Assessing Writing 51 (2022) 100603

Contents lists available at ScienceDirect

Assessing Writing

journal homepage: www.elsevier.com/locate/asw

Dependency distance measures in assessing L2 writing proficiency

Jinghui Ouyang [a], Jingyang Jiang [b,*], Haitao Liu [b,c]

[a] School of Foreign Languages, Tongji University, Shanghai, China
[b] Department of Linguistics, Zhejiang University, Hangzhou, China
[c] Center for Linguistics & Applied Linguistics, Guangdong University of Foreign Studies, Guangzhou, China

ARTICLE INFO

Keywords:
Syntactic complexity
Writing assessment
L2 writing proficiency
Dependency distance measures
Dependency syntactic annotation

ABSTRACT

Syntactic complexity is one of the key research foci in writing assessment. This study combines traditional syntactic complexity (SC) measures with newly-proposed dependency distance measures to better assess second language (L2) SC development. Based on a syntactically-annotated corpus of 400 compositions, we aim to investigate the extent to which the traditional SC measures and dependency distance measures can differentiate the writing proficiency of beginner, intermediate, and advanced learners. In terms of traditional SC measures, length-based measures fare better when pinpointing the different writing proficiency levels, but none of these can significantly differentiate all adjacent levels. As for dependency distance measures, the overall mean dependency distance can significantly discriminate all pairs of adjacent proficiency levels, serving as the best metric explored in our study. Moreover, dependency distance measures can further explain the findings of the traditional SC measures from the perspective of language processing.

1. Introduction

Syntactic complexity has long been an area of interest in the research surrounding writing assessment. In the past two decades, a significant number of studies have investigated the relationships between syntactic complexity (SC) in second language (L2) writing and L2 proficiency (see Ai & Lu, 2013; Martinez, 2018; Ortega, 2003; Vyatkina, 2013; Wolfe-Quintero, Inagaki, & Kim, 1998) or the quality of L2 writing (see Taguchi, Crawford, & Wetzel, 2013; Yang, Lu, & Weigle, 2015).

Language testing research and application strongly relies on quantitative measures of SC in order to effectively predict the level or the developmental stage of learners (see Egbert, 2017; Lu, 2017; Kyle & Crossley, 2017). Although some commonly used traditional SC measures (e.g., omnibus measures, such as the mean length of sentence) can be effective in assessing L2 writing proficiency (see Jiang, Bi, & Liu, 2019; Norris & Ortega, 2009), they are not based on linguistically-interpreted analyses of syntax and therefore cannot accurately reflect the syntactic characteristics of L2 writing or the nature of L2 writing development (Biber, Gray, Staples, & Egbert, 2020). More recent research has utilized more fine-grained measures, such as the frequency of relative clauses, in order to investigate SC development (see Jiang et al., 2019; Biber et al., 2020). It has been found that SC has different developmental patterns at different stages of language learning, and L2 writings exhibit different syntactic features at different stages of acquisition (Biber et al., 2020; Biber, Gray, & Poonpon, 2011). The SC measures suitable for advanced learners may not apply to or even be relevant for beginners

* Correspondence to: Department of Linguistics, School of International Studies, Zhejiang University, No. 866 Yuhangtang Road, Hangzhou, China.
E-mail address: jjy_203@yeah.com (J. Jiang).

https://doi.org/10.1016/j.asw.2021.100603
Received 30 August 2020; Received in revised form 9 December 2021; Accepted 16 December 2021
Available online 27 December 2021

(Verspoor, Lowie, Chan, & Vahtrick, 2017), and vice versa. For example, the non-finite relative clause is a syntactic feature of the advanced learning stage (Biber et al., 2011), but seldom appears at the intermediate or primary learning stages.

To summarize, the omnibus measures, i.e., MLS and the fine-grained measures such as the frequency of relative clauses have their advantages for effective assessment of L2 writing levels or L2 language proficiency (De Clercq & Housen, 2017), as well as for the description of the syntactic characteristics of L2 writings based on comprehensive syntactic information (Biber et al., 2020). Neither of these measures, however, adequately explains why they reflect the SC from the perspective of language processing. Both L1 and L2 acquisition studies have shown that long-distance syntactic relations cause processing difficulties (Fang & Liu, 2018; Futrell, Mahowald, & Gibson, 2015; Slavkov, 2015). Dependency distance, a quantitative measure from dependency syntax, presents an effective cross-linguistic metric of L1 SC, based on both experimental evidence (see Fedorenko, Piantadosi & Gibson, 2013) and corpus-based quantitative studies (see Liu, 2008). The modern thoughts of dependency syntax were first proposed by Tesnière (1959), in which the sentence structure was analyzed using dependency relations between words in a sentence (Hudson, 2007, 2010; Nivre, 2006; Tesnière, 1959). "Dependency distance" specifically refers to the distance between two syntactically-related words in a sentence and is also referred to as "dependency length" in some studies (see Gildea & Temperley, 2010; Temperley & Gildea, 2018).[1]

In order to complement traditional SC measures, we propose the use of dependency distances as SC measures to examine whether they can provide more in-depth findings. Based on a corpus of 400 compositions with syntactical annotations across five writing proficiency levels, classified according to the Common European Framework of Reference for Languages (CEFR), we aim to investigate the extent to which the traditional SC measures and the dependency distance measures can gauge different writing proficiency levels.

In the following section, we review the definition of SC, provide a theoretical background of dependency distance, and assess the advantages of dependency distance for measuring SC in SLA.

2. Background

2.1. Defining syntactic complexity

Previous definitions of SC and the broader concepts of linguistic complexity imply a variety of seemingly different notions, which can be broadly related to either absolute complexity or relative complexity (De Clercq & Housen, 2017). Absolute complexity is defined as the inherent language properties of a language feature or (sub)system and is typically actualized according to the number and types of the discrete components of which a language feature consists, and the number and nature of their internal relationships and interactions with other features (Bulté & Housen, 2012). Relative complexity refers to how costly, taxing, or difficult it is for language users and learners to learn a language feature or system of features in a given learning environment (Bulté & Housen, 2012). Although most studies of SC adopt the absolute perspective, these two approaches are clearly linked to each other (De Clercq & Housen, 2017). This research is designed to explore the SC measures that can best differentiate and predict writing proficiency. Thus, the most general definition of complexity is used in this study: namely, a quantitative notion related to the quantity and variety of the constituent elements (of an item) in an entity or system, and the relationships and interactions among the constituent parts (Rescher, 1998:1).

2.2. Dependency distance: a metric of syntactic complexity

Assuming that the syntactic structures of a sentence are composed of dependencies between individual words (Hudson, 2007; Nivre, 2006), a syntactic dependency relation comprises three core features: (1) It is a binary relation between two words; (2) It is usually asymmetrical, one of the two words serving as the governor (or head) and the other as the dependent; (3) It is classified according to the scope of general syntactic relations, as conventionally shown by the label at the top of the arc connecting the two words (Hudson, 1990, 2007; Tesnière, 1959). On the basis of these three features, we can build a directed dependency graph to represent the syntactic structures of a sentence. Fig. 1 is a dependency analysis of the sentence: "She always does homework on the weekend." The connections of all words in the sentence are actualized by syntactic relations. The main verb is the root of a sentence, which directly or indirectly governs the other elements (Tesnière, 1959). Specifically, the main verb dominates the subject and the object, the noun dominates the adjective, and so on. In each pair of two connected words, one is named the dependent and the other the governor. The arc with the label points from the governor to the dependent.

The dependency distance is the linear distance between the governor and the dependent (Heringer, Bruno, & Rainer, 1980; Hudson, 1995). For example, the dependency distance between the subject "she" and the verb "does" is two. The empirical evidence from both existing psychological experiments (see Fedorenko, Woodbury, & Gibson, 2013; Gibson & Ko, 1998; Levy & Keller, 2013) and corpus-based investigations (see Liu, 2008; Futrell et al., 2015) demonstrate that dependency distance is held as an important index of memory burden and an indicator of syntactic complexity (Liu, Xu, & Liang, 2017) or the linguistic complexity of sentence processing mechanism (Gibson, 1998). The close relationship between dependency distance and syntactic complexity has a cognitive basis. When the sentence is syntactically parsed, the words in the sentence are continuously stored in the working memory. Only when the governor of a word appears can it be deleted (Liu, 2008). According to the dependency locality theory, the longer a predicted syntactic category must be kept in memory before the prediction is satisfied, the greater the memory cost of maintaining the prediction. In addition, the

[1] In this study, we adopted the term "dependency distance."

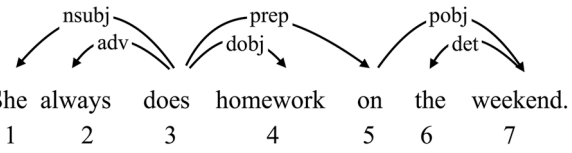

Fig. 1. Dependency analysis of the sentence "She always does homework on the weekend".

longer the distance between an incoming word and the most local governor or dependent to which it attaches, the greater the integration cost (Gibson, 1998). In other words, the longer the dependency distance, the longer the word is stored in the working memory. The capacity of human working memory is limited, which means that when the number of stored words exceeds the memory capacity, it will cause difficulties in comprehension. In this way, long-distance dependencies cause processing difficulties (Fang & Liu, 2018; Futrell et al., 2015).

The mean dependency distance (MDD) of a text is a comparative metric of linguistic complexity (Liu, 2008). Previous studies (Liu, 2008, 2009) have exemplified the rationality and high-efficiency of dependency distance as a cross-linguistic metric of SC. For example, based on dependency treebanks (corpora with dependency annotation) of 20 languages, Liu (2008) investigated the average dependency distance and other features of 20 languages and concluded that dependency distance can be used as a measure of linguistic complexity or language comprehension difficulty. With regard to the relationship between the MDD and L2 SC, Jiang and Ouyang (2017) first proposed mean dependency distance as a metric of SC in the studies of interlanguage. They stated that the SC of L2 learners' interlanguage can also be measured by dependency distance. By calculating the MDDs of Chinese EFL learners' English compositions, Jiang and Ouyang (2018) found that the MDDs of their compositions have the tendency to augment with the increase of learners' grades. This study preliminarily explored how the MDDs change with L2 learners' language proficiency. In the following section, we present the advantages of dependency distance for measuring L2 SC.

2.3. Advantages of dependency distance for measuring L2 SC in application

The advantages of using dependency distance as a measure of L2 SC are as follows: (1) Dependency distance is a cross-linguistic metric; (2) It is efficient and convenient to extract dependency relations (types) and calculate dependency distance; (3) Dependency distance is applicable to beginners' language materials with run-on sentences.

First, as a cross-linguistic metric, dependency distance reveals the ways in which cross-linguistic factors influence SC. Dependency distance is generally regarded as an important indicator of memory burden and an index of SC across languages. Because the constraint of memory burden is universal, there is a general trend for dependency distance minimization (DDM) or dependency length minimization (DLM), both of which shape a variety of syntactic patterns in human languages (Futrell et al., 2015; Liu et al., 2017). As a cross-linguistic metric of SC, dependency distance makes it convenient to study the potential impact of cross-linguistic factors (such as the language typology of native language), which has received little attention in previous L2 research (Housen, De Clercq, Kuiken, & Vedder, 2019). For instance, due to the different positions of adverbials in English and Chinese, the negative language transfer in the word order from Chinese will affect the acquisition of adverbials for beginners (Jiang, Ouyang, & Liu, 2019). Therefore, the beginners are inclined to produce sentences with adverbials in incorrect word order, as shown in Fig. 2 (the first sentence). The traditional length-based SC measures (e.g., the length of the sentence) are the same in the two sentences in Fig. 2, though one sentence is in the correct order, and the other is in the wrong order. In addition, the commonly-used quantity-based measure (the frequency of the adverbial clause modifier "with my classmate") is still incapable of reflecting the syntactic difference between the two sentences. Even so, the different dependency distances of the adverbial clause modifier between the predictive verb "go" and the preposition "with" in two sentences can clearly reflect this syntactic difference. Moreover, the different linear order of "go" and "with" in the two sentences further reflects the language transfer in word order from Chinese.

Second, it is both efficient and convenient to extract dependency relations and calculate dependency distance. As NLP technology develops, many syntactic parsers are available today, such as the Tree tagger (Schmid, 1994), and the Stanford dependency parser (Chen & Manning, 2014). All these parsers are designed to annotate the L1, but the use of annotated L2 learners' corpora has become a trend in SLA (Jiang et al., 2019). After the syntactic annotation, the final dependency treebank can be stored as a spreadsheet (Fig. 3), which is advantageous for calculating dependency distance. Since the syntactic annotation systematically identifies the target

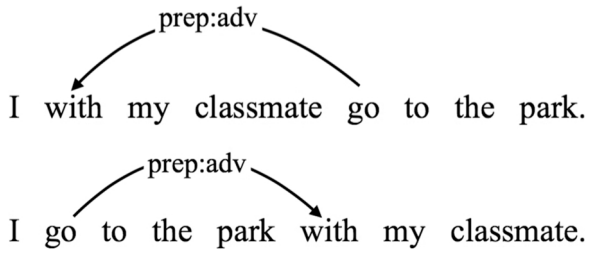

Fig. 2. Dependency analysis of adverbial "with my classmate".

structures with high efficiency (Meurers & Dickinson, 2017), it is also possible to extract and calculate the dependency distance of certain types of syntactic relations, such as the dependency distance of a subject-predicate structure. For example, the dependency distance of the subject-predicate structure between the subject "she" and the predicate verb "does" can be easily extracted and calculated. In this way, both the MDD of all syntactic structures in a text as well as the MDDs of more specific structures can be easily obtained.

Finally, dependency distance is useful in assessing the beginners' compositions with run-on sentences. The vast majority of SC research has focused on advanced learners, but it has also been noted that the SC measures for advanced learners may not be applicable to beginners (Ishikawa, 1995; Verspoor et al., 2017). It is widely acknowledged that L2 English beginner learners often produce comma splices or run-on sentences in writing (Bardovi-Harlig, 1992). Too many run-on sentences would lead to inaccurate calculations of some commonly used traditional SC measures, such as an increase in sentence length (Jiang et al., 2019). Despite being a length-based measure, dependency distance is not influenced by run-on sentences. As illustrated in Fig. 4, "I like PE class, I can play" is a run-on sentence in A1 level composition. Because of the lack of conjunctions, there is no dependency relation (dotted line) between "like" and "play," which should appear when the sentence is grammatically correct. The MDD of this run-on sentence is the same as when it is split into two independent and grammatically correct sentences ("I like PE class. I can play"). Compared with the correct version with the conjunction ("I like PE class, because/and I can play"), the syntactic complexity of two split independent sentences ("I like PE class. I can play") can better reflect the true syntactic competence of low-proficiency learners, as they haven't mastered the correct usage of conjunctions. In this case, the MDD accurately reflects the syntactic complexity of the run-on sentence.

2.4. Current research

This research assesses L2 writing by exploring the SC of L2 compositions of beginners, intermediate, and advanced learners, adopting both traditional SC measures and MDD measures. Consensus holds that SC is a multi-dimensional construct with the following four dimensions: the length of production unit, amount of subordination, amount of coordination, and degree of phrasal sophistication (see Ai & Lu, 2013; Lu, 2017; Norris & Ortega, 2009). In this study, we also adopt these four dimensions as traditional SC measures because they have been commonly used in previous studies (see Jiang et al., 2019; Ai & Lu, 2013; Lu, 2017). As for dependency distance measures, besides the overall MDD of each writing proficiency level, MDD measures of subordination, coordination, and noun phrases are also included. This is because subordination-related and coordination-related measures are those most commonly employed in other L2 SC studies (Bulté & Housen, 2012; Lu & Ai, 2015; Ortega, 2003). Furthermore, until recently, the use of noun phrases has also been treated as the main SC feature of L2 writings (Biber et al., 2011; Lu, 2011; Parkinson & Musgrave, 2014). We collected English compositions of Chinese learners ranging from primary schools to universities in this study, encompassing a total of fifteen grades. Based on CEFR, learners were classified into five different levels according to their compositions. The L2 Syntactic Complexity Analyzer (L2SCA) (Lu, 2010) and the dependency syntactically-annotated corpus was used to collect the data of traditional SC measures and MDD measures, respectively. The research questions are as follows:

RQ1. . To what extent can traditional SC measures differentiate among writing proficiency levels of beginner, intermediate, and advanced learners?

RQ2. . To what extent can the MDD measures differentiate among writing proficiency levels of beginner, intermediate, and advanced learners?

3. Methods

3.1. Participants and materials

The participants included 800 students from an elementary school, a middle school,[2] and a university in Zhejiang Province, in eastern China. They ranged from the fourth grade of elementary school to English majors in the fourth year of undergraduate, representing learners at different English learning stages in China (Table 1). Their mother tongue was Chinese, and they had no long-term experience (more than six months) of living or studying in an English-speaking country. The compositions were rated according to the CEFR written assessment grid (Council of Europe, 2018, pp. 173–174). The grid consists of the descriptor scales for five linguistic competences: (vocabulary) range, coherence, (grammatical) accuracy, description, and argument at levels A1 (breakthrough), A2 (waystage), B1 (threshold), B2 (vantage), C1 (effective operational proficiency), and C2 (mastery).

After scoring all 800 compositions, we found that most students' writing proficiency levels were concentrated at Pre-A1 (99), A1 (90), A2 (124), B1 (188), B2 (224), and C1 (70) (Pre-A1 means lower than A1 level). The distribution of student compositions across five levels and different learning stages are presented in Table 1. It is less relevant to calculate the SC of Pre-A1 compositions, as there are too many L1 words and language mistakes. Additionally, the number of C2 compositions was too small (only five). Thus, the Pre-A1 and C2 compositions were excluded. Thus, this study focused on A1 to C1, covering five levels of student compositions, effectively incorporating most Chinese English learners of different writing proficiency levels. To balance the number of compositions across the five levels, compositions of almost the same amount at each level were randomly chosen regardless of the grade, which guaranteed the

[2] The L2 compositions of middle school students in this study were selected from the same pool of a previously published study in the *Journal of Second Language Writing.*

Text number	Sentence number	Word order	Word	POS	Governor order	Governor	Dependency relation	Dependency distance
t10	s7	1	She	PRP	3	does	nsubj	2
t10	s7	2	always	RB	3	does	advmod	1
t10	s7	3	does	VBZ	3	does	root	0
t10	s7	4	homework	NN	3	does	dobj	1
t10	s7	5	on	IN	3	does	prep	2
t10	s7	6	the	DT	7	weekend	det	1
t10	s7	7	weekend	NN	5	on	pobj	2
t10	s7	8	.	.	3	does	punct	5

Fig. 3. Screenshot of the spreadsheet form of the dependency treebank.

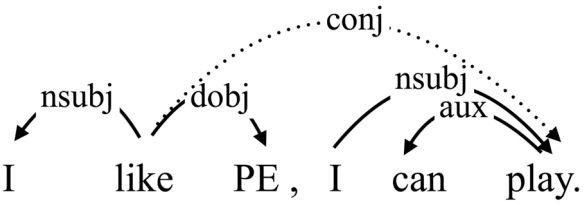

Fig. 4. Dependency analysis of the run-on sentence.

Table 1
The distribution of students' compositions across seven levels and different learning stages.

Grade	Pre-A1	A1	A2	B1	B2	C1	C2	Total
Fourth year elementary school	51	4	0	0	0	0	0	55
Fifth year elementary school	33	21	0	0	0	0	0	54
Sixth year elementary school	15	40	0	0	0	0	0	55
First year junior high school	0	17	38	0	0	0	0	55
Second year junior high school	0	8	46	0	0	0	0	54
Third year junior high school	0	0	33	20	0	0	0	53
First year senior high school	0	0	7	46	0	0	0	53
Second year senior high school	0	0	0	43	11	0	0	54
Third year senior high school	0	0	0	45	8	0	0	53
First year undergraduate (non-English major)	0	0	0	19	35	0	0	54
Second year undergraduate (non-English major)	0	0	0	15	36	1	0	52
First year undergraduate (English major)	0	0	0	0	37	16	0	53
Second year undergraduate (English major)	0	0	0	0	35	17	0	52
Third year undergraduate (English major)	0	0	0	0	33	17	2	52
Fourth year undergraduate (English major)	0	0	0	0	39	19	3	51
Total	99	90	124	188	224	70	5	800

randomness. In total, we used 400 compositions, and the distribution of learner texts and the total numbers of words per CEFR level are shown in Table 2.

The self-built corpus contains 400 English compositions with a total of 61,699 words. For a better vertical comparison, we controlled the genre and topic of the compositions. With the exception of those in elementary school, students (middle school, high school, and college) were required to write the narratives. The assigned topics were about the students' own experiences, with titles such as "An Unforgettable Experience", "An Interesting Experience", and "My Weekend." We acknowledge that familiar topics, like personal experiences, may require less causal reasoning (Robinson & Gilabert, 2007; Robinson, 2011) and topics requiring less causal reasoning may result in less complex writing (Yang et al., 2015). Considering the fact that the L2 levels of the participants spanned a wide range, however, and that the L2 level of some of the participants in the study was very low, these relatively simple and familiar topics encouraged the low-proficiency participants to write more text. The elementary school students, who had limited English vocabulary and little training in writing, were asked to describe a colorful picture entitled "My Weekend" in English. We allowed them to resort to L1 if necessary in order to encourage them to write as much as they could. In the scoring process, these compositions with L1 words were all classified as being at the Pre-A1 level.

3.2. Procedure

In English class, the students were given 30 min to write their compositions. In turn, they were not allowed to use any reference materials, as though they were taking a formal exam. The compositions were then collected by their teachers and sent to the

Table 2
The basic information of five self-built sub-corpora.

	Number of Texts	Total Number of Words
A1	90	6680
A2	85	8037
B1	80	11675
B2	75	17399
C1	70	17908
Total	400	61699

researchers. Before participating, they were told that their compositions would be used for academic purposes, and the students consented to their participation in the study. Subsequently, each participant received a gift as a reward. The students' compositions were input into the computer in their original form, including spelling, punctuation, capitalization, and language mistakes.

Each composition was initially given a rating by two scorers, and then a third scorer in case of discrepancy. These scorers were Chinese English teachers with extensive experience grading compositions. Before the formal scoring took place, the three scorers familiarized themselves with the *CEFR for Languages: Learning, Teaching, and Assessment* (Council of Europe, 2018), scored the writing samples from *The CEFR Grids for Writing* developed by ALTE members (Cambridge University Press, 2019), and agreed upon the scoring standard. In the formal scoring process, the scorers were required to first assign a CEFR level for each linguistic competence, and then assign an overall CEFR level for each composition, in order to provide a general assessment on the proficiency in each composition. In terms of the overall CEFR level, the two scorers were perfectly aligned on 580 compositions (72.5% of the 800 texts). They disagreed by one level on 216 compositions (27%) and by two levels on 4 compositions (0.5%), all of which were then submitted to the third scorer. For these compositions, the final level was the level agreed on by two of the three raters. This rating method was previously adopted by Paquot (2018), in whose research the rate of perfect agreement was 55%. Therefore, in this study, the rate of perfect agreement 72.5% was relatively high. After rating and random selection, the five sub-corpora were built and annotated.

In this study, we used semi-automatic annotation, machine annotation plus manual revision, instead of complete machine annotation to tag all the words and sentences in the text. We did this for two reasons. First, syntactic annotation parsers (e.g., Stanford Parser) are designed for L1 annotation. Thus, the reliability is questionable in the annotation of L2 learners' compositions, and in order to achieve better accuracy we used semi-automatic annotation (Jiang et al., 2019). Second, we added some more detailed sub-types of the original dependency relations based on the annotation manual of Stanford Parser 3.6.0, such as the acl:relcl:attr (adjective relative clause). This step also required manual annotation. After Stanford Parser 3.6.0 automatically did the POS (Part-of-Speech) and dependency relation annotation, two postgraduate students of applied linguistics examined and modified the automatic annotations.

To guarantee the consistency and the accuracy of manual modifications, the two postgraduate students examined and modified the same 100 compositions after the automatic annotation process. Then a senior linguist specialized in SLA and another specialized in dependency syntax assisted with the revisions. After that, the postgraduate students and linguists discussed the annotations of these 100 compositions and agreed upon the modifications. Later, the two postgraduate students continued to examine and manually-annotate the remaining 300 compositions. The two postgraduate students' agreement rate on the revisions reached 89.5%, indicating the inter-annotator agreement rate was acceptable. Any inconsistent revisions were submitted to the senior linguists for final revisions. Finally, the dependency syntactically-annotated corpus was stored as an Excel spreadsheet, as shown in Fig. 3.

3.3. Measures

3.3.1. Traditional SC measures

We adopted L2SCA, a program specially designed for L2 linguistic materials, by Lu (2010), to calculate the traditional SC measures. Reportedly, L2SCA can provide highly reliable results for the coding of SC measures in L2 linguistic materials (Lu, 2010). Although it is still to be determined whether it will generate reliable data from less grammatically correct language materials, studies (see Jiang et al., 2019) have already shown that L2SCA also works well for these language materials. L2SCA can generate the data of 14 measures, involving four dimensions: the length of production unit; the amount of subordination; the amount of coordination; and the degree of phrasal sophistication (Lu & Ai, 2015). As some of the fourteen measures are redundant (Norris & Ortega, 2009), one measure from each dimension is chosen, for a total of four measures. In addition, all three length-based measures (MLS, MLC and MLT) were

Table 3
The seven traditional SC measures (adopted from Lu & Ai, 2015).

Dimensions	Measures	Codes	Definitions
Length of production unit	Mean length of clause	MLC	# of words/# of clauses
	Mean length of sentence	MLS	# of words/# of sentences
	Mean length of T-unit	MLT	# of words/# of T-units
Amount of subordination	Dependent clauses per clause	DC/C	# of dependent clauses/# of clauses
Amount of coordination	Coordinate phrases per clause	CP/C	# of coordinate phrases/# of clauses
	T-unit per sentence	T/S	# of T-units/# of sentences
Degree of phrasal sophistication	Complex nominal per clause	CN/C	# of complex nominal/# of clauses

maintained, because although they all measure the length of grammatical units, they target at different grammatical levels. As a result, seven measures were used in this study: mean length of clause (MLC); mean length of T-unit (MLT); mean length of sentence (MLS); dependent clauses per clause (DC/C); coordinate phrases per clause (CP/ C); T-units per sentence (T/S); and complex nominal per clause (CN/C) (Table 3).

3.3.2. MDD measures

We adopted the calculation proposed by Liu (2009) to calculate MDD measures in each composition. To begin, we rendered all the words in a sentence into the form of a word string. For the dependency relation between two words Wa and Wb ('a′ is the word order of Wa; 'b′ is the word order of Wb), if Wa is the governor and Wb is its dependent, then the dependency distance between them is the difference in the value of 'a-b′. When 'a′ is larger than 'b′, the dependency distance is a positive value, indicating that the governor is after the dependent; when 'a′ is smaller than 'b′, the dependency distance is a negative value, which means that the governor is before the dependent. However, when calculating the dependency distance, the absolute value is used. The MDD of a sentence or a composition is the mean value of all dependency distances. For example, a set number of dependency distances can be obtained from the sentence in Fig. 1 as follows: 2, 1, 1, 2, 2, and 1. The MDD of this sentence is $9/6 = 1.5$.

Furthermore, we modified this calculation method in order to calculate the MDD of a certain type of dependency relation in a text. The dependency relations concerning the subordination, coordination, and noun modifier are listed in Table 4. The measures used to calculate the MDDs of these dependency types are presented in Table 5, and the specific calculation processes are shown in Appendix A with a student composition provided as an example. Four MDD measures are used in this study: the overall mean dependency distance (OMDD); the mean dependency distance of subordinations (SMDD); the mean dependency distance of coordinations (CMDD); and the mean dependency distance of noun modifiers (NMDD).

3.4. Data analysis

The Shapiro-Wilk test and Q-Q plots indicated that only one measure, the overall MDD, followed the normal distribution. The one-way ANOVA test was used to determine whether there was significant difference in the OMDDs across the five levels. The Bonferroni post hoc test was then used to examine whether a significant difference could be traced between every two adjacent levels. As for other SC and MDD measures, the Kruskal-Wallis test was employed to analyze the group differences among five levels. Finally, paired-comparisons were conducted in order to determine whether the differences were significant in these SC and MDD measures between every adjacent levels.

4. Results

In this section, we present the statistical results for traditional SC measures and MDD measures.

4.1. Traditional SC measures

In Table 6 and Fig. 5, only the MLC, the MLS, the MLT, and the DC/C increase with the development of writing proficiency. The T/S fluctuates with no regular changes on the writing proficiency level. The CN/C grows from A2 to C1, but with a decrease from A1 to A2. The CP/C descends from A1 to B1, and then increases from B1 to C1.

Kruskal-Wallis revealed significant effects of writing proficiency levels on four SC measures: MLC (X^2=230.993 , $p = .000$, η^2=.575), MLT (X^2=257.087 , $p = .000$, η^2=.641), MLS (X^2=277.239 , $p = .000$, η^2=.692), and DC/C (X^2=208.223 , $p = .000$, η^2 =.517). The effect sizes of MLC, MLS, MLT, and DC/C all exceed 0.5, indicating over 50% of the variances of the four measures could be explained by writing proficiency (Fritz, Morris, & Richler, 2012). Among them, the MLS has the biggest effect size, reaching nearly 0.7. Follow-up paired-comparisons (Table 7) show that there were significant differences between two pairs of adjacent levels in the MLC, the MLS, and the MLT, and one pair of adjacent levels in the DC/C.

Among all traditional SC measures, length-based measures (MLC, MLS, MLT) perform best in gauging the writing proficiency of beginner, intermediate, and advanced learners. However, none can significantly differentiate all adjacent writing proficiency levels.

4.2. MDD measures

As shown in Table 8 and Fig. 6, the OMDD, the SMDD and the NMDD all rise with the increase of writing proficiency. However, the CMDD increases at the start but then slightly decreases. An one-way ANOVA revealed a significant effect of writing proficiency on the OMDDs (F(4, 395)= 154.119 , $p = .000$, η^2= .609). Kruskal-Wallis revealed significant effects of writing proficiency levels on the other two MDD measures: NMDD (X^2=166.161 , $p = .000$, η^2=.411) and SMDD (X^2=67.365 , $p = .000$, η^2=.219).

The results of Bonferroni post hoc test (Table 9) show that there were significant differences between all four pairs of adjacent levels in the OMDDs. The results of follow-up paired-comparisons (Table 9) indicate that there were significant differences between two pairs of adjacent levels in the NMDDs, and there was a significant difference between only one pair of adjacent levels in the SMDDs. Based on these results, the OMDD has the potential to best evaluate the writing proficiency of L2 learners and can significantly differentiate every pair of adjacent levels.

Table 4
Dependency relations concerning subordination, coordination and noun modifier.

Dimensions	Dependency Relations	Codes	Examples
Subordination	Adverbial clause	advcl	If you know who did it, you should tell the teacher.
	Sentential relative clause	acl:relcl1	He walked slowly, which made me impatient.
	Clausal subject	csubj	What she said makes sense.
	Clausal complement	ccomp	I know that you like me.
Coordination	Conjunct	conj	He is clever and honest.
Noun modifier	Determiner	det	the man
	Possessive modifier	nmod:poss	their office
	Adjectival modifier	amod	beautiful garden
	Numeric modifier	nummod	three apples
	Compound noun	compound	school department
	Prepositional phrase as attributes	prep:attr	captain of team
	Adjectival relative clause	acl:relcl:attr	the thing that I need
	Participle modifier	acl:attr	a poor student named Tom

Table 5
The four MDD measures.

Measures	Codes	Definitions
Overall mean dependency distance	OMDD	the total dependency distances/the total number of dependency relations
Mean dependency distance of subordinations	SMDD	the dependency distances of subordinations/the number of subordinations
Mean dependency distance of coordinations	CMDD	the dependency distances of coordinations/the number of coordinations
Mean dependency distance of noun modifiers	NMDD	the dependency distances of noun modifiers/the number of noun modifiers

Table 6
Descriptive statistics for traditional SC measures.

Measure	A1		A2		B1		B2		C1	
	M	SD	M	SD	M	SD	M	SD	M	SD
MLC	5.846	1.368	6.599	0.995	6.791	0.956	8.694	1.398	10.227	1.981
MLS	8.128	3.258	8.630	2.132	12.958	3.824	15.954	3.520	18.505	4.074
MLT	7.249	2.508	7.990	1.568	11.172	2.798	14.256	2.947	16.857	3.978
DC/C	0.121	0.121	0.142	0.109	0.323	0.095	0.350	0.103	0.364	0.103
CP/C	0.158	0.120	0.120	0.097	0.102	0.087	0.205	0.128	0.304	0.148
T/S	1.138	0.392	1.078	0.138	1.160	0.178	1.123	0.124	1.106	0.109
CN/C	1.157	0.394	0.361	0.196	0.517	0.211	0.838	0.260	1.157	0.394

5. Discussion

This section addresses the two research questions on the basis of the above findings.

5.1. The changes of traditional SC measures

Compared with other traditional SC measures, three measures based on length (MLC, MLS, and MLT) are likely better indexes of the writing proficiency of L2 learners' narrative writing, because they have larger effect sizes, and can differentiate two pairs of adjacent writing proficiency levels. Both the MLS and the MLT can distinguish waystage users and threshold users, as well as the two levels of independent users. This finding is different from Jiang et al. (2019) research results regarding the MLS and the MLT, in which the MLS and the MLT can significantly differentiate adjacent proficiency levels of beginners and intermediate learners. In Jiang et al. (2019) study, they split each run-on sentence into two or more independent sentences in order to reduce effects of run-on sentences on the values of traditional SC measures. By doing this, the values of MLS and MLT decreased more in beginners' compositions than those at higher levels, which could cause significant differences between adjacent proficiency levels of beginners. The MLS and the MLT were likely not as applicable to beginners due to the effect of run-on sentences.

Despite being used as a length-based measure, the MLC was different than the MLS and the MLT. The MLC could significantly distinguish between the two levels of basic users (A1 & A2) and the two levels of independent users (B1 & B2). However, the MLC could not discriminate significantly between waystage users and threshold users or vantage users and proficient users. The different performances of MLC compared to MLS and MLT, is likely because the MLC measures the SC at the clause-level complexity level, which primarily assesses the use of phrases in clauses (Alexopoulou, Michel, Murakami, & Meurers, 2017; Bulté & Housen, 2012). Comparatively, the MLT measures sentence-level complexity (Jiang et al., 2019) that primarily assesses the complexity of clauses and phrases (Kyle & Crossley, 2018). In contrast, the MLS serves as a holistic complexity measure on the sentence-level (Jiang et al., 2019), assessing all syntactic structures that lead to sentence lengthening, including phrases, coordinations, and subordinations. The MLS and

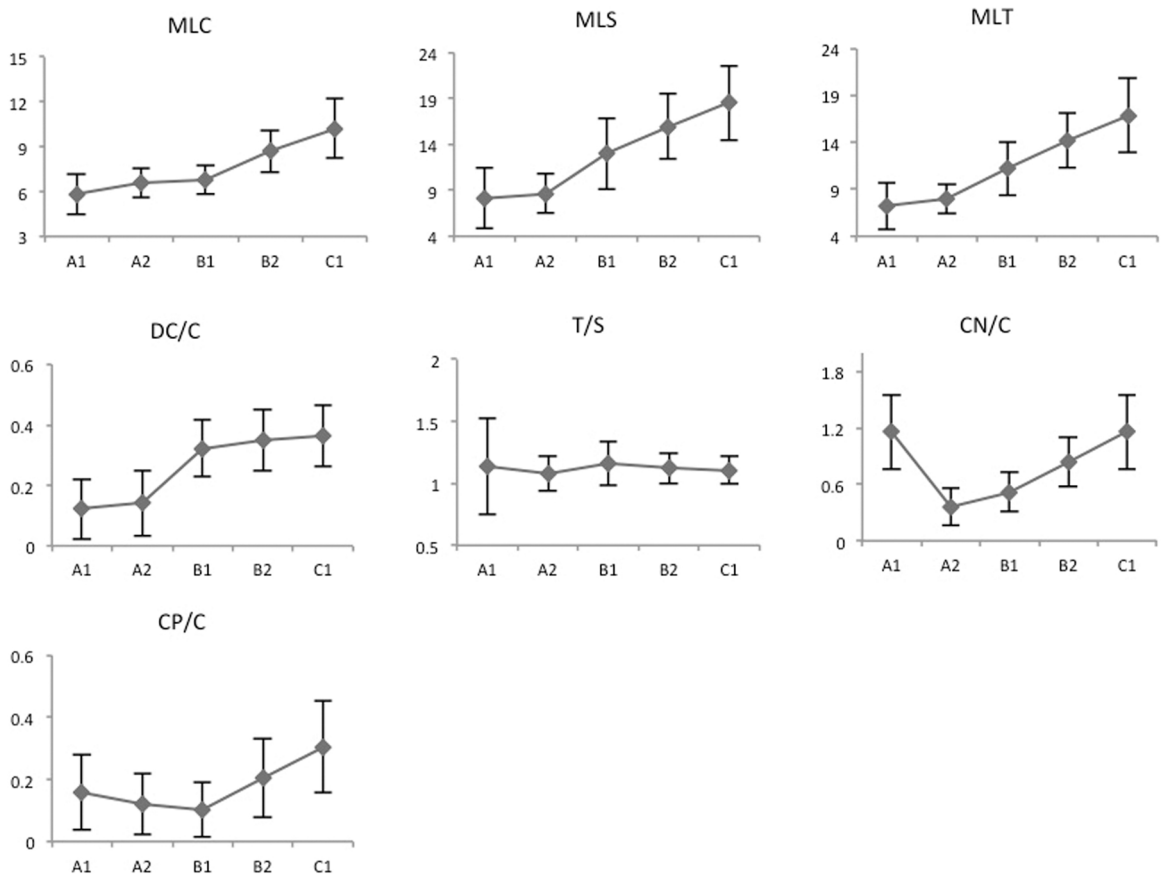

Fig. 5. Development trends of seven traditional SC measures.

Table 7
Follow-up paired-comparisons on traditional SC measures.

	MLC	MLS	MLT	DC/C
A1-A2	0.040*	1.000	0.565	1.000
A2-B1	1.000	0.000**	0.000**	0.000**
B1-B2	0.000**	0.004*	0.001*	1.000
B2-C1	0.068	0.397	0.462	1.000

* $p<0.05$.
** $p<0.001$.

Table 8
Descriptive statistics for four MDD measures.

MDD Measures	A1		A2		B1		B2		C1	
	M	SD	M	SD	M	SD	M	SD	M	SD
OMDD	1.687	0.227	1.917	0.216	2.122	0.216	2.309	0.192	2.421	0.216
SMDD	4.636	2.280	4.732	1.761	5.481	1.212	6.515	2.031	7.298	2.158
CMDD	3.646	1.686	4.361	1.873	5.706	2.343	5.516	1.63	5.317	1.513
NMDD	1.190	0.177	1.212	0.155	1.312	0.115	1.460	0.174	1.503	0.183

the MLT are therefore susceptible to the handling of run-on sentences, while the MLC is not. The MLC could not significantly distinguish A2 and B1 levels, which may be related to the change in complexity at the phrase level when A2 level rises to B1 level. As illustrated in Fig. 5 and Table 5, when the writing proficiency of L2 learners' narrative writing develops from A2 to B1, the coordinate-phrase-level complexity measure CP/C drops sharply, which may cause the MLC to only slightly increase from A2 to B1; that is, the MLC cannot significantly distinguish between A2 and B1 levels.

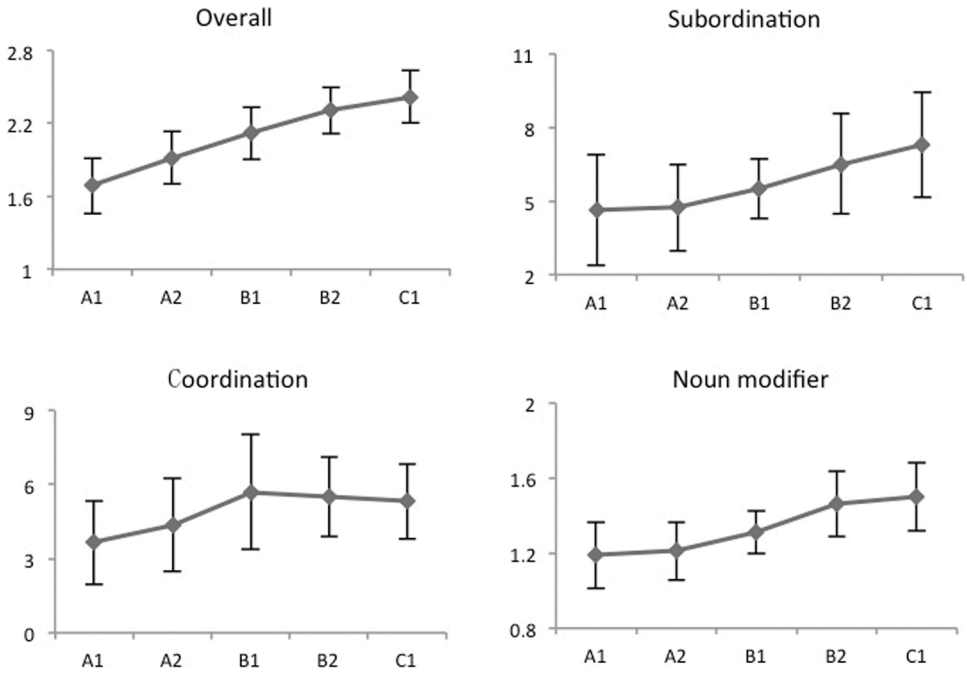

Fig. 6. Development trends of the MDD measures.

Table 9
Post hoc test on the MDD measures.

	OMDD	SMDD	NMDD
A1-A2	0.000**	1.000	1.000
A2-B1	0.000**	0.126	0.001*
B1-B2	0.000**	0.005*	0.000**
B2-C1	0.019*	1.000	1.000

* $p<0.05$.
** $p<0.001$.

In addition, our research found that none of the three length-based measures could distinguish between higher levels (B2 and C1). Paquot (2018) research reached a similar conclusion that length-based measures MLC, MLS, and MLT could not adequately discriminate between high-level and higher-level learners (B2, C1, and C2).

As for the clause-related measure, DC/C continues to increase with the improvement of writing proficiency in L2 narrative writing and can significantly differentiate between A2 and B1 writing levels. This result is consistent with Khushik and Huhta (2019) finding that measures based on subordination can more effectively measure the SC development of beginners and intermediate learners.

5.2. The changes of the MDD measures

In order to better characterize the development of L2 SC and verify the practicability of DD measures to assess the proficiency levels of L2 learners' narrative writings, we combined DD and traditional SC measures to assess the development of SC in L2 learners' narrative writing in groups.

Among the traditional SC measures and MDD measures used in our study, the overall MDD served as the best metric for measuring writing proficiency of L2 beginner, intermediate, and advanced learners' narrative writing. By manifesting the development of holistic SC in learners' narrative writing, it could significantly differentiate all pairs of adjacent proficiency levels. By contrast, the MLS, even though it is a holistic complexity measure (Jiang at al., 2019) and the best traditional SC measures in our study, it could not significantly discriminate between A1 and A2 levels and B2 and C1 levels. As we demonstrated in Section 2.3, the MDD and the MLS are differently influenced by run-on sentences, making the OMDD a more valid measure for beginners. In addition, as learners' writing

proficiency improved to the advanced level, the OMDDs kept increasing even when sentence length stayed stable. To convey complex meanings,[3] intermediate learners began using more subordination (Biber et al., 2011; Jiang et al., 2019), such as attributive clauses (sentence 1 in Fig. 7), and advanced learners employed complex sentences with participles of verbs (Biber et al., 2011), as seen in sentence 2 in Fig. 7. The dependency analysis of these two sentences is presented in Fig. 7. In sentence 1, there are 8 dependency relations, and the dependency distances are 1, 3, 2, 1, 3, 2, 1, and 1, respectively. In sentence 2, there are 7 dependency relations and the dependency distances are 3, 2, 1, 6, 2, 1, and 1, respectively. By using the formula in Table 5, we obtained the OMDDs of sentence 1 and 2; 1.75 and 2.29, respectively. Although the length of sentence 1 (9) is larger than that of sentence 2 (8), the OMDD of sentence 2 (2.29) is higher than that of sentence 1 (1.75). Thus, compared with the MLS, the OMDD is better at distinguishing between the intermediate and advanced learners.

As discussed above, the subordination-related measure DC/C rises slowly from the B1 to C1 levels, as advanced learners tend to use noun phrases instead of subordinations to increase the complexity of the sentence. In addition, the SMDDs of subordinations demonstrate that while the number of subordinations increases slowly, their complexity continues to increase. The governors and the dependents of the adverbial clauses are in bold in Excerpts 1 and 2.[4] The values of dependency distances equal the linear distance between the dependent and the governor in a dependency relation. Therefore, the dependency distance of the adverbial clause in B1 is 5, and the dependency distance of the adverbial clause in C1 is 14. Though there is only one adverbial clause in both two sentences, the second one is much more complex than the first one.

"At first I didn't **take** it serious, until he **called** the waiter and asked for his food." (Excerpt 1 from B1 level).

"Besides, it **was** the first time I saw my mother making Zongzi, since she had never **mentioned** that she could do it before." (Excerpt 2 from C1 level).

The L2 measure concerning complex nominal, CN/C experienced a dramatic decrease from A1 to A2, though it can significantly discriminate three pairs of adjacent writing proficiency levels (A2-C1) in the narrative writing of L2 learners. In comparison, the NMDD continues to increase along with the learners' writing proficiency, but it cannot significantly distinguish intermediate (B2) and advanced (C1) learners. Combining the CN/C and the NMDD could provide a more comprehensive view of how learners use complex noun phrases at different learning stages. In the following two excerpts, noun modifiers are in italics. The value of NMDD in Excerpt 3 is 1, while the value of NMDD in Excerpt 4 is 1.6. Thus, not only the number of noun modifiers, reflected by CN/C, but also the complexity of noun modifiers at the C1 level, reflected by the NMDD, is higher than those at the A1 level.

"Hi, *My name* is Jone… Than I eat *my breakfast*. After that, I brush *my teeth* and go to school at seven ten…. I get home from school at five and do *my homework*…." (Excerpt 3 from A1 level).

"For me, *the biggest pleasure of life* is going shopping and having *delicious food* with *my dear families*, especially after entering *the university where I spent tough and stressful days* as *a freshman*…" (Excerpt 4 from C1 level).

5.3. General discussion

Traditional quantity-based measures (such as CN/C, CP/C, DC/C) quantify the amount of certain syntactic structures and help reveal which syntactic structures are helpful for the construction of SC at different L2 learning stages. However, one type of syntactic structures, which are of the same amount at different learning stages, can have varying degrees of complexity. Therefore, not only the number, but also the complexity of target structures (corresponding MDD measures) should be taken into consideration when measuring the syntactic complexity. In addition, traditional quantity-based measures seem to be only suitable for measuring L2 learners' writing proficiency of narrative writing at certain learning stages. For example, DC/C can significantly distinguish between beginners and intermediate learners, but loses its effectiveness for intermediate and advanced learners. Traditional length-based measures (MLS, MLT, MLC) perform better than traditional quantity-based measures when differentiating different writing proficiency. The OMDD performs best in distinguishing different writing proficiency among all traditional and MDD measures. As we elaborated in the Section 2.2, dependency distance has been established as an important index of memory burden (Liu et al., 2017) and the linguistic complexity of sentence processing mechanisms (Gibson, 1998). Due to the close relationship between the sentence length and the dependency distance, longer sentences are supposed to have larger dependency distances and also larger MDDs (Jiang & Liu, 2015).

Therefore, from the deep perspective of sentence processing mechanism, the MDDs help explain why the traditional length-based measures, such as MLS can perform well in distinguishing writing proficiency levels in L2 learners' narrative writing. In addition, SC refers to how costly or difficult it is for language learners to process syntactic structures (Bulté & Housen, 2012). Syntactic structures with longer dependency distances are more cognitively demanding. Therefore, having found that the longer MDD indicates higher syntactic complexity, the current study further supports previous research findings (see Slavkov, 2015), namely that the longer-distance dependency relations demand more cognitive effort in L2.

Nevertheless, neither the OMDD nor the traditional length-based measures can reflect which specific syntactic structures constitute the SC at different L2 learning stages. Thus, only by combining the development patterns of traditional SC measures (both length-based

[3] In this study's data, the learners began to use the complex sentences with attributive clauses from the B1 level (Normalized Frequency: 0.53), and they started to employ the complex sentences with participles of verbs at the C1 level (Normalized Frequency: 0.39). The frequency of the complex sentences was normalized to 100 words to eliminate the influence of text length on frequency values.

[4] These excerpts are from our corpus. Because the language mistakes in the excerpts do not influence our analysis, they were not marked or corrected.

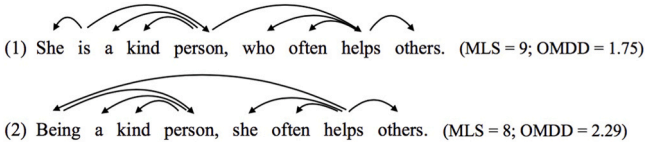

Fig. 7. Dependency analysis of two sentences.

and quantity-based) with MDD measures, can the development of the SC in learners' narrative writing be represented more clearly and comprehensively.

6. Conclusion

By employing traditional syntactic complexity measures and dependency distance measures, we have explored the syntactic complexity in a dependency syntactically-annotated corpus of 400 narrative compositions written by Chinese English learners across five writing proficiencies. The results suggest that among all traditional measures of syntactic complexity, length-based measures (mean length of clause, mean length of sentences, mean length of T-unit) perform best in differentiating the writing proficiency of beginner, intermediate, and advanced learners. However, none of these measures can significantly differentiate all adjacent writing proficiency levels. In terms of dependency distance measures, the overall mean dependency distance can significantly discriminate all pairs of adjacent proficiency levels. This makes it the best metric to assess the writing proficiency levels of beginner, intermediate and advanced learners Moreover, the MDD measures further explain the findings of the traditional SC measures from the perspective of language processing.

Our study serves as a pilot study for the combination of traditional SC measures and dependency distance measures in gauging the writing proficiency of L2 learners. Indeed, this study reveals that dependency distance not only measures the syntactic complexity of L1writers, but also those of L2. This finding supports the idea that longer dependency relations demand more cognitive effort in L2 (Slavkov, 2015). The findings of this study can also offer new methods and measures for large-scale writing assessments. Lastly, writing instructors could use this research to design exercises that more effectively help L2 learners practice using long-distance dependency relations and help learners pay attention to the language mistakes caused by long-distance dependency relations.

At the same time, however, it is important to acknowledge some of the limitations of this study, some of which may be addressed through future research. Firstly, L2 learners from different language backgrounds should be taken into consideration in order to test the practicability of dependency distance in future studies. In addition, most compositions in our study are in the form of narrative writing, which may affect the generalizability of some of the findings.

Funding

This work is supported by the National Social Science Foundation of China [grant number 17AYY021].

Declare of Conflicting Interest

The authors declare that there is no conflict of interest.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Appendix A. Illustrating the calculation processes of the SMDDs, CMDDs, and NMDDs

To illustrate how to calculate the SMDDs, CMDDs, and NMDDs of a composition or a corpus, we took a student's composition as an example.

Student 0402 (B1 level).

Everyone will meet *annoying things* in *our daily life*. Today I want to share *an annoying thing* that *happened* on me.

On *the last Friday*, we had *an English test*. It is important for *every classmate*. When we were **doing** *the paper*, *my best friend* Jack **asked** me for *the answer*. I **said** no to him because I do not **think** it **is** good for *each other*.

After *the test*, Jack didn't want to say anything to me. He seemed to be quiet and unhappy. But I didn't do anything wrong. And then he told someone *the thing* that *happened* in class. It made me angry. I won't **forgive** him, unless he **comes** to me and say sorry to me. That's all. Do you **think** I **did** wrong in *this thing*?

As shown above, the subordinations are in bold, the coordinations are underlined, and the noun modifiers are in italics. Only the dependent and the governor of the dependency relations (subordinations, coordinations and noun modifiers) are marked in order to clearly see the dependency distances of these dependency relations.

The values of dependency distances equal the linear distance between the dependent and the governor in a dependency relation.

The values of dependency distances are embedded in brackets. There are 5 subordinations: When we were doing the paper, my best friend Jack asked me…(7); I **said** no to him because I do not **think**…(8); I do not **think** it **is** good…(2); I won't **forgive** him, unless he **comes** to me…(4); Do you **think** I **did** wrong…(2). By using the formula in the Table 5, we can obtain the value of the SMDD, that is 4.6 ((7 +8 +2 +4 +2)/5). With the help of the statistical tools, such as Excel (Fig. 3), the mean values of certain dependency relations can be conveniently calculated. In the same way, the values of the CMDD and NMDD of this composition can also be calculated.

There are 2 coordinations: quiet and unhappy (2); comes to me and say sorry to me (4). The CMDD is 3 ((2 +4)/2).

There are 20 noun modifiers: *annoying things* (1); *our* daily *life* (2); *daily life* (1); *an* annoying *thing* (2); *annoying thing* (1); *thing* that *happened* (2); *the* last *Friday* (2); *last Friday* (1); *an* English *test* (2); *English test* (1); *every classmate* (1); *the paper* (1); *my* best *friend* (2); *best friend* (1); *the answer* (1); *each other* (1); *the test* (1); *the thing* (1); *thing* that *happened* (2); *this thing* (1). The NMDD is 1.35 ((1 +2 +1 +2 +1 +2 +2 +1 +2 +1 +1 +1 +2 +1 +1 +1 +1 +1 +2 +1)/20).

References

Ai, H., & Lu, X. (2013). A corpus-based comparison of syntactic complexity in nns and ns university students' writing. In A. DïAz-Negrillo, N. Ballier, & P. Thompson (Eds.), *Automatic treatment and analysis of learner corpus data* (pp. 249–264). Amsterdam: John Benjamins.

Alexopoulou, T., Michel, M., Murakami, A., & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning, 67*(S1), 180–208. https://doi.org/10.1111/lang.12232

Bardovi-Harlig, K. (1992). A second look at T-unit analysis: Reconsidering the sentence. *TESOL Quarterly, 26*(2), 390–395.

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly, 45*(1), 5–35. https://doi.org/10.5054/tq.2011.244483

Biber, D., Gray, B., Staples, S., & Egbert, J. (2020). Investigating grammatical complexity in L2 English writing research: Linguistic description versus predictive measurement. *Journal of English for Academic Purposes, 46*, Article 100869. https://doi.org/10.1016/j.jeap.2020.100869

Bulté, B., & Housen, A. (2012). Defining and operationalizing l2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 Performance and Proficiency: Complexity Accuracy and Fluency in SLA* (pp. 21–46). Amsterdam: John Benjamins.

Cambridge University Press. (2019). The CEFR Grids for Writing, developed by ALTE members.

Chen, D., Manning, C.D. (2014). A fast and accurate dependency parser using neural networks. In Y. Marton (Ed.), The 2014 Conference on Empirical Methods in Natural Language Processing (pp. 740–750). Stroudsburg, PA: Association for computational linguistics.

Council of Europe. (2018). Common European Framework of Reference for Languages: Learning, Teaching, Assessment.

De Clercq, B., & Housen, A. (2017). A cross-linguistic perspective on syntactic complexity in L2 development: Syntactic elaboration and diversity. *The Modern Language Journal, 101*(2), 315–334. https://doi.org/10.1111/modl.12396

Egbert, J. (2017). Corpus linguistics and language testing: Navigating uncharted waters. *Language Testing, 34*, 555–564.

Fang, Yu., & Liu, H. (2018). What factors are associated with dependency distances to ensure easy comprehension? A case study of ba sentences in mandarin Chinese. *Language Sciences, 67*, 33–45. https://doi.org/10.1016/j.langsci.2018.04.005

Fedorenko, E., Woodbury, R., & Gibson, E. (2013). Direct evidence of memory retrieval as a source of difficulty in non-local dependencies in language. *Cognitive Science, 37*(2), 378–394. https://doi.org/10.1111/cogs.12021

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology General, 141*(1), 2–18.

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences of the United States of America, 112*(33), 10336–10341. https://doi.org/10.1073/pnas.1502134112

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition, 68*(1), 1–76. https://doi.org/10.1016/S0010-0277(98)00034-1

Gibson, E., Ko, K. (1998). An integration-based theory of computational resources in sentence comprehension. Paper presented at the Fourth Architectures and Mechanisms in Language Processing Conference, University of Freiburg, Germany.

Gildea, D., & Temperley, D. (2010). Do grammars minimize dependency length? *Cognitive Science, 34*(2), 286–310. https://doi.org/10.1111/j.1551-6709.2009.01073.x

Heringer, H., Bruno, S., & Rainer, W. (1980). *Syntax: Fragen, Lösungen, Alternativen [Syntax: Issues, Solutions, Alternatives]*. Munich: Wilhelm Fink Verlag,.

Housen, A., De Clercq, B., Kuiken, F., & Vedder, I. (2019). Multiple approaches to complexity in second language research. *Second Language Research, 35*(1), 3–21. https://doi.org/10.1177/0267658318809765

Hudson, R. (1990). *An English word grammar*. Oxford: Basil Blackwell.

Hudson, R. (1995). Measuring syntactic difficulty. Unpublished paper. Retrieved October 4, 2016 from ⟨http://dickhudson.com/wp-content/uploads/2013/07/Difficulty.pdf⟩.

Hudson, R. (2007). *Language networks: The new word grammar*. Oxford: Oxford University Press.

Hudson, R. (2010). *An introduction to word grammar*. Cambridge: Cambridge University Press.

Ishikawa, S. (1995). Objective measurement of low-proficiency EFL narrative writing. *Journal of Second Language Writing, 4*(1), 51–69.

Jiang, J., Bi, P., & Liu, H. (2019). Syntactic complexity development in the writings of EFL learners: Insights from a dependency syntactically-annotated corpus. *Journal of Second Language Writing, 46*(100666). https://doi.org/10.1016/j.jslw.2019.100666

Jiang, J., & Liu, H. (2015). The effects of sentence length on dependency distance, dependency direction and the implications-based on a parallel English–Chinese dependency treebank. *Language Sciences, 50*, 93–104. https://doi.org/10.1016/j.langsci.2015.04.002

Jiang, J., & Ouyang, J. (2017). Dependency distance: A new perspective on the syntactic development in second language acquisition. *Physics of Life Reviews, 21*, 209–210. https://doi.org/10.1016/j.plrev.2017.06.018

Jiang, J., & Ouyang, J. (2018). Minimization and probability distribution of dependency distance in the process of second language acquisition. In J. Jiang, & H. Liu (Eds.), *Quantitative Analysis of Dependency Structures* (pp. 167–190). De Gruyter Mouton.

Jiang, J., Ouyang, J., & Liu, H. (2019). Interlanguage: A perspective of quantitative linguistic typology. *Language Sciences, 74*, 85–97. https://doi.org/10.1016/j.langsci.2019.04.004

Khushik, G. A., & Huhta, A. (2019). Investigating syntactic complexity in EFL learners' writing across common European framework of reference levels A1, A2, and B1. *Applied Linguistics*. https://doi.org/10.1093/applin/amy064

Kyle, K., & Crossley, S. A. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing, 34*, 513–535. https://doi.org/10.1177/0265532217712554

Kyle, K., & Crossley, S. A. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *Modern Language Journal, 102*(2), 333–349. https://doi.org/10.1111/modl.12468

Levy, R., & Keller, F. (2013). Expectation and locality effects in German verb-final structures. *Journal of Memory and Language, 68*(2), 199–222. https://doi.org/10.1016/j.jml.2012.02.005

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

Liu, H. (2009). Using a Chinese treebank to measure dependency distance. *Corpus Linguistics and Linguistic Theory, 5*(2), 161–174. https://doi.org/10.1515/CLLT.2009.007
Liu, H., Xu, C., & Liang, J. (2017). Dependency distance: a new perspective on syntactic patterns in natural languages. *Physics of Life Reviews, 21*, 171–193. https://doi.org/10.1016/j.plrev.2017.03.002
Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*, 474–496. https://doi.org/10.1075/ijcl.15.4.02lu
Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly, 45*, 36–62. https://doi.org/10.5054/tq.2011.240859
Lu, X. (2017). Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. *Language Testing, 34*(4), 493–511. https://doi.org/10.1177/0265532217710675
Lu, X., & Ai, H. (2015). Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing, 29*, 16–27. https://doi.org/10.1016/j.jslw.2015.06.003
Martinez, A. C. (2018). Analysis of syntactic complexity in secondary education EFL writers at different proficiency levels. *Assessing Writing, 35*, 1–11. https://doi.org/10.1016/j.asw.2017.11.002
Meurers, D., & Dickinson, M. (2017). Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics. *Language Learning, 67*(S1), 66–95.
Nivre, J. (2006). *Inductive dependency parsing*. Dordrecht: Spring.
Norris, J., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*, 555–578. https://doi.org/10.1093/applin/amp044
Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics, 24*, 492–518.
Paquot, M. (2018). Phraseological competence: A missing component in university entrance language tests? Insights from a study of EFL learners' use of statistical collocations. *Language Assessment Quarterly, 15*(1), 29–43. https://doi.org/10.1080/15434303.2017.1405421
Parkinson, J., & Musgrave, J. (2014). Development of noun phrase complexity in the writing of English for Academic Purposes students. *Journal of English for Academic Purposes, 14*, 48–59. https://doi.org/10.1016/j.jeap.2013.12.001
Rescher, N. (1998). *Complexity: A philosophical overview*. New Brunswick: Transaction Publishers.
Robinson, P. (2011). Second language task complexity, the cognition hypothesis, language learning, and performance. In P. Robinson (Ed.), *Second language task complexity: Researching the cognition hypothesis of language learning and performance* (pp. 3–38). Amsterdam: John Benjamins.
Robinson, P., & Gilabert, R. (2007). Task complexity, the cognition hypothesis and second language learning and performance. *IRAL-International Review of Applied Linguistics in Language Teaching, 45*(3), 161–176. https://doi.org/10.1515/iral.2007.007
Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. Proceedings of the international conference on new methods in language processing. ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger1.pdf.
Slavkov, N. (2015). Long-distance wh-movement and long-distance wh-movement avoidance in L2 English: Evidence from French and Bulgarian speakers. *Second Language Research, 31*(2), 211–237.
Taguchi, N., Crawford, W., & Wetzel, D. Z. (2013). What linguistic features are indicative of writing quality? A case of argumentative essays in a college composition program. *TESOL Quarterly, 47*(2), 420–430. https://doi.org/10.1002/tesq.91
Temperley, D., & Gildea, D. (2018). Minimizing syntactic dependency lengths: Typological/cognitive universal? *Annual Review of Linguistics, 4*(1), 67–80. https://doi.org/10.1146/annurev-linguistics-011817-045617
Tesnière, L. (1959). *Eléments de la syntaxe structurale [Elements of Structural Syntax]*. Paris: Klincksieck.
Verspoor, M., Lowie, W., Chan, H. P., & Vahtrick, L. (2017). Linguistic complexity in second language development: Variability and variation at advanced stages. *Recherches Eno¯sis didactique des langues et des cultures, 14*, 1–27.
Vyatkina, N. (2013). Specific syntactic complexity: Developmental profiling of individuals based on an annotated learner corpus. *The Modern Language Journal, 97*, 11–30. https://doi.org/10.1111/j.1540-4781.2012.01421.x
Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity*. Honolulu: University of Hawaii, Second Language Teaching & Curriculum Center.
Yang, W., Lu, X., & Weigle, S. A. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing, 28*, 53–67. https://doi.org/10.1016/j.jslw.2015.02.002

Jinghui Ouyang is an Assistant Professor in linguistics and applied linguistics at Tongji University. Her research interests include applied linguistics, dependency syntax and quantitative linguistics. She is the author of about 10 scientific publications about language and linguistics.

Jingyang Jiang is a Professor of Linguistics and Applied Linguistics at Zhejiang University. Her research interests include applied linguistics, dependency syntax and quantitative linguistics. She has published about 40 research papers in influential journals, is the author of two monographs and the editor-in-chief of one book in a book series published by De Gruyter.

Haitao Liu is a Qiushi distinguished professor of linguistics and applied linguistics at Zhejiang University. His research interests include dependency syntax, quantitative linguistics and applied linguistics. He is the author of more than 200 scientific publications about language and linguistics.