason that rhetorical structures are different in different genres. The present study is different from the study of Sun and Xiong (2019) in methodology, corpus and purpose. We use two corpora (RST-DT and GUM) to carry out our task, but Sun and Xiong (2019) only used RST-DT. Sun and Xiong (2019) intended to assess textual complexity. What is of interest to us here is whether discourse distance also varies in different genres.

2.2 Network theory and discourse network

Network science has recently become widely applied in many different disciplines. In research carried out in the last two decades, network analysis has proven capable of yielding an abundance of nuanced results in the investigation of complex systems in multiple academic fields (Barabási 2016; Newman 2018). For example, the investigation of lexicon and syntax drawing upon a network approach has been used in the majority of language studies cases (Ferstl et al. 2008; Siew et al. 2019). This approach can yield striking insights into connections between linguistic units and how those connections constitute topological relationships.

A network (or a graph) is generally considered to be a mathematical object made up of individual units—vertices (v) or nodes—that are linked by edges (e). The data on RST dependency representations can also be employed as network data. A discourse network can be used to visualize such network data, and can better observe the global connection of discourse units. Information about the RST connection between EDUs can be employed to reveal the topological relation of every node in the discourse. This might be the first study to make this kind of investigation. However, our ongoing study is not the only one to use RST networks. RST networks from other perspectives have been widely applied in language engineering. For instance, Gerani et al. (2014, 2019) used an RST network to conduct automatic discourse summarization. These studies show that discourse network approaches are reliable and effective in studies on language and language engineering.

The data on discourse dependency representations can be converted into network data. The perspective of the network is quite helpful in examining the discourse structure by helping make further distinctions among genres. Sun and Xiong (2019) proposed that merging the two methods yields an effective computational model for measuring discourse complexity. This merged method can be used not only for measuring discourse comprehension and complexity, but also for measuring the distinctions among genres.

3 Materials and methods

3.1 Materials

RST-DT Rhetorical structure theory was used to annotate the RST-DT (Carlson and Marcu 2001; Carlson et al. 2002). The treebank is composed of texts from the Penn Treebank that were acquired from the *Wall Street Journal* (WSJ). The corpus topics include finance, world news, and the arts. The RST-DT data were arranged into a training set of 347 documents and a test set of 38 documents for a total of 385 texts. The corpus is a scholarly resource that has been extensively used by researchers, and the RST-DT (Carlson et al. 2002) is recognized as a fairly homogeneous and consistent effort of RST analyses.

Webber (2009) classified the documents of the Penn Treebank into four types. Although the RST-DT also used WSJ documents in the Penn Treebank, the number of WSJ documents is larger than that of RST-DT. The 385 documents in the RST-DT were classified into four types: essays, highlights, letters and news, following Plank (2019).[2] However, a few RST-DT documents are contained in the list of this

2 http://www.let.rug.nl/~bplank/metadata/genre_files_updated.html.

classification. Ultimately, the total number of RST-DT documents in the present study is 346. The genre differences are clearly explained in Webber (2009). Letters and readers' letters belong to the opinion category. Essays include a range of forms, including press reviews, and press editorials. Highlights are similar to the "summaries" mentioned in Webber (2009). However, some headlines are also included in the highlights. The four genres are supposed to be distinct. The size of each genre corpus can be seen in Table 2 in subsection 4.1. However, the four genres in the RST-DT may be highly imbalanced, that is, the "news" genre comprises the majority of the documents. To overcome this limitation, we use a second RST-annotated corpus, the GUM.

GUM The GUM is an open-source multilayer corpus of richly annotated web texts from a number of genres (Zeldes 2017). The selection of genres in this corpus can represent different communicative purposes. The same texts in this corpus were annotated at different layers. The layers include annotations for multiple POS tags, dependency syntax, entity, coreference annotation and RST relations.

The RST annotations in the GUM are mostly consistent with those in the RST-DT. The GUM RST annotations include eight genres, which is greater than the number in the RST-DT. We used seven of eight genres included in the corpus. The seven genres are academics, biographies, fiction, interviews, news, travel guides (voyage), and how-to guides (whow). The genre of "conversation" is excluded because the conversation discourse may follow the patterns (such as speech acts, turn-taking organization etc.) which will not be applicable in the other seven genres. The size of each genre can be seen in Table 2 in subsection 4.1. The distribution of texts across the seven genres in GUM is fairly even, which is better than the distribution in the RST-DT.[3] The advantage of using two corpora is that it allows us to cross-verify the corpora with each other and thoroughly evaluate the validity of our methods.

3.2 Methods

RST Treebank converted into dependency tree As noted in the Introduction and Background sections, an RTS tree can be converted into a syntactic dependency tree, as has been investigated in many studies (Morey et al. 2018; Sagae 2009; Zhang and Liu 2016). The concern of the present study is not determining possible improvements to the algorithm used in the conversion process. Several algorithms (Hirao et al. 2013; Li et al. 2014) have recently been proposed to convert RST

3 With regard to the RST-DT, there are some data on sentence numbers for the four genres: essays (819), highlights (420), letters (242) and news (6310).

relations into dependency representations. The dependency conversion method from Hirao et al. (2013) is based on the idea of assigning each EDU in an RST-DT a unique selected head. A satellite leaf (child node) can easily identify its nearest nucleus leaf (child node) as its head. However, a nucleus leaf (child node) cannot easily identify its head. Their method can solve this problem. Traversing each nonterminal node in a bottom-up manner, the head assignment procedure determines the head based on its children in the following manner: the head of the leftmost child node with the nucleus is the head; if no child node is the nucleus, the algorithm alliteratively seeks the leftmost child node with nucleus until finds a nucleus node. Following this method, we ensure that each child node finds its unique head (a different child node). There is only one difference between Hirao et al. (2013) and Li et al. (2014): the process that finds the highest non-terminal node to which each EDU must be assigned as the head. For example, nodes "7" and "8" in Figure 1 form a multi-nuclear "contrast" relation. After the algorithm of Hirao et al. (2013) is used, "8" would be "background" to "3", shown in Table 1. By contrast, when the method of Li et al. (2014) is implemented, "8" would be "contrast" to "7". Li et al. (2014)'s method treats the first node as "head" when the nodes form multi-nuclear relation. In view of the nature of dependencies and the overall performance of the two algorithms on the two corpora, the present study uses the method from Hirao et al. (2013) to convert the RST annotations into dependency representations.[4]

We look at how an RST tree is converted into a (discourse) dependency tree. The RST example in Figure 1 illustrates how to convert RST relations into dependency representations. Child node "1" (satellite) has a "preparation" relation with the nonterminal node ("2–11"). Within this nonterminal node, the leftmost child node with a nucleus is node "3", so child node "1" has its head as "3". Child node "2" has a rhetorical relation with nonterminal node "3–11", where the leftmost child node with a nucleus is still node "3", so child node "2" has its head as "3". Child node "3" has ROOT as its head. Child node "4" (nucleus) establishes a nonterminal node ("4–5") with child node "5", and the nonterminal node has its

4 The method in Sun and Xiong (2019) might miss some dependency representations. Currently we have two algorithms to choose for conversion. Using the algorithm of Hirao et al. (2013) means that the dependency structure might not always represent the multi-nuclear relations, such as "List". However, for a multi-nuclear relation, the method of Li et al. (2014) intentionally assigns the leftmost child node as a nucleus node. Such a nucleus node is not a real "head" because all nodes in a multi-nuclear relations are equal. This means that the structures derived from multi-nuclear relations using the method of Li et al. (2014) are not real dependencies. More details on the differences between the two algorithms can be seen in Hayashi et al. (2016). Further, whichever algorithm we use, genres with many multi-nuclear relations might will be affected to the extent that the ability of accurately calculating the discourse distance could be somewhat reduced.

Table 1: Discourse dependency representations in the example of Figure 1.

Satellite (dependent)	Nucleus (head/governor)	Dependency distance	Frequency	Relation
1	(2–11)3	2	1	Preparation
2	(3–11)3	1	1	Circumstance
3	0	NA	1	ROOT
4	3	1	1	Background
5	4	1	1	Result
6	(3–5)3	3	1	Background
7	(3–6)3	4	1	Background
8	(3–6)3	5	1	Background
9	(3–8)3	6	1	Background
10	(9–11)3	7	1	Background
11	10	1	1	Concession

leftmost child with a nucleus (i.e., "3"), so child node "4" has its head as "3". Child nodes "7" and "8" are nuclei, and they constitute a nonterminal node ("7–8") and establish a rhetorical relation with another nonterminal node ("3–6"). The nonterminal node ("3–6") has its leftmost child with a nucleus, which is "3", so child nodes "7" and "8" both have the same head of "3". The information on dependency representations from this RST tree is shown in Table 1. The RST tree in Figure 1 is converted into a (discourse) dependency tree, as shown in Figure 2. All RST-DT annotation files can be drawn by both RST trees and dependency trees.

Discourse distance Dependency grammar (Hudson 2007) shows that dependencies can be adjacent or nonadjacent when a hierarchical structure is organized in a linear fashion into a sequence of words in a sentence. This implies that the two words that constitute a dependency can be found next to one another. However, they can also be separated by intervening words. This gives us the notion

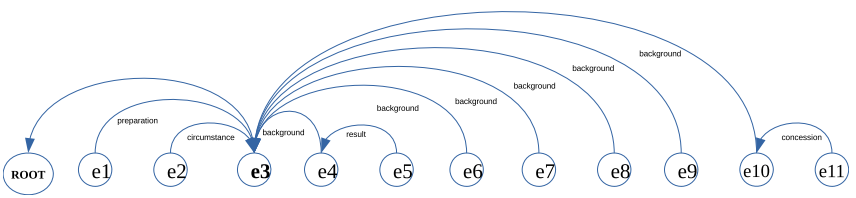

Figure 2: A dependency tree that is a convert from the RST tree of Figure 1. An EDU in RST is similar to a word in syntactic dependency analysis. The node where the arrow point is located is a *head (or governor)*. Here, "e" denotes "EDU".

of dependency distance. The number of intervening words situated between two syntactically related words or the difference in their linear position within the sentence itself determines the dependency distance.

A discourse relation is similar to a syntactic dependency. With regard to dependency representations, an EDU in discourse is similar to a word in sentence. Linear distance, that is, the distance between a "dependent" node and "head" node, can be defined as **discourse distance** for RST relations. The calculation of dependency distance in what follows uses the *dependency distance* algorithm.

Liu (2008) used the term *dependency distance*, and calculated the *mean dependency distance* (MDD) of a sentence or a text, using the following two formulas:

$$\text{MDD (the sentence)} = \frac{1}{n-1}\sum_{i=1}^{n}|DD_i| \quad (1)$$

$$\text{MDD (the text/corpus)}\frac{1}{n-s}\sum_{i=1}^{n}|DD_i| \quad (2)$$

In Equation (1), n is the number of words in the sentence, and *DDi* is the dependency distance of the *i*th syntactic link of the sentence. In Equation (2), n is the total number of words in the text, and s is the total number of sentences in the text. Ferrer-i-Cancho (2004) used a similar method to calculate the MDD of sentences. $|DD_i|$ is the absolute summation of the dependency length, which is influenced by the length of the text to some degree (i.e., total "dependency length" in some studies [Futrell et al. 2015]). However, when $|DD_i|$ is divided by $n-1$, it ensures that the mean distance is not influenced by the sentence/text length.

The RST relations in a text can be converted into dependency representations. In this sense, the dependency distance algorithm is applied to discourse dependency representations. In the following, we utilize the first equation to determine the discourse distance in every RST text. The values of discourse distance are then added up and divided by the number of RST texts. This yields the **mean discourse distance** in the RST corpus. Discourse distance, like dependency distance, is highly useful in quantitatively assessing the structure of a discourse.

Because of the conversion between an RST tree and a syntactic dependency tree, the algorithm of dependency distance can be applied in calculating the linear distance between two nodes representing the "head" and the "dependent" in RST discourse. As shown in the analysis above, we can thus obtain useful information (in Table 1) for calculating the discourse distance from the example used in Figure 1. Having acquired these data, we then use our suggested algorithm to calculate its *discourse distance* in this text: $|(3-1)| + |(3-2)| + |(3-4)| + |(4-5)| + |(3-6)| + |(3-7)| + |(3-8)| + |(3-9)| + |(3-10)| + |(10-11)|/(11-1) = 31/10 = 3.1$. The data on the dependency containing "0" in the

nucleus (i.e., a ROOT relation) are not computed in our algorithm. The programming script was written with a view to obtain the information about each of 345 WSJ texts and 130 GUM texts. This information is extremely useful not only in the calculation of discourse distance but also in carrying out investigations from a network perspective.

Network and community As discussed in the Background section, the data on discourse dependency representations can be treated as the network data. The perspective of the network greatly helps to examine the discourse structure by making further distinctions among genres.

The data shown in Table 1 (for one text) allow us to treat them as the data on network relations from this text. The network data from a number of texts contained in each genre can be assembled to form the network data for this genre. The network data contain nodes that represent individual EDUs. The cardinal number marking each node represents its position in all EDUs contained in the document. The largest node number in the RST-DT is 240. As shown in Table 1, EDUs are vertices (*v*) (or nodes), and the two EDUs are linked by an edge (*e*). The frequency of the occurrence of two EDUs represents the weight of this edge. In this way, the data on dependency representations are converted into discourse network data. There are some common parameters for describing and measuring network characteristics (Kolaczyk and Csárdi 2014). The eight parameters are employed to examine whether they are adequate to distinguish the network for each genre (Table 1 in SM simply explains these parameters).

The quantitative links or connections between the discourse units can be assessed using the topological relation for all nodes. We can view a large network as one that can be divided into different communities whose properties may exhibit a high degree of difference and variation from the average properties found in the network. We can better understand how topological relations are built from discourse unity by examining communities in the network of a discourse structure. This study not only collects basic information about the characteristics of the network, but also represents different networks with communities by using different algorithms, such as "edge.betweenness", and "walktrap" (Yang et al. 2016). The analysis of network data can be easily implemented using the R package "igraph" (Csardi and Nepusz 2006).

Statistical methods We wanted to examine whether there is a significant difference or a similarity between the network parameters in the genres for RST-DT and GUM. To this end, we employ two statistical methods. The two statistical methods will cross-verify with each other and ensure the validity of the results.

The first is a frequency-based method called ANOVA. Based on ANOVA, pairwise *t* test and Tukey's test with pairwise contrasts are performed to examine the differences in the data on network parameters among these genres.

Additionally, the data on genres are split into two pseudogenres to examine the effect size measures for cross-genre comparisons, and such tests examine whether there are significant differences among the genre validation comparisons.

The second method is Bayesian ANOVA (Gelman 2005; Gelman et al. 2019). There is, as a general rule, a significant difference between the two groups of data if the p-value in the ANOVA test is lower than 0.05. The claim that these genres are truly different in their network parameters is thus supported by the first method. However, it must be noted that p-values have been the subject of strong criticism (Nuzzo 2014). Using Bayesian measures that are roughly analogous to the frequentist p-values can make this examination process easier for investigating the hypothesis when the data size is not too large. With regard to our tests on the small-scale data, Bayesian test is a good choice. It is usually necessary to set up priors on all parameters to implement multiple Bayesian tests (see Section 4 in SM). We use the "brms" package (Bürkner 2017) to carry out the Bayesian tests.

4 Results and discussion

4.1 Results

Mean discourse distance in different genres The RST-DT this study uses contains 346 texts (or stories), and each text usually consists of at least one paragraph. The GUM contains 130 texts, each of which consists of multiple paragraphs. The RST relations were annotated throughout each text that might be composed of several paragraphs. The method of discourse distance is more appropriate here than the algorithm of dependency length (Futrell et al. 2015) mentioned above because of the length of the texts concerned. We calculated the value of discourse distance for each text using Equation (1). In this way, each text has its own discourse distance, so each genre containing a number of texts has its own "mean discourse distance". The average value of the mean discourse distance for multiple genres is 4.36 in RST-DT and 5.19 in GUM. When the discourse distance value is large, greater attention might be needed to process the text. In this respect, processing discourse distance is similar to processing dependency distance in syntax (Gibson 1998; Ferrer-i-Cancho 2004; Liu et al. 2017).

Using the same method, we calculated the mean discourse distance for each genre in two corpora. The results are shown in Table 2. Note that some texts can be classified into two genres in the RST-DT, so the total number of texts is more than 346. In Table 2, we see that each genre has its own mean discourse distance, which is also different from the mean discourse distance (4.36) in the RST-DT. Here, the mean discourse distance is the average of all texts of the RST-DT rather than the

average of the four values of discourse distance. Similarly, the discourse distance of each genre in the GUM is also distinct from this mean discourse distance (5.19). The large disparity in the GUM data distribution reflects that each genre is quite distinct in discourse distance (i.e., values range from 2.55 to 9.28). All of this could indicate that the genre types vary greatly from each other in terms of discourse distance.

According to the data on the mean discourse distance in the RST-DT, "highlights" has the largest mean discourse distance, but the "news" has the shortest. In a similar vein, the mean discourse distance data in the GUM show that "biography" has the largest mean discourse distance, but "voyage" (travel guide) has the smallest. In addition, the mean discourse distance of news in the RST-DT is quite close to that of "news" in the GUM (4.23 vs. 4.47). This consistency provides evidence that our algorithm is reliable. The discourse distance of "news" in the RST-DT is shorter than that of the majority of genres. Despite this, the discourse distance of "news" is in the middle of all genres in RST-DT and GUM.

Discourse network After converting the data on the dependency representations into network data, we used "igraph" to compute the various parameters for each genre. Each EDU is assigned as a cardinal number according to the order of its occurrence in a text. For the RST-DT, the largest number of nodes among the 346 texts is 240. Most nodes have numbers lower than 240 because of the length of the text. With regard to the GUM, the largest number of nodes among the 130 texts is 215. It can be deduced that the connectedness of those nodes with smaller numbers is denser than that of nodes with larger numbers. The reason for this is that the frequency of a small number in texts is larger than that of a larger number. Using a

Table 2: Genre and its mean discourse distance.

Genre	Text number	Mean discourse distance
RST-DT_essays	28	5.12
RST-DT_highlights	12	7.19
RST-DT_letters	12	4.42
RST-DT_news	304	4.23
	(Total) 350	**(Mean) 4.36**
GUM_academics	16	6.14
GUM_biography	20	9.28
GUM_fiction	18	3.82
GUM_interview	19	3.59
GUM_news	21	4.47
GUM_voyage	17	2.55
GUM_whow	19	2.6
	(Total) 130	**(Mean) 5.19**

network thus helps in examining these features in RST relations. However, the connectedness also depends on the frequency of the occurrence of two nodes (i.e., the weight of the edge) to some extent (see Table 1). The eight network parameters in each genre are shown in Tables 2 and 3 in SM.

In the following, we carried out three types of frequency-based statistical tests, including individual ANOVA, pairwise t test and Tukey's test, to compare the data on network parameters for an individual genre in the two corpora. The first test is individual ANOVA. The test results show that genres in the four RST-DT groups are significantly different ($F = 13.76$, $p = 1.06e{-}05$), and the genres in the seven GUM groups are also significantly different ($F = 3.411$, $p = 0.00683 < 0.05$). Pairwise *t* test shows that five RST-DT pairs (among six pairs) are significantly different ($p < 0.05$) and twelve GUM pairs (among twenty one pairs) are significantly different ($p < 0.05$). In the RST-DT, according to a Tukey's test output with pairwise contrasts, the difference between "news" and "essay" is significant, and the difference between "news" and "highlights" and the difference of "news" and "letters" are both significant, with an adjusted *p*-value of <0.001. It indicates that "news" is significantly different from the other genres. In the GUM, a Tukey's test output shows that the difference between "biography" and "news" as well as the difference between "fiction" and "news" are both significant, with an adjusted *p*-value of < 0.01; the differences between "interview" and "academics" and between "interview" and "biography" are both significant. The other three types of differences are also significant.

The Tukey's results are visualized in Figure 3, where it is easier to observe the differences among these genre pairs in the two corpora. In Figure 3, we find that each pair of genres is distinctly positioned. There are three pairs without zero effect size among six groups for RST-DT (the left panel) ($p < 0.05$), while eight pairs without zero effect size can be found among 21 groups for GUM (the right panel) ($p < 0.05$). "News" genre in the RST-DT is significantly different from the other three genres. By contrast, the GUM has only eight genre pairs which are significantly distinct. The reason for this is that the number of texts (between 16 and 21) in each genre of GUM examined is much smaller than that of the "news" genre (304) in the RST-DT. The "news" of RST-DT is significantly distinct from the other six genres in the GUM given that it is put into the GUM genres. It shows that the number size of texts in one genre plays a crucial role in Tukey's tests. When the text number of each genre was as large as that of "news" of RST-DT, more genre pairs could be significantly different in Tukey's tests. Despite this, currently the data on network parameters are still helpful in distinguishing different genres according to Tukey's tests.

In order to examine effectiveness and stability of discourse network parameters in distinguishing genres, we tested them on half of the data. After the data on

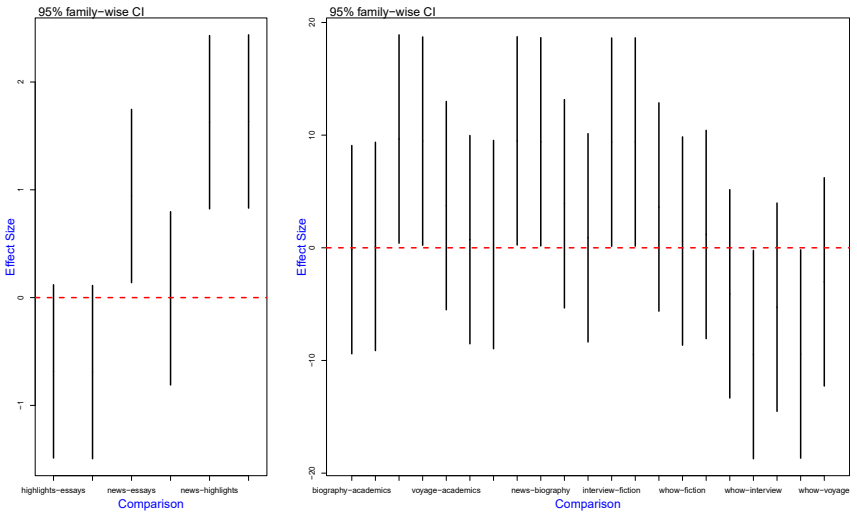

Figure 3: The size effect of Tukey's test for the data on network parameters in RST-DT and GUM. The left panel is RST-DT and the right panel is GUM. The *x*-axis represents one pair of genres (e.g., highlights-essays, biography-academics), and *y*-axis stands for effect size. When the size effect for one pair of genres contains zero, it indicates that the distinction between two genres is not so significant (i.e., with an adjusted *p*-value of >0.05). Overall, there are three significant pairs in the RST-DT and eight significant pairs in GUM.

genres were split into two pseudogenres, we examined the effect size measures for cross-genre comparisons. Half of the data for each genre are stably effective in distinguishing from each other (seven/six genre pairs can be significantly distinguished in half data. More details are provided in Section 3 in SM). Overall, all these test results show that the data on discourse network parameters are able to distinguish among all different genres in either corpus.

To perform Bayesian tests, we might set up the priors.[5] In the "brms" package, the function of "mcmcpvalue" was created to check an empirical *p*-value and thus test the hypothesis that the columns of sample have a mean of zero, as opposed to a general multivariate distribution with elliptical contours. In this sense, this function can exhibit differences from the mean standardized in the observed variance-covariance factor.[6] If the value is *zero*, it means that one group of data has *no effect* on the other; i.e., the two groups of data are truly statistically distinct.

After this test, it turns out that the "mcmcpvalue" for any two of the four groups is always *zero* in RST-DT. This means that the effects of news-essays, news-

5 The parameters of priors can be seen in SM.

6 The R script can be seen at http://www.flutterbys.com.au/stats/tut/tut7.5b.html.

highlights, news-letters, highlights-letters, highlights-essays, and essays-letters are all zero. This indicates that the parameters among the four genres are *significantly different*. With regard to the GUM, the "mcmcpvalue" for any two of the seven genres is always *zero*. This suggests that the effects of twenty-one pairs are all zero, which indicates that the parameters among the seven genres are *significantly different*.

The results from both the frequency-based tests and the Bayesian tests are consistent, thus confirming the hypothesis that the four groups of data for the RST-DT are quite distinct, and the seven groups of data for the GUM are also distinct. This means that from a network perspective, the four genres in the RST-DT differ greatly, and the seven genres in the GUM also differ markedly.

Discourse network communities In the following, we adopt network community algorithms to examine the topography and communities in the different genres in the two corpora. After inputting the data into "igraph", we plotted various discourse networks for each genre in terms of the algorithm of network communities (Yang et al. 2016). Here, we present only networks selected by "walktrap" (Pons and Latapy 2005) (see Section 5 in SM) due to space concerns.

Using the "walktrap" algorithm, each genre in the RST-DT and GUM is plotted, as shown in Figure 4 and Figure 4 of SM, respectively. These networks show that the EDUs (represented by cardinal numbers) that are proximally located have a closer relationship than those that are widely dispersed. There are a number of communities arranged in different colors for each genre in Figure 4 and Figure 4 of SM. Regardless of the genre, the first EDU (marked with No. "1") should be classified in the first community because of its location. The centralization parameter of the first EDU is the largest among all nodes. Furthermore, some nodes are shared by two community groups. In these plots, we can see that the first EDU not only has a particular role in the whole discourse network but is also closely linked with the other communities. Based on its role in the network, we can argue that the first EDU in a given discourse is absolutely essential, probably because it can be the topic sentence of a paragraph or a couple of paragraphs. Furthermore, by observing each community, we find that the EDUs with numbers in the middle are located in the center if these EDUs have formed a community.

Generally, the EDUs located in the central part are more important than those at the other locations of the discourse. This is because the EDUs in the central part are more likely to obtain connections than the EDUs at the beginning or end. The central part therefore becomes more important for comprehending this discourse. It also suggests that the block in the middle of a paragraph should be more informative. All of this can be verified by individuals with professional experience in writing English discourse (Juzwiak 2009: 107–203; Zinsser 2006: 55–67).

The four genre networks with communities in the RST-DT are clearly different. As shown in Figure 4, we find that essays and highlights are clearly distinct from

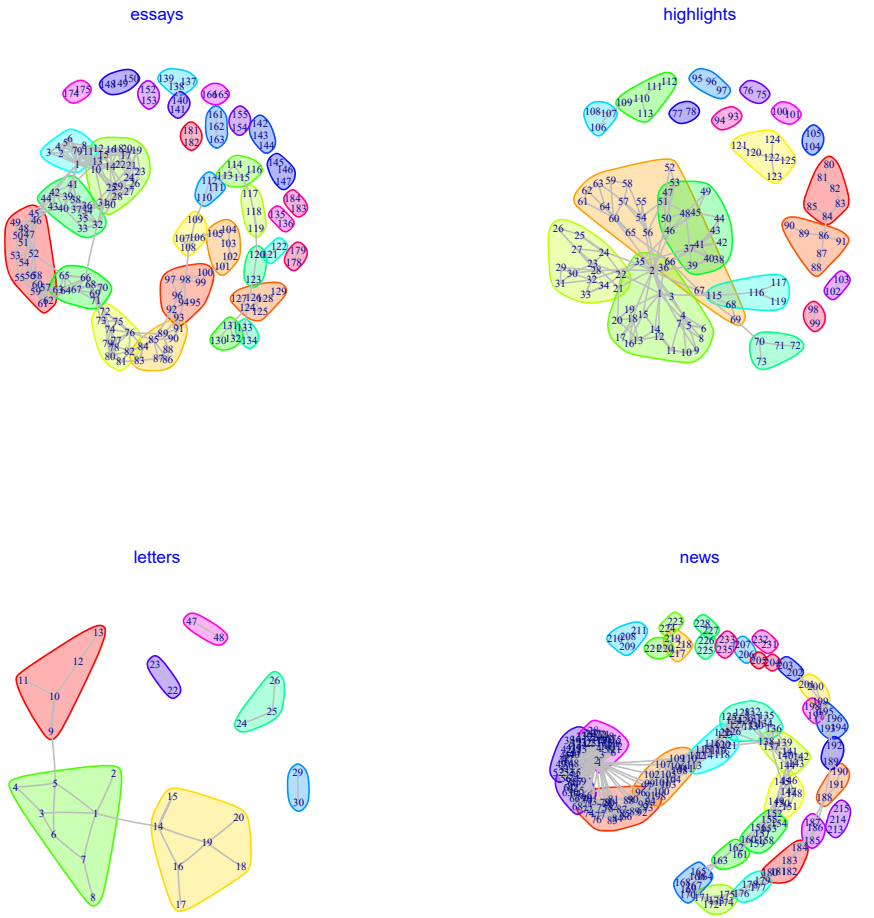

Figure 4: Four networks and communities of RST-DT with the random walk algorithm.

news in terms of shape. There is an empty area at the center of both essays and letters. In contrast, the news network has no empty center. The nodes are clustered with a high density, and the clusters are closely linked but do not form any empty areas. In any case, the node clusters are closely linked in the three genres. By contrast, the node clusters in highlights and essays are scattered and linked more loosely than those in the other two genres, which is why the former two have a longer discourse distance than the latter two. Nonetheless, we found similarities between the four genres: the nodes in the outer areas are scattered and loosely connected to the dense inside clusters. In a similar vein, the seven genre networks in the GUM are also different (more details are provided in Section 5 of SM).

4.2 Discussion

Discourse distance in different genres *Textual complexity*, conceptually similar to syntactic or lexical complexity, can be used as a measure of the complexity of textual structure for a text. *Minimal discourse distance* could be valuable for measuring textual complexity, which is analogous to measuring syntactic complexity using syntactic dependency distance.[7] Gibson (1998) proposed that the processing complexity of a sentence is related to the length of its syntactic dependencies, which has been supported by many experimental studies (Phillips et al. 2005; Temperley 2007), that is, longer dependencies are more difficult to process. Similarly, a text with a discourse distance that is greater than the minimum could indicate that the text is more complicated and harder to process. Therefore, the larger (smaller) the discourse distance of a text is, the higher (lower) its textual complexity might be.

Some texts may be read with ease and appear consistent, grammatical and lexically correct to a given reader. However, the same reader may find other texts (in a different genre) too complicated to follow. Table 2 potentially indicates the processing difficulty hierarchy using the mean discourse distance for the RST-DT: highlights > essays > letters > news. The hierarchy is likely to be consistent with a number of studies and with our predictions. The news may have become easier to read because this kind of discourse follows specific patterns and norms and because it has become more conventionalized over time (Kolodzy 2006; Van Dijk 1985). The news data and position in the hierarchy of GUM genres on discourse distance also support this argument. In contrast, the central idea of essays and letters cannot be as easily grasped as those in texts from the other two genres. Additionally, highlights are short summaries, which are difficult to process. This hierarchy is also consistent with the network communities in the four genres. The networks of highlights and essays can be similarly divided using algorithms of network communities. However, the news network community is quite distinctly separated, which is different from the network communities in the other three genres.

Table 2 potentially indicates the processing difficulty hierarchy using the mean discourse distance for the GUM: biography > academics > news > fiction > interview > whow > voyage. The hierarchy of these genres, based on discourse distance, might indicate that the hierarchy of textual complexity for the

7 Syntactic complexity can be assessed using different measures. For example, in formal theoretical linguistics (especially in the generative paradigm), the question of complexity differences among different languages does not arise. In contrast, cross-linguistic complexity differences have long been at the heart of functionalist and usage-based linguistics (Givón and Shibatani 2009). However, second language studies adopt different approaches to investigate syntactic complexity (Housen et al. 2019). Similarly, textual complexity can also be evaluated using different measures.

same genres can be arranged in this way. The genres of whow (how-to guides) and voyage (travel guides) might not be as complex as those of academics and biography. As we expected, academic writing has longer discourse distance than most of the seven genres. This indicates that academic texts do have more complicated textual structure than the other genres. That is why we usually feel that academic texts are quite more formal and difficult to process than other genres. Moreover, in both corpora, biographies have the highest discourse distance among all genres. However, few studies have paid attention to linguistic and stylistic features of biographies in comparison with news, academics and fiction. This hierarchy will be useful to make further research on genre complexity and readability. For example, we can use this hierarchy of genre discourse distance to investigate the relationship between discourse distance and other measures (e.g., lexical/syntactic complexity, text readability) for the same genre texts. Additionally, the hierarchy is also consistent with the network communities in the seven genres to some degree (see Section 5 of SM).

Each genre has a distinct discourse distance. As mentioned previously, syntactic dependency distance varies greatly in the different genres (Wang and Liu 2017). The discourse distance perspective allows us to better understand how and why the different genres are distinct. The studies of Webber (2009), Palmer and Friedrich (2014) partly support the findings in the current study. Webber (2009) found differences between genres with respect to three types of discourse relations: intrasentential discourse connectives, intersentential discourse connectives and intersentential discourse relations that are not lexically marked. For example, the percentage of implicit connectives in essays, letters and news is much higher than that in summaries. In contrast, the percentage of explicit connectives in summaries is higher than that in the other genres. The use of explicit connectives rather than implicit connectives might make reading easier, but a higher use of implicit connectives can cause the text to be more difficult to read. Palmer and Friedrich (2014) explored the relationship between the genre of a text and the types of situations introduced by the clauses in that text. This examination was carried out from the perspective of discourse mode theory (Smith 2003). Palmer and Friedrich (2014) concluded that news/jokes are different from essays/persuasive texts, which is partly consistent with our finding. However, the two studies described above did not focus on the global textual relationship in discourse.

Network communities in different genres The discourse network provides another perspective for understanding how discourse units are connected with each other and how something flows in a network. A text can consist of several network communities. Paragraphs in a texts perform like communities in a discourse network. The transitional nodes and network communities can reflect an entire network comprising several small parts, and the transitional nodes are able

to connect these small parts. These transitional nodes and network communities can reflect the importance of paragraphing in text.

Additionally, similarities can be detected between any two network figures. For example, the centrality in the discourse network types of the first EDU is confirmed. This phenomenon substantiates a narrative discourse macrostructure (Van Dijk 2019). The macrostructure or main topic in discourse is not explicitly mentioned by the RST itself. However, when we want to find the "core" sentence, that is, the central statement, the RST-based network analysis used in the present study can be very useful. In addition, as the four genre/seven genre networks exhibit, it would be a mistake to neglect the transitional nodes in each of the communities. This further indicates that, irrespective of genre types, something must be written in the text that connects two paragraphs to each other.

However, as analyzed through statistical tests, the network parameters in all of the genres are truly different. The network communities also show that these genres have formed their own different networks and communities. Despite this, different genres share similarities in terms of textual structure. For instance, discourse distance for each genre has a range (usually between 2 and 7). With regard to the network, the center cluster is very similar in various genres. The first node is essential among all genres. All of this is consistent with the data concerning discourse distance. Despite the fact that discourse distance and discourse network have two different dimensions, they are mutually complementary in revealing the distinctiveness of various genres in a comprehensive manner. The merging of the two methods can be applied in genre-based writing. For instance, the manipulation of the two measures can be successfully applied to facilitate the understanding of L1 and L2 genre-based writing (more seen in Section 6 of SM).

5 Conclusion

This study extracted data from the RST-DT and GUM by using discourse dependency representations. The algorithms of discourse distance and discourse network were taken to process the data. We found that the discourse distance for each genre has its own range. This could indicate that the difficulty of processing each genre varies. A network approach revealed that some discourse units play a more important role in the discourse structure network. The network parameters in each genre also show statistically significant differences. The network shape and communities for each genre also vary greatly from each other. We also found that the data on discourse distance for each genre are consistent with the discourse network of the genre. It is the first time that genre distinctions have been explored from the perspective of discourse structure by using discourse distance and

network algorithms. This quantitative method can effectively assess genre differences, and the merged method has revealed aspects of genre distinctions that have not been previously disclosed.

In the future, this method can be used to cross-linguistically assess more genres and can be treated as a potentially effective method in genre-based writing. Additionally, the cognitive cost of processing discourse relations might be influenced and reflected by discourse distance, which needs to be supported by psycholinguistic experiments. This is another interesting topic to explore further.

Acknowledgments: We would like to thank the three anonymous reviewers (particularly the first reviewer) for their insightful and constructive comments on the paper. We also express our sincere gratitude to the Editor-in-Chief for her great helps and generosity in improving this paper. The first author thanks his little son for his cooperation during this difficult time.
Research funding: This work was supported by the ERC (European Research Council) advanced grant (No. 742545). " The second and third authors were funded by "Important Humanities and Social Science Research Project of Zhejiang Higher Education (Fund No. 2018QN071)" and "Beijing Municipal Natural Science Foundation (Fund No.16YYB018)" respectively."

References

Asher, Nicholas & Alex Lascarides. 2003. *Logics of conversation*. Cambridge: Cambridge University Press.
Barabási, Albert-László. 2016. *Network science*. Cambridge: Cambridge University Press.
Bax, Stephen. 2010. *Discourse and genre: Using language in context*. London: Palgrave Macmillan.
Beliankou, Andrei, Reinhard Köhler & Sven Naumann. 2012. Quantitative properties of argumentation motifs. In *Methods and applications of quantitative linguistics, selected papers of the 8th international conference on quantitative linguistics*, 35–43. Belgrade: University of Belgrade.
Berzlánovich, Ildikó & Gisela Redeker. 2012. Genre-dependent interaction of coherence and lexical cohesion in written discourse. *Corpus Linguistics and Linguistic Theory* 8(1). 183–208.
Biber, Douglas & Susan Conrad. 2019. *Register, genre, and style*. Cambridge: Cambridge University Press.
Bürkner, Paul-Christian. 2017. brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software* 80(1). 1–28.
Carlson, Lynn & Daniel Marcu. 2001. *Discourse tagging reference manual. Technical Report ISI-TR-545*. University of Southern California Information Sciences Institute.
Carlson, Lynn, Daniel Marcu & Mary E. Okurowski. 2002. *RST discourse treebank (RST-DT). LDC2002T07*. Philadelphia: Linguistic Data Consortium.
Cong, Jin & Haitao Liu. 2014. Approaching human language with complex networks. *Physics of Life Reviews* 11(4). 598–618.

Csardi, Gabor & Tamas Nepusz. 2006. The igraph software package for complex network research. *InterJournal, Complex Systems* 1695(5). 1–9.

Das, Debopam & Maite Taboada. 2018. Signalling of coherence relations in discourse, beyond discourse markers. *Discourse Processes* 55(8). 743–770.

Eder, Maciej, Rybicki Jan & Mike Kestemont. 2016. Stylometry with R: A package for computational text analysis. *R Journal* 8(1). 107–121.

Ferrer-i-Cancho, Ramon. 2004. Euclidean distance between syntactically linked words. *Physical Review E* 70(5). 056135.

Ferstl, Evelyn E., Jane Neumann, Carsten Bogler & D. Yves von Cramon. 2008. The extended language network: a meta-analysis of neuroimaging studies on text comprehension. *Human Brain Mapping* 29(5). 581–593.

Fludernik, Monika. 2000. Genres, text types, or discourse modes? Narrative modalities and generic categorization. *Style* 34(2). 274–292.

Futrell, Richard, Kyle Mahowald & Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences* 112(33). 10336–10341.

Gelman, Andrew. 2005. Analysis of variance—why it is more important than ever. *The Annals of Statistics* 33(1). 1–53.

Gelman, Andrew, Ben Goodrich, Jonah Gabry & Vehtari Aki. 2019. R-squared for Bayesian regression models. *The American Statistician* 73(3). 307–309.

Gerani, Shima, Giuseppe Carenini & Raymond T. Ng. 2019. Modeling content and structure for abstractive review summarization. *Computer Speech & Language* 53. 302–331.

Gerani, Shima, M. Yashar Mehdad, Giuseppe Carenini, Raymond T. Ng & Bita Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1602–1613. Doha, Qatar: Association for Computational Linguistics.

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

Givón, Thomas & Masayoshi Shibatani. 2009. *Syntactic complexity: Diachrony, acquisition, neurocognition, evolution*. Amsterdam: John Benjamins.

Gruber, Helmut & Peter Muntigl. 2005. Generic and rhetorical structures of texts: Two sides of the same coin? *Folia Linguistica* 39(1–2). 75–113.

Hayashi, Katsuhiko, Tsutomu Hirao & Masaaki Nagata. 2016. Empirical comparison of dependency conversions for rst discourse trees. In *Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue*, 128–136. Los Angeles: Association for Computational Linguistics.

Hirao, Tsutomu, Yasuhisa Yoshida, Masaaki Nishino, Norihito Yasuda & Masaaki Nagata. 2013. Single-document summarization as a tree knapsack problem. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, 1515–1520. Seattle, USA: Association for Computational Linguistics.

Housen, Alex, Bastien De Clercq, Folkert Kuiken & Ineke Vedder. 2019. Multiple approaches to complexity in second language research. *Second Language Research* 35(1). 3–21.

Hudson, Richard. 2007. *Language networks: The new word grammar*. Oxford: Oxford University Press.

Hyland, Ken. 2012. Genre and discourse analysis in language for specific purposes. In Carol Chapelle (ed.), *The encyclopedia of applied linguistics*. Oxford: Wiley-Blackwell.

Iruskieta, Mikel, Iria da Cunha & Maite Taboada. 2015. A qualitative comparison method for rhetorical structures: identifying different discourse structures in multilingual corpora. *Language Resources and Evaluation* 49(2). 263–309.

Juzwiak, Chris. 2009. *Stepping stones: a guided approach to writing sentences and paragraphs*. Boston: Bedford/St. Martins.

Kolaczyk, Eric D. & Gábor Csárdi. 2014. *Statistical analysis of network data with R*. Heidelberg: Springer.

Kolodzy, Janet. 2006. *Convergence journalism: Writing and reporting across the news media*. Lanham, Maryland: Rowman & Littlefield.

Lee, David Y. W. 2001. Genres, registers, text types, domain, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology* 5(3). 37–72.

Li, Sujian, Liang Wang, Ziqiang Cao & Wenjie Li. 2014. Text-level discourse dependency parsing. In *Proceedings of the 52nd annual meeting of the Association for Computational Linguistics*, 25–35. Baltimore, Maryland: Association for Computational Linguistics.

Liu, Haitao. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science* 9(2). 159–191.

Liu, Haitao, Chunshan Xu & Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews* 21. 171–193.

Mann, William C. & Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text* 8(3). 243–281.

Mehler, Alexander, Andy Lücking, Sven Banisch, Philippe Blanchard & Barbara Job. 2016. *Towards a theoretical framework for analyzing complex linguistic networks*. Heidelberg: Springer.

Morey, Mathieu, Philippe Muller & Nicholas Asher. 2018. A dependency perspective on rst discourse parsing and evaluation. *Computational Linguistics* 44(2). 198–235.

Newman, Mark. 2018. *Networks*. New York: Oxford University Press.

Nuzzo, Regina. 2014. Statistical errors: P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. *Nature* 506(7487). 150–153.

Palmer, Alexis & Annemarie Friedrich. 2014. Genre distinctions and discourse modes: Text types differ in their situation type distributions. In *Workshop on frontiers and connections between argumentation theory and natural language processing*. Italy: Forlì-Cesena, July 21–25.

Phillips, Collin, Nina Kazanina, & Shani H. Abada. 2005. ERP effects of the processing of syntactic long-distance dependencies. *Cognitive Brain Research* 22(3). 407–428.

Pons, Pascal & Matthieu Latapy. 2005. Computing communities in large networks using random walks. In Pinar Yolum, Tunga Güngör, Fikret Gürgen & Can Özturan (eds.), *Computer and information sciences – ISCIS 2005*, 284–293. Heidelberg: Springer.

Sagae, Kenji. 2009. Analysis of discourse structure with syntactic dependencies and data driven shift-reduce parsing. In *Proceedings of the 11th international conference on parsing technologies*, 81–84. Paris: Association for Computational Linguistics.

Sanders, Ted & Carel van Wijk. 1996. Pisa — A procedure for analyzing the structure of explanatory texts. *Text* 16(1). 91–132.

Sanders, Ted J., Demberg Vera, Jet Hoek, Merel C. J. Scholman, Fatemeh Torabi Asr, Sandrine Zufferey & Jacqueline Evers-Vermeul. 2018. Unifying dimensions in coherence relations: How various annotation frameworks are related. *Corpus Linguistics and Linguistic Theory*. https://doi.org/10.1515/cllt-2016-0078.

Siew, Cynthia S., Dirk U. Wulff, Nicole M. Beckage & Yoed N. Kenett. 2019. Cognitive network science: A review of research on cognition through the lens of network representations, processes, and dynamics. *Complexity* 2019. 24.

Smith, Carlota S. 2003. *Modes of discourse: The local structure of texts*. Cambridge: Cambridge University Press.

Stede, Manfred, Stergos Afantenos, Andreas Peldszus, Nicholas Asher & Jérémy Perret. 2016. Parallel discourse annotations on a corpus of short texts. In *Proceedings of the tenth*

international conference on Language Resources and Evaluation (LREC'16), 1051–1058. Portorož, Slovenia: European Language Resources Association.
Sun, Kun & Wenxin Xiong. 2019. A computational model for measuring discourse complexity. *Discourse Studies* 21(6). 690–712.
Sun, Kun & Lili Zhang. 2018. Quantitative aspects of PDTB-style discourse relations across languages. *Journal of Quantitative Linguistics* 25(4). 342–371.
Swales, John. 1990. *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.
Taboada, Maite & Julia Lavid. 2003. Rhetorical and thematic patterns in scheduling dialogues: A generic characterization. *Functions of Language* 10(2). 147–178.
Taboada, Maite & William C. Mann. 2006. Rhetorical structure theory: Looking back and moving ahead. *Discourse Studies* 8(3). 423–459.
Temperley, David. 2007. Minimization of dependency length in written English. *Cognition* 105(2). 300–333.
Upton, Thomas A. 2002. Understanding direct mail letters as a genre. *International Journal of Corpus Linguistics* 7(1). 65–85.
Van Dijk, Teun A. 1985. Structures of news in the press. In Teun A. van Dijk (ed.), *Discourse and communication: New approaches to the analysis of mass media discourse and communication*, 69–93. Berlin: De Gruyter.
Van Dijk, Teun A. 2019. *Macrostructures: An interdisciplinary study of global structures in discourse, interaction, and cognition*. London: Routledge.
Wang, Yaqin & Haitao Liu. 2017. The effects of genre on dependency distance and dependency direction. *Language Sciences* 59. 135–147.
Webber, Bonnie. 2009. Genre distinctions for discourse in the Penn treebank. In *Proceedings of the joint conference of the 47th annual meeting of the ACL*, 674–682. Singapore: Association for Computational Linguistics.
Williams, Sandra & Ehud Reiter. 2003. A corpus analysis of discourse relations for natural language generation. In *Proceedings of corpus linguistics*, 28–31. U.K.: Lancaster University.
Yang, Zhao, René Algesheimer & Tessone J Claudio. 2016. A comparative analysis of community detection algorithms on artificial networks. *Scientific Reports* 6. 30750.
Zeldes, Amir. 2016. rstWeb – A browser-based annotation interface for rhetorical structure theory and discourse relations. In *Proceedings of the 2016 conference of the North American chapter of the Association for Computational Linguistics*, 1–5. San Diego, CA: Association for Computational Linguistics.
Zeldes, Amir. 2017. The gum corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation* 51(3). 581–612.
Zeldes, Amir. 2018. *Multilayer corpus studies*. London: Routledge.
Zhang, Hongxin & Haitao Liu. 2016. Rhetorical relations revisited across distinct levels of discourse unit granularity. *Discourse Studies* 18(4). 454–472.
Zinsser, William. 2006. *On writing well: The classic guide to writing nonfiction*. New York, NY: HarperCollins.

Supplementary Material: The online version of this article offers supplementary material (https://doi.org/10.1515/cllt-2020-0064).