Journal of Second Language Writing 46 (2019) 100666

Contents lists available at ScienceDirect

Journal of Second Language Writing

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

Syntactic complexity development in the writings of EFL learners: Insights from a dependency syntactically-annotated corpus

Jingyang Jiang[a], Peng Bi[a], Haitao Liu[a,b,c,*]

[a] Department of Linguistics, Zhejiang University, Hangzhou, China
[b] Institute of Quantitative Linguistics, Beijing Language and Culture University, Beijing, China
[c] Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, Guangzhou, China

ARTICLE INFO

Keywords:
Syntactic complexity
L2 English writing
Beginner and intermediate learner
Dependency syntactic annotation

ABSTRACT

This study investigates the syntactic complexity, as measured by both large-grained and fine-grained measures, of 410 narrative writings across four writing proficiency levels written by beginner and intermediate L2 English learners. By exploring the differences in syntactic complexity between writings at different proficiency levels, the study is purposed to find out the measures that can best discriminate and predict writing proficiency. The L2 Syntactic Complexity Analyzer and the dependency syntactically-annotated corpus are used respectively to collect the data for large-grained and fine-grained measures. With regard to large-grained measures, it is found that students with higher writing proficiency tend to produce longer language units, more subordinate clauses, more coordinate clauses, and more noun phrases in their writings; mean length of T-unit, mean length of sentence, and dependent clauses per clause can better predict writing proficiency than other traditional large-grained measures. As for fine-grained measures, it is found that three types of subordinate clauses, that is, adverbial clauses, complement clauses and relative clauses, and two types of noun modifiers, that is, prepositional phrases and adjectival relative clauses, occur more frequently in the writings of more proficient learners; the frequency of compound nouns correlates negatively with writing proficiency.

1. Introduction

Syntactic complexity (hereafter, SC) is held as an important construct to assess second language (L2) writing quality (De Clercq & Housen, 2017; Housen & Kuiken, 2009; Lu, 2011). To this end, over the past 40 years, substantial studies have attempted to establish effective and reliable SC "yardsticks" to capture the linguistic development in L2 writing and to evaluate learners' writing proficiency (Ansarifar, Shahriari, & Pishghadam, 2018).

In earlier studies, the most frequently used measures of SC include mean length of sentence (MLS), mean length of clause (MLC), and ratio of subordinate clauses (Hunt, 1965). However, compared with these measures, T-unit (the minimal terminable unit) length is proposed to be a more effective measure of syntactic maturity (Hunt, 1965, 1970). From this measure are derived some relevant measures of SC as indicators of the amount of subordination, such as clauses per T-unit (C/T). Lately, the use of noun phrases is regarded as a main SC characteristic of the writings of advanced learners (Biber, Gray, & Poonpon, 2011; Lu, 2011; Parkinson & Musgrave, 2014). Despite the rich variety of SC measures in the literature, most studies from the 1960s to the 2000s employed only one or two of these measures, mainly subordination-related and coordination-related measures (Bulté & Housen, 2012; Lu & Ai, 2015;

[*] Corresponding author.
E-mail address: lhtzju@yeah.net (H. Liu).

https://doi.org/10.1016/j.jslw.2019.100666
Received 12 February 2019; Received in revised form 15 August 2019; Accepted 26 August 2019
Available online 04 September 2019

Ortega, 2003). It is not until recently that researchers have reached the consensus that SC is a multi-dimensional construct, normally involving the following four dimensions: length of production unit, amount of subordination, amount of coordination, and degree of phrasal sophistication (Ai & Lu, 2013; Lu, 2017; Norris & Ortega, 2009). In other words, the measurement methods of SC in L2 writing studies are under constant improvement with various linguistic units being integrated into the measurement to unveil the whole picture of SC development. More importantly, four new trends seem to have emerged in the latest SC studies, that is, the emphasis on learners at lower levels of proficiency (Ortega, 2015), the utilization of more fine-grained measures (Kyle & Crossley, 2018), the growing application of annotated learner corpora (Vyatkina, Hirschmann, & Golcher, 2015), and the dynamic nature of SC development (Verspoor, Schmid, & Xu, 2012). However, very few studies venture to integrate all these four trends, which is what the present study attempts. Based on a syntactically annotated corpus with 410 compositions across four writing proficiency levels, we intend to explore to what extent the large-grained measures (length of language unit, amount of subordination, amount of coordination, and degree of phrasal sophistication) and the fine-grained measures (different types of subordinate clauses and noun modifiers) differ across the four levels with the purpose to identify the SC measures that can better gauge the writing proficiency of young high school EFL learners.

In the next section, we will provide a review of SC. Sections 3 and 4 will introduce the research questions and methods. Then the results and discussion will be presented in Sections 5 and 6. Finally, the conclusions will be drawn in Section 7.

2. Research background

This section will be devoted to a brief introduction of SC and the four latest trends in SC studies.

2.1. Syntactic complexity and its operationalization

The term complexity is polysemous in second language acquisition (SLA) studies (Pallotti, 2015), since it can refer to either cognitive complexity or linguistic complexity (Bulté & Housen, 2012). The former is the cognitive cost or difficulty of learning and processing a linguistic construction (Bulté & Housen, 2012), susceptible to multiple learner-related factors, such as aptitude, motivation and working memory. Constructions difficult for some learners may be easy for others. So, cognitive complexity is a concept somewhat relative and subjective. In comparison, linguistic complexity (also referred to as structural complexity) is more objective, defined "as the number of discrete components that a language feature or a language system consists of, and as the number of connections between the different components" (Bulté & Housen, 2012, p. 24). The confusion of these two complexities would result in circular reasoning (for more details, see Housen & Simoens, 2016). The loose definition of complexity would also render it difficult to compare the results of different studies. To avoid these problems, this study adheres to the taxonomy of L2 complexity proposed by Bulté and Housen (2012), measuring SC in objective and quantitative terms, such as the length of production unit and the amount of different syntactic components.

2.2. The emphasis on young beginner and intermediate EFL learners

Recently, it has been noted that different levels of writing proficiency may lead L2 learners to employ different grammatical devices to construct complex structures (Ryshina-Pankova, 2015; Wolfe-Quintero, Inagaki, & Kim, 1998). In the past decades, SC studies have overwhelmingly focused on advanced learners, neglecting less proficient learners. However, the SC measures applicable to advanced learners may not be suitable and even irrelevant for beginner and intermediate learners (Ishikawa, 1995; Verspoor, Lowie, Chan, & Vahtrick, 2017). Thus, it is worthwhile to conduct some studies that pursue "empirical descriptions for the best matching of targeted areas for complexification with relevant proficiency levels" (Ortega, 2015, p. 90). Only by taking into account beginner and intermediate learners can we come to understand what types of measures should be adopted to assess the writing development of learners with distinct proficiency levels. It should be acknowledged that there have already existed several SC studies targeting at lower-level learners, particularly for the acquisition of an L2 other than English, but most of the participants are adult learners (e.g. Jiang, 2013 on L2 Chinese; Spoelman & Verspoor, 2010 on L2 Finnish; Vyatkina et al., 2015 on L2 German). However, since young students constitute the majority of beginner and intermediate L2 English learners, who are cognitively and affectively different from adult learners (Michel, Kormos, Brunfaut, & Ratajczak, 2019), it might be of more significance to study the SC development in the productions of young L2 English learners with low and intermediate levels. In recent years, relevant studies have begun to emerge. Verspoor et al. (2012) conducted a comparative investigation into the writings of Dutch-speaking learners of English from two secondary grades (the first year and the third year of secondary education), so as to explore the changes in 64 syntactic and lexical complexity variables. Their results suggested that the total number of dependent clauses and the proportion of present tense were two effective discriminators of language proficiency. Similarly, another study probed into the writings of EFL learners from two grades (the third year and the fourth year of secondary education), and found that 13 out of 14 SC measures, with the exception of compound-complex sentence ratio, could differentiate the two groups of writers (Lahuerta Martínez, 2018). These studies are significant in the sense that they have made some attempts to find the objective SC measures to gauge the writing development of young beginner and intermediate L2 learners. However, in a general sense, they are still understudied. Moreover, the relevant studies only target at students from a few grades, that is, the temporal interval between the grades is only two or three years. Nonetheless, most young ESL or EFL learners might learn English for approximately six years in the K-12 context. Thus, the present study is designed to explore the SC development in the writings of young L2 English learners across a wider time span.

2.3. The adoption of more fine-grained measures

The second trend is the adoption of more fine-grained SC measures. Biber et al. (2011) pioneering work has inaugurated a consensus that some large-grained measures, such as MLT, seem to mask and obscure the SC development: Grammatical devices at both the sentential and the clausal levels could result in changes of T-unit length (Kyle & Crossley, 2018). Accordingly, from the 2010s, quite a few fine-grained measures of noun modifiers have been introduced and employed to capture some of the instantiations of noun phrases (e.g. Kyle & Crossley, 2018; Parkinson & Musgrave, 2014). These studies found that there existed variations in the developmental patterns of different noun modifier types for university-level students. Similarly, subordination covers three distinct elaboration processes in English: complement clauses, adverbial clauses and relative clauses. Moreover, they surface at different stages of acquisition and relative clauses emerge later than complement and adverbial clauses (Schmid, Verspoor, & MacWhinney, 2011). Therefore, it is also hypothesized that the three types of subordinate clauses do not develop synchronously. Put differently, the granularity of subordination, which is neglected in the existing studies, also deserves equal attention as noun modifiers. As a complement to the widely-used large-grained measures, more fine-grained measures under the dimension of noun modifiers and subordinate clauses should be incorporated into the current studies in order to see which kinds of detailed constructions seem to develop at each development level. By so doing, the development of large-grained measures would be better interpreted (Kyle & Crossley, 2018).

2.4. The annotated corpora of learners

The need for fine-grained measures (the second trend) has partly led to the proliferation of studies based on annotated corpora, namely the third trend, for we are not able to search specific syntactic structures such as complement clauses accurately without an annotated corpus (Gablasova, Brezina, McEnery, & Boyd, 2017). Syntactic annotation enables us to identify the target structures systematically and efficiently (Meurers & Dickinson, 2017). With the development of Natural Language Processing (NLP) technology, many syntactic annotation systems are available today, such as the Stanford dependency parser (Chen & Manning, 2014), the Tree tagger (Schmid, 1994) and the Biber tagger (Biber, 1988). All these systems are virtually intended for the annotation of first language (L1). Their reliability appears to be problematic when it comes to annotating linguistic materials from L2 learners. It has been pointed out that the accuracy rate of the Tree tagger "drops dramatically while tagging learner data" (Vyatkina et al., 2015, p. 34). To cope with this problem, Vyatkina et al. (2015) adopted semi-automatic annotation, that is, the automatic annotation plus manual-revisions. However, most studies lack the procedure of manual-revisions, relying exclusively on automatic annotation (e.g. Kyle & Crossley, 2018), which may considerably compromise the reliability of L2 annotation.

2.5. The dynamic nature of SC development

Language is a complex adaptive system and language learning is a dynamic process (Bulté & Housen, 2018; Larsen-Freeman, 2006; Verspoor et al., 2012). Due to the self-organization and the interaction effects of different subsystems, variability is an innate property of language development (Spoelman & Verspoor, 2010; Verspoor, Lowie, & van Dijk, 2008). The development of SC is no exception: The findings of a series of empirical studies under a Dynamic System Theory (DST) have pointed to the fact that the development of SC measures is characterized by more or less variability (Bulté & Housen, 2018; Spoelman & Verspoor, 2010; Verspoor et al., 2012; Vyatkina et al., 2015). Furthermore, since L2 development is a dynamic process, "at different moments in the developmental process (at different proficiency levels) the very make-up of the learners L2 interlanguage system is different" (Verspoor et al., 2017, p. 20-21). In other words, as learners go through stages of SLA, their syntactic repertoire also presents dynamic changes, which may be captured by different SC measures (the first trend).

The above are the four new trends in SC studies. However, there are limitations of the studies that have ushered in these trends. For example, the studies on the SC development of young learners take into account students from only two grades and the temporal interval is quite short, hardly covering the full development period from beginning to intermediate levels. Therefore, the question remains largely unanswered which large-grained SC measures can effectively assess the L2 writing development of beginner and intermediate learners. In addition, fine-grained measures, such as different types of subordinate clauses, have not received due attention. To the best of our knowledge, few studies have delved into the developmental patterns of the three types of subordinate clauses in learners' writings (for an exception, see Vyatkina et al., 2015). Instead, most studies primarily addressed large-grained measures like MLT and C/T. Another limitation lies in annotation, which is largely automatically implemented, compromising the reliability, especially in the case of learner language.

3. Current study

Taking into account these limitations, this research attempts to probe into the SC of high-school students' narrative writings spanning six grades by adopting both large-grained measures under four dimensions (length of production unit, amount of subordination, amount of coordination, and degree of phrasal sophistication), and fine-grained measures under two dimensions (different types of subordinate clauses and noun modifiers). The writing proficiency of learners is operationalized in terms of their writing scores, on the basis of which, they are grouped into four levels. The L2 Syntactic Complexity Analyzer (L2SCA) (Lu, 2010) and the dependency syntactically-annotated corpus have been utilized as instruments to collect respectively the data of large-grained and fine-grained measures. To note, our corpus is annotated with a combination of machine parsing, and manual checking and revisions

Table 1
The distribution of writing samples across four levels and six grades.

	G7	G8	G9	G10	G11	G12	Total
Level 1	50	26	27	0	0	0	103 (89)
Level 2	17	35	30	7	12	3	104 (122)
Level 3	1	5	21	31	24	20	102 (154)
Level 4	0	0	9	24	30	38	101 (168)
Total	68 (91)	66 (104)	87 (129)	62 (160)	66 (168)	61 (150)	410

Note. G7 stands for the first grade in junior high school and G12 the last grade in senior high school. The mean length of compositions in each level or grade is shown in the brackets.

under dependency grammar so that the target grammatical structures can be retrieved efficiently and accurately. The research questions are as follows:

> RQ1. To what extent could the large-grained measures discriminate among different levels of beginner and intermediate learners' writings?
>
> RQ2. To what extent could the fine-grained measures discriminate among different levels of beginner and intermediate learners' writings?

4. Research methods

4.1. Participants and material

The language material of this study consists of 410 compositions by students from three grades in one junior high school and three grades in one senior high school in Zhejiang province, eastern China. These compositions were collected in the spring semester of 2015. The age of participants, whose L1 is Chinese, ranges from 13 to 18. They were required to write a narrative essay with the prompt "a(n) happy/annoying/embarrassing thing or my last weekend." The composition must be written within 30 mins in the class without referring to any other written materials. They all understood that their compositions would be used for academic purposes, and permitted us to do so. Each received a present as a reward for participation.

In China, students normally begin English learning at the fourth grade in the primary school. During this stage, the goal is merely to acquaint them with alphabets and some daily dialogues. It is not until in the junior high school that they start to formally learn vocabulary and grammar. Thus, it is plausible to deem the first and second grade students in the junior high school as beginner learners. According to the English teaching syllabus for senior high schools in China, the English proficiency of students of higher grades in the senior high school should reach an intermediate level. Put simply, Chinese high school students are assumed to have low or intermediate English proficiency, that is, our participants are beginner and intermediate EFL learners.

To serve our research purposes, the 410 compositions are grouped into four levels of writing proficiency according to the four quartiles of their writing scores. The writing proficiency is operationalized as writing scores and the development of writing proficiency is conceptualized as the growth of scores (Yoon, 2018). The rating rubric is adapted from The Preliminary English Test of Cambridge (PET), which is an analytical scoring rubric, including four sub-scales: content, communicative achievement, organization, and language use. The total score is the sum of the four sub-scores. This rubric is chosen for two reasons. For one thing, PET is designed for learners below B1 level, whose English proficiency approximately matches that of our participants. For another, an analytical scoring rubric may better reflect the multiple facets of writing proficiency, including linguistic and discourse knowledge (Yoon, 2018).

Two experienced high school teachers scored all these compositions. The inter-rater reliability for all the four sub-scores, and the total scores between the two raters was strong, especially the language use ($r = .811$; $p = .000$), and the total scores ($r = .883$; $p = .000$). As described above, corresponding to the four quartiles of the total scores, the compositions are categorized into four writing proficiency levels. The distribution of writing samples across grades and writing proficiency levels is shown in Table 1.[1]

Table 2 presents the descriptive statistics of writing scores across four proficiency levels and six grades. One-way analysis of variance (ANOVA) test demonstrated significant differences in scores among different levels ($F = 1289.557$, $p = .000$, $\eta^2 = .905$). Additionally, the Bonferroni post hoc tests affirmed significant differences in scores between each pair of adjacent levels (levels 1-2: $p = .000$; levels 2-3: $p = .000$; levels 3-4: $p = .000$). Each dataset contains 103, 104, 102, and 101 compositions respectively. All the 410 compositions have a total of 54,472 word tokens. As shown in Table 1, the average length of compositions in each level is 89, 122, 154, and 168 respectively.

[1] It is observed that there exist big differences of writing scores and mean length of composition between G9 and G10. For one thing, the institution difference might be at work, because students of G9 and G10 come from two different schools. For another, students of G10 have just taken the high school entrance examination, and this could help them improve their English proficiency quickly.

Table 2
Descriptive statistics for writing scores across four levels and six grades.

	Writing proficiency levels				Grade levels					
	Level 1	Level 2	Level 3	Level 4	G7	G8	G9	G10	G11	G12
N	103	104	102	101	68	66	87	62	66	61
M	10.621	13.514	15.711	17.767	11.316	12.656	13.477	16.282	16.212	17.057
SD	.124	.071	.055	.075	1.734	1.734	2.517	1.240	1.829	1.469
Range	7–12	12.5–14.5	15–16.5	17–20	7–15.5	8–16	7.5–19	13.5–19.5	12–20	12.5–19.5

4.2. Instruments and measures

The instrument for coding traditional large-grained measures is L2SCA designed by Lu (2010), which can yield the data of 14 measures concerning length of production unit, amount of subordination, amount of coordination, and degree of phrasal sophistication. Some of these measures are redundant (Norris & Ortega, 2009), so from these 14 measures only seven are chosen for the present study: MLC, MLT, MLS, dependent clauses per clause (DC/C), T-units per sentence (T/S), coordinate phrases per clause (CP/C), and complex nominals per clause (CN/C)[2] (see Table 3). The definitions of these grammatical terms are available in Lu (2011), who, for instance, defines clauses as "structures with a subject and a finite verb, including independent, adjective, adverbial, and nominal clauses, but not nonfinite (including gerund, infinitive, and participle) verb phrases" (Lu, 2011, p. 44). What is noteworthy is that low-level L2 English learners tend to use run-on sentences or comma splices in their writings (Bardovi-Harlig, 1992). These mistakes are particularly severe in the compositions of Chinese students, because Chinese clauses can simply occur in juxtaposition without any coordinator or subordinator. Too many run-on sentences would increase the sentence length, but reduce the number of sentences, leading to inaccurate calculations of MLS and T/S. As a result, we revised the run-on sentences by splitting each of them into two or more independent sentences. For example, the run-on sentence: "We are friends, we often have a walk in the park" was divided into two independent sentences: "We are friends" and "We often have a walk in the park." Though L2SCA is reported to be able to yield highly reliable results for the coding of SC measures (Lu, 2010), it remains a question whether it would generate similarly reliable data from "not so-grammatically-correct" language materials. Thus, we hand-coded 40 randomly-selected scripts from our corpus, ten from each subset. Then, the same 40 scripts were automatically coded by L2SCA. Pearson correlation test showed high correlations between them (MLC: r = .835, p = .000; MLT: r = .971, p = .000; MLS: r = .968, p = .000; DC/C: r = .837, p = .000; T/S: r = .846, p = .000; CP/C: r = .986, p = .000; CN/C: r = .832, p = .000). These results prove that L2SCA also functions well for our data.

As has been mentioned, the fine-grained measures are extracted from the corpus syntactically annotated with dependency grammar (the dependency treebank), which defines syntax in terms of the binary and asymmetric dependency relations between two grammatical units, typically two words, one being the governor and the other the dependent (Hudson, 2010; Liu, 2008; Liu, Xu, & Liang, 2017; Tesnière, 1959). Fig. 1 shows the dependency structure of the sentence "He is a lazy student who always gets up late", in which are presented the grammatical function and the word class of each word. For example, nsubj stands for noun subjects and det for determiners. Limited space precludes a more detailed and extensive introduction to dependency grammar (for further information, refer to Hudson, 2010; Jiang & Liu, 2018; Liu et al., 2017). The key issue is that the annotation of dependency relations would facilitate the extraction of fine-grained measures for the current study.

The annotation is implemented through two procedures. During the first procedure, the scripts (after run-on sentence splitting) were parsed by the Stanford dependency parser 3.6.0 (Chen & Manning, 2014) and the results were saved as an excel spreadsheet. Then, two postgraduate students of applied linguistics checked and revised the automatic annotation according to an updated annotation manual of the parser, to which were added more specific sub-types of some original dependency relations, such as prep:attr (prepositions as attributives) and prep:adv (prepositions as adverbials). In order to ensure the accuracy and the consistency of manual checking and revisions, the two postgraduate students first checked and revised the same 130 machine-annotated compositions, about 1/3 of the whole. Then, together with two senior linguists (one specializing in dependency grammar and the other in SLA), they compared and discussed the revisions they made on the annotations of these 130 scripts and reached a consensus on the revisions. After this pilot trial, they went on to check and revise all the remaining annotations, about 1/3 for each of them. Cohen's kappa test indicated that the inter-annotator agreement score between the two postgraduates for the revision was acceptable (k = .789, p = .000). The finalized dependency treebank is also stored as a spreadsheet, as illustrated in Fig. 2.

Altogether, eight fine-grained measures under the dimension of subordinate clauses and noun modifiers are identified in our treebank (see Table 4). The frequency of each target structure was normalized to 100 words to eliminate the influence of text length on frequency values.

In sum, 15 SC measures are selected for this study, including seven large-grained measures (MLC, MLT, MLS, DC/C, T/S, CP/C,

[2] It is acknowledged that to some extent, MLC, MLT and MLS all capture the complexity of grammatical unit length. Thus, only one of them should be used to avoid the redundancy of measures. However, it is argued that MLC targets at the clausal-level complexity; MLT measures the sentential-level complexity; and MLS is a holistic SC index. From this perspective, the three measures conceptualize different aspects of SC. For this reason, all of the three have been adopted in this study.

Table 3
The seven traditional large-grained SC measures.
(modified from Lu, 2010, p. 479)

Dimension	Label	Definition	Calculation
Length of production unit	MLC	Mean length of clause	Number of words/Number of clauses
	MLT	Mean length of T-unit	Number of words/Number of T-units
	MLS	Mean length of sentence	Number of words/Number of sentences
Amount of subordination	DC/C	Dependent clauses per clause	Number of dependent clauses/Number of clauses
Amount of coordination	T/S	T-units per sentence	Number of T-units/Number of sentences
	CP/C	Coordinate phrases per clause	Number of coordinate phrases/Number of clauses
Degree of phrasal sophistication	CN/C	Complex nominals per clause	Number of complex nominals/Number of clauses

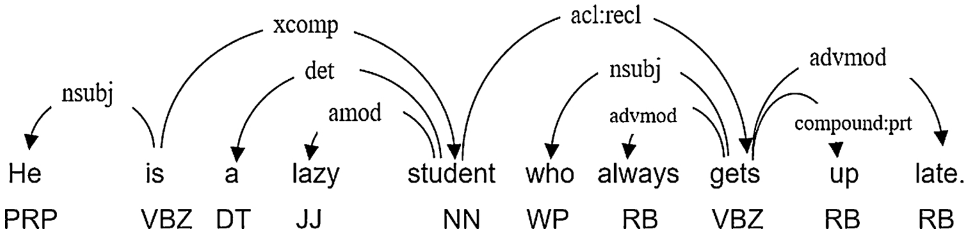
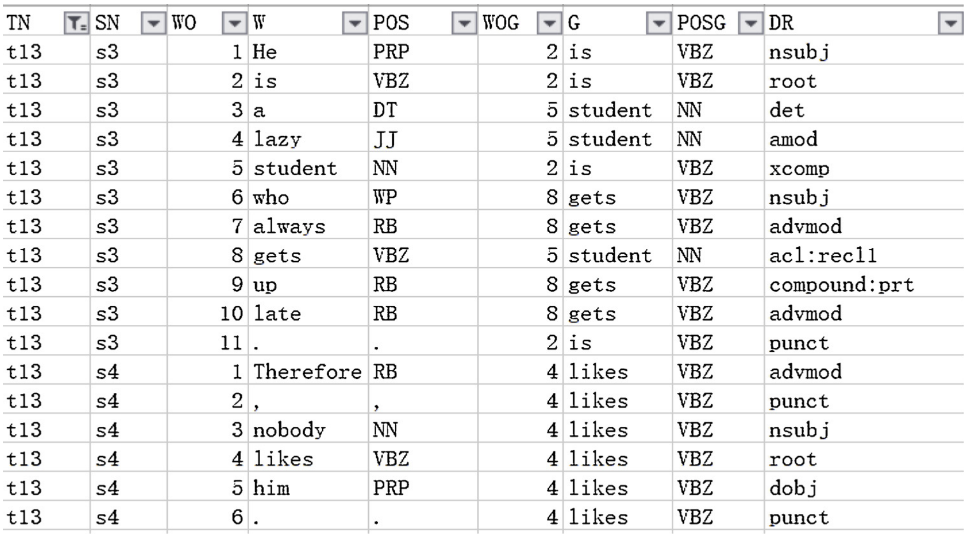


Fig. 1. The syntactic dependency structure of one sentence.



TN	SN	WO	W	POS	WOG	G	POSG	DR
t13	s3	1	He	PRP	2	is	VBZ	nsubj
t13	s3	2	is	VBZ	2	is	VBZ	root
t13	s3	3	a	DT	5	student	NN	det
t13	s3	4	lazy	JJ	5	student	NN	amod
t13	s3	5	student	NN	2	is	VBZ	xcomp
t13	s3	6	who	WP	8	gets	VBZ	nsubj
t13	s3	7	always	RB	8	gets	VBZ	advmod
t13	s3	8	gets	VBZ	5	student	NN	acl:recl1
t13	s3	9	up	RB	8	gets	VBZ	compound:prt
t13	s3	10	late	RB	8	gets	VBZ	advmod
t13	s3	11	.	.	2	is	VBZ	punct
t13	s4	1	Therefore	RB	4	likes	VBZ	advmod
t13	s4	2	,	,	4	likes	VBZ	punct
t13	s4	3	nobody	NN	4	likes	VBZ	nsubj
t13	s4	4	likes	VBZ	4	likes	VBZ	root
t13	s4	5	him	PRP	4	likes	VBZ	dobj
t13	s4	6	.	.	4	likes	VBZ	punct

Fig. 2. Screenshot of the spreadsheet form of the dependency treebank.

and CN/C) and eight fine-grained measures (advcl, ccomp, acl:recl, nmod:poss, compound, amod, prep:attr, and acl:recl1).

4.3. Statistical analyses

Q–Q plots indicated the absence of normal distributions of these 15 measures. Levene's tests further pointed to the absence of the homogeneity in their variance. After the log transformation, seven measures (i.e. MLC, MLT, MLS, DC/C, CN/C, nmod:poss, and amod) met the assumptions of normality and homogeneity of variance. A one-way multivariate analysis of variance (MANOVA) test was applied to ascertain whether there were significant differences in these seven measures among the four groups. Then, Bonferroni post hoc tests were employed to make certain whether the differences in these measures were significant between every two adjacent proficiency levels. The descriptive statistics for the seven measures were still presented by the untransformed values. For the rest eight measures which still had the problem of skewness after transformations (i.e. T/S, CP/C, advcl, ccomp, acl:recl, compound, prep:attr, and acl:recl1), the Kruskal-Wallis test was used to analyze group differences.

Table 4
The eight fine-grained SC measures.

Dimensions	Grammatical structures	Dependency relations	Examples from our treebank
Subordinate clauses	Adverbial clauses	advcl	We were just looking at each other when the bus drove away.
	Complement clauses	ccomp	My friends think I am a good boy.
	Relative clauses	acl:recl[a]	He speaks so loudly, which makes me uncomfortable.
Noun modifiers	Possessive modifiers	nmod:poss	his mother
	Compound nouns	compound	cartoon books
	Adjectival modifiers	amod	a good boy
	Prepositional phrases as attributes	prep:attr	the skill about debate
	Adjectival relative clauses	acl:recl1[a]	the food that I had ordered

[a] *Note*. According to Biber, Johansson, Leech, Conrad, and Finegan (1999)), there are two types of relative clauses in English: adjectival relative clauses and sentential relative clauses. The former is used to modify the head nouns while the latter is used to comment on the previous clauses (e.g. He speaks so loudly, *which makes me uncomfortable*). It should be noted that only adjectival relative clauses can function as noun modifiers, so it is this type that is included in the fine-grained measures concerning noun modifiers.

5. Results

This section will present the statistics for seven large-grained and eight fine-grained SC measures.

5.1. Large-grained measures

As can be seen from Table 5 and Fig. 3, the large-grained measures all augment with the increase of writing proficiency, except CP/C, whose mean value ranges between .115 and .139.

Pillai's trace in MANOVA revealed a significant main effect of writing proficiency on large-grained SC measures (F = 15.376, p = .000). Follow-up univariate comparisons affirmed significant effects of writing proficiency on five measures: MLC (F = 11.078, p = .000, η^2 = .076), MLT (F = 82.247, p = .000, η^2 = .378), MLS (F = 91.076, p = .000, η^2 = .402), DC/C (F = 70.580, p = .000, η^2 = .343), and CN/C (F = 32.767, p = .000, η^2 = .195). The effect sizes of MLT, MLS and DC/C are quite large, all exceeding .3. That is, over 30% of the variances of the three measures could be explained as resulting from different writing proficiency levels. In comparison, the effect size of MLC is smaller. However, it is still acceptable because according to Cohen (1969), the size of .076 is a medium one. Table 6 shows that between at least two pairs of adjacent levels were found significant differences in MLT, MLS, and DC/C. Thus, it can be concluded that these three measures have the potential to best evaluate writing proficiency of high school students. The Kruskal-Wallis test reached a similar result for T/S (χ^2 = 38.042, p = .000).

In brief, these statistical results suggested a tendency for more proficient young high school students to produce longer clauses, longer T-units, longer sentences, more coordinate and subordinate clauses, and more noun phrases. Of these measures, MLT, MLS, and DC/C may better index the writing proficiency of high school English learners.

5.2. Fine-grained measures

5.2.1. Fine-grained measures concerning subordinate clauses

As can be seen from Table 7 and Fig. 4, there are very few relative clauses in the compositions at level 1, though, generally speaking, all the three types of subordinate clauses occur rather infrequently in the compositions at level 1. The advancement of writing proficiency seems to bring about a constant increase in the frequencies of advcl (adverbial clauses) and acl:recl (relative clauses). The frequency of ccomp (complement clauses) ascends to the peak when the writing proficiency reaches level 3 and then decreases. Kruskal-Wallis tests revealed a significant effect of writing proficiency on advcl (χ^2 = 79.348, p = .000), ccomp (χ^2 = 58.344, p = .000), and acl:recl (χ^2 = 147.599, p = .000). In conclusion, all the three types of subordinate clauses are more frequently used by students with higher writing proficiency.

5.2.2. Fine-grained measures concerning noun modifiers

As mentioned in the method section, the present study investigates five types of noun modifiers, including three types of pre-

Table 5
Descriptive statistics for seven large-grained SC measures.

	MLC		MLT		MLS		DC/C		T/S		CP/C		CN/C	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
1	6.701	.127	7.684	.162	7.981	.182	.13	.011	1.04	.008	.133	.012	.382	.021
2	6.708	.091	9.091	.215	10.007	.276	.21	.012	1.10	.014	.115	.009	.440	.019
3	7.004	.105	10.834	.235	12.209	.307	.30	.010	1.12	.013	.119	.009	.571	.022
4	7.515	.134	11.833	.250	13.122	.262	.33	.010	1.12	.011	.139	.010	.647	.025

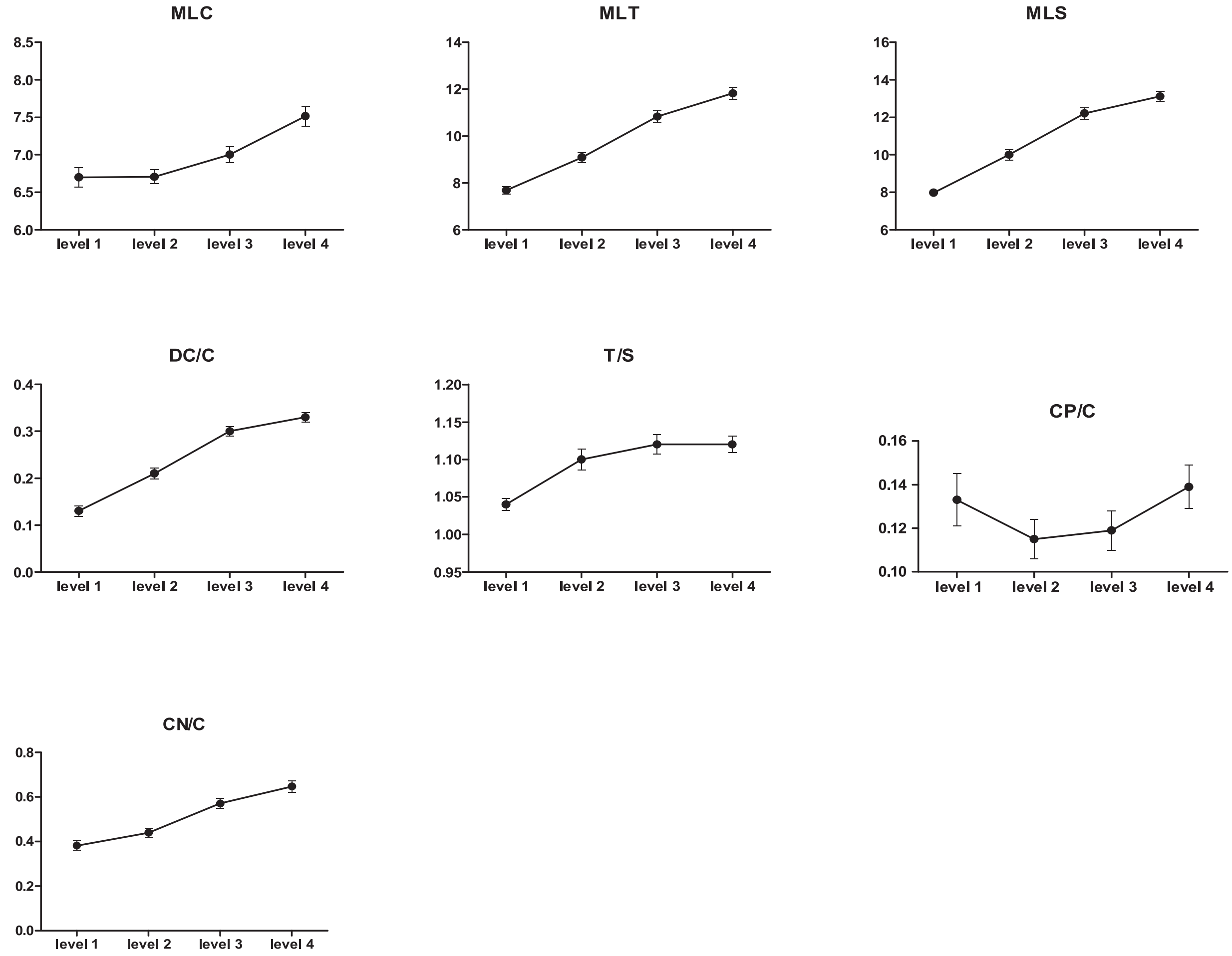


Fig. 3. Development trends of seven large-grained measures.

Table 6
Bonferroni post hoc tests on the five large-grained measures.

	MLC	MLT	MLS	DC/C	CN/C
1-2		.000	.000	.000	
2-3		.000	.000	.000	.000
3-4	.020	.019			

Table 7
Descriptive statistics for three fine-grained measures concerning subordinate clauses.

	advcl		ccomp		acl:recl	
	M	SD	M	SD	M	SD
1	.61	.100	.68	.101	.01	.008
2	1.28	.117	1.49	.144	.31	.060
3	1.56	.092	1.89	.131	.79	.086
4	1.92	.111	1.76	.135	1.01	.074

modifiers: nmod:poss (possessive modifiers), compound (compound nouns), and amod (adjectival modifiers), and two types of post-modifiers: prep:attr (prepositional phrases) and acl:recl1 (adjectival relative clauses). Table 8 shows that the frequencies of these five types of noun modifiers are unevenly distributed across the four proficiency levels. High school students, particularly those at lower proficiency levels, often rely heavily on pre-modifiers, especially the nmod:poss and the amod, to construct noun phrases. Though post-modifiers occur less frequently than pre-modifiers at all the four levels, their frequencies increase steadily as learners advance in

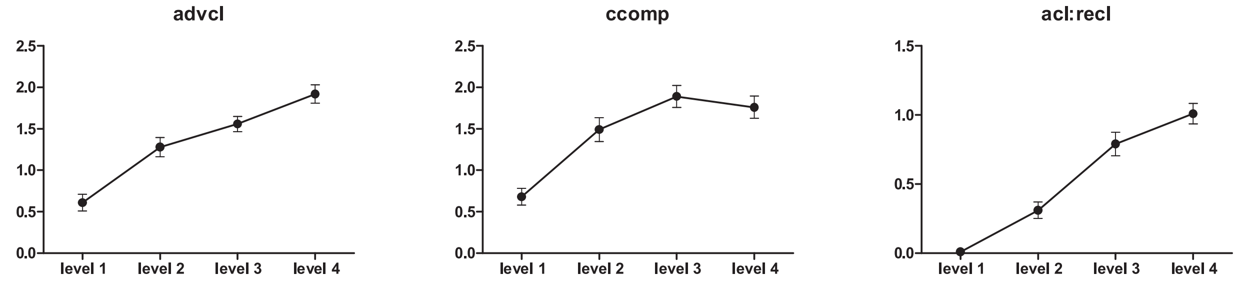


Fig. 4. Development trends of three fine-grained measures concerning subordinate clauses.

Table 8
Descriptive statistics for five fine-grained measures concerning noun modifiers.

	Pre-noun modifiers						Post-noun modifiers			
	nmod:poss		compound		amod		prep:attr		acl:recl1	
	M	SD	M	SD	M	SD	M	SD	M	SD
1	3.634	.207	2.329	.189	5.057	.305	.423	.076	.01	.008
2	3.959	.213	1.713	.170	4.854	.224	.997	.104	.28	.057
3	3.632	.213	1.328	.151	4.974	.205	1.355	.115	.67	.081
4	3.466	.183	.853	.085	4.496	.205	1.773	.143	.83	.068

proficiency. According to Fig. 5, the mean values of compound and amod present an overall descending tendency whereas those of prep:attr and acl:recl1 show a general ascending tendency. The frequency of nmod:poss rises as the writing proficiency advances from level 1 to level 2, but then drops. The MANOVA test and the Kruskal-Wallis test showed significant differences among proficiency levels in compound (χ^2 = 39.245, p = .000), prep:attr (χ^2 = 74.917, p = .000), and acl:recl1 (χ^2 = 123.885, p = .000), but not in amod (F = .371, p = .774, η^2 = .003) or nmod:poss (F = .848, p = .468, η^2 = .006). To sum up, the frequencies of prepositional phrases and adjectival relative clauses correlate positively with writing proficiency while the frequency of compound nouns correlates negatively with it.

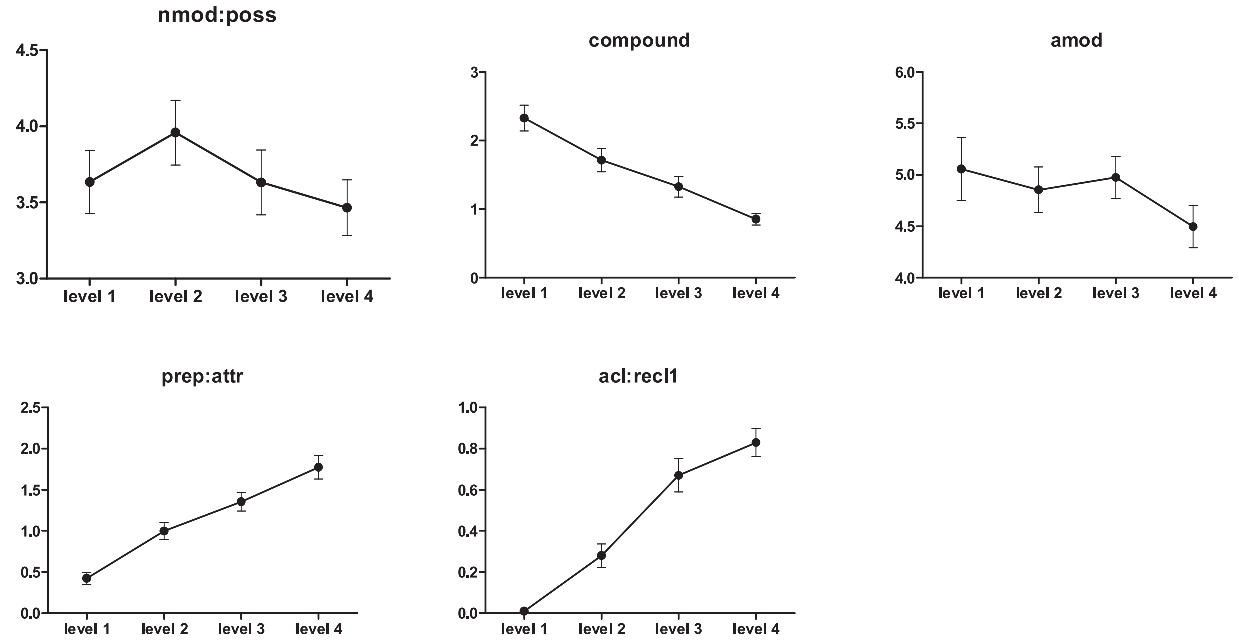


Fig. 5. Development trends of five fine-grained measures concerning noun modifiers.

6. Discussion

This section will attempt, on the basis of above findings, to answer the two research questions and provide some explanations to the new trends in light of other related previous studies.

6.1. The two research questions

The two research questions are mainly concerned with which large-grained and fine-grained measures can best predict writing proficiency. The statistical results suggest that the values of all the large-grained measures, except CP/C, tend to be larger in the writings of higher levels. Compared with other large-grained measures, MLT, MLS and DC/C are probably better indicators of L2 writing proficiency, because they have larger effect sizes, and could distinguish between several adjacent proficiency levels. Interestingly, MLC is a less effective measure of writing proficiency than MLT and MLS, especially for students at levels 1 and 2: The MLC of compositions at level 2 is not conspicuously longer than that at level 1 (see Fig. 3). Some studies hold that MLC, as a complexity measure at the clausal level, mainly captures the use of phrases within a clause (e.g. Alexopoulou, Michel, Murakami, & Meurers, 2017; Bulté & Housen, 2014). Instead, MLT is a complexity measure at the sentential level, primarily capturing clausal and phrasal elaboration (Kyle & Crossley, 2018). Our results also found that MLC behaves quite differently from the other two length-based measures: MLT and MLS. Presumably, the three of them do not tap into the same dimension of complexity, which may be illustrated by the following three sentences.

(1) He is a lazy student who always gets up late (MLC = 5; MLT = 10; MLS = 10).
(2) Being a lazy student, he always gets up late (MLC = 9; MLT = 9; MLS = 9).
(3) He is a lazy student, and he always gets up late (MLC = 5.5; MLT = 5.5; MLS = 11).

In both sentences (1) and (2), the MLT and the MLS are almost the same. However, the MLC of sentence (1) is 5 while that of sentence (2) is 9. Clearly, the longer MLC of sentence (2) is due to the use of the participle phrase “being a lazy student.” The considerable difference in MLT between sentence (3) and sentences (1) and (2) is caused by the absence of subordinate clauses in sentence (3). It is thus suggested that MLC is more likely to be influenced by the use of phrases and MLT is more likely to be influenced by clausal and phrasal elaboration. The MLS in all the three sentences is similar, which, however, consist of different syntactic structures. That is, all the structures, including subordinate clauses, coordinate clauses, and different kinds of phrases, could lead to the increase of sentence length, suggesting that MLS is a holistic syntactic complexity measure. To conclude, MLC is more of a clausal-level complexity measure; MLT is a sentential-level complexity measure, and MLS is a holistic complexity measure. From this perspective, the absence of MLC difference between level 1 and level 2 may be attributed to the slow growth of CN/C and the sharp decrease of CP/C, among other factors (see Fig. 3).

Evidence concerning the development patterns of large-grained measures can be found by scrutinizing the following two excerpts[3] from our corpus. Students at level 1 are only capable of producing very simple sentences in their compositions, and most sentence structures are constructed around verbs instead of nouns. However, students at level 4 can generate more compound and complex sentences (e.g. coordinate and subordinate clauses) and more complex noun phrases, which results in much longer grammatical units.

Excerpt 1 (a script from level 1)

Hi, **my name** is Jone. I am fifteen years old. I am very healthy every day. On **last weekend**, I get up at six. And I exercise from six fifteen to six forty. Than (Then) I eat **my breakfast**. After that, *I brush* ***my teeth*** *and go to school* at seven ten. I eat lunch twenty to twelve. *I get home from school at five and do* ***my homework***. I go to bed at ten. This is **my last weekend**. And you? What did you do on **last weekend**?

Excerpt 2 (a script from level 4)

As **a member of society**, we are supposed to *behave ourselves and do* ***our utmost*** *to help others*. Yet **one thing which happened yesterday** made me feel desperately angry.

I was in **a bus** at **that time, which was so crowed (crowded)** that everyone couldn't even breathe. ***The bus*** *moved slowly and finally I saw* ***the bus station***. ***Few passengers*** *got off* ***the bus*** *and I had* ***a seat***. Then I realized **an old lady** stood next to me. Just at **that time when I was giving the seat to that lady**, **a young man** *pushed me away and had* ***that seat***! *I was really angry, but he just ignored me and didn't feel ashamed at all*. **The scene** replayed in **my mind** all **the time**. I just can't emphasize **the importance of helping others** too much. I believe that no one will accept **that young man** unless he realize **what mistake** had he done.

Note. Coordinate phrases and clauses are in italics; subordinate clauses are underlined; noun phrases are in bold.

The above findings are roughly consistent with Khushik and Huhta (2019), whose findings also demonstrated that measures based on length, subordination, and noun modifiers could reflect the SC development of learners with CEFR levels from A1, A2 and B1, namely, beginner and intermediate learners. Meanwhile, some length-based (e.g. MLS and MLC) and noun-phrase-density-based measures have also been found to be effective measures of SC development of advanced learners (Alexopoulou et al., 2017; Lu, 2011). In view of these findings, it might be proposed that length-related and noun-phrase-related measures are valid measures of SC in L2 writing, regardless of the proficiency level. In the same vein, Verspoor et al. (2017, p. 18) concluded that “these general length

[3] The language mistakes in these two excerpts are not marked or corrected since they do not influence our analysis. Only two misspellings are corrected in brackets because they block our understanding.

measures do well to trace development at a wide spectrum of proficiency levels."

However, although subordination is a vital tool for SC in the writings of beginner and intermediate learners as found in Khushik and Huhta (2019) and the present study, subordination-related measures such as C/T and DC/C seem invalid in gauging the SC in the writings of college-level learners (Bulté & Housen, 2014; Lu, 2011; Yoon, 2017). Instead, the writings of advanced learners tend to be more characterized by noun phrasal elaboration instead of subordination (Biber et al., 2011; Parkinson & Musgrave, 2014). Together with the findings of previous studies, our results have confirmed Lu (2011) and Ortega (2015) assertion that the SC development of L2 learners may follow this overall pattern: To convey complex meanings, beginner L2 learners mainly depend on coordination; intermediate learners resort more to subordination; and advanced learners typically employ noun-phrase elaboration.

When it comes to fine-grained measures, the frequencies of adverbial clauses and relative clauses tend to rise in the compositions at higher levels. From level 1 to level 3, there is a spurt in the frequency of complement clauses, which, nevertheless, decreases when students advance from level 3 to level 4. One possible reason for this boom of complement clauses between levels 1 and 3 is the frequent use of mental verbs, such as *think*, *believe* and *know*, which usually take complement clauses (Beers & Nagy, 2009). From level 3 to level 4, the decrease of complement clauses is probably caused by the higher-level learners' ability to use other syntactic means and constructions for clausal integration. As evidenced by our empirical data and teaching experience, more proficient writers would employ such expressions as "in my opinion" and "as far as I know" to replace the verb frames like "I think/know." As for the five types of noun modifiers, it is found that students at higher proficiency levels tend to rely less on pre-modifiers such as adjectival modifiers and compound nouns, increasingly using post-modifiers, including prepositional phrases and relative clauses. It could be summarized that the use of noun modifiers by young high school learners progresses from pre-attributive modifiers to post-phrasal and clausal modifiers. Such examples abound in our treebank, which can illustrate this developmental pattern of noun modifiers. Learners with lower writing proficiency would repeatedly produce noun phrases such as "my friend" and "the cat food." In comparison, students at higher levels are capable of using more diverse and more complex noun phrases, such as "a friend I know" and "the food for cat." This development pattern of noun modifiers is also found by Jiang, Ouyang, and Liu (2019) and corroborates the hypothesized developmental stages of complexity features proposed by Biber et al. (2011). Similarly, the frequency of prepositional phrases can also predict the writing quality of advanced learners (Taguchi, Crawford, & Wetzel, 2013). However, some studies reported that the writings of advanced learners (e.g. college- and graduate- levels learners) featured the use of more compound nouns instead of more relative clauses (Parkinson & Musgrave, 2014), which is contrary to our findings based on high school students.

6.2. The new trends

The results of the current study have further expanded and validated the new trends mentioned in the research background section. In the first place, our results demonstrate that to a large extent, the SC development in the writings of beginner and intermediate learners is different from that in the writings of advanced learners in several measures, namely, dependent clauses per clause, the normalized frequency of adjectival relative clauses and compound nouns. Therefore, different measures are needed to characterize the SC at different proficiency levels. Accordingly, young L2 English learners at low and intermediate levels warrant more attention because they have been overlooked and understudied. In the second place, the fine-grained measures have enabled us to trace the changes in some specific types of subordinate clauses and noun modifiers. Despite being under the same dimension of subordinate clauses or noun modifiers, with the advancement of writing proficiency, some sub-types show upward patterns, some downward patterns, and some no clear developmental patterns (see Figs. 4 and 5). Put simply, there exist variations on the development trends of different sub-types of subordinate clauses or noun modifiers. Therefore, a dearth of more fine-grained measures would mask some important developmental features and thus blur the development process. Our research results could also lend evidence to the dynamic nature of SC development. Although some of the measures seem to develop linearly (e.g. MLT, MLS, advcl, acl:recl, and prep:attr), other measures show apparent non-linear development patterns, such as spurs and leveling-off (e.g. MLC, T/S, ccomp, and amod). Those measures showing less degrees of variability could be more useful in measuring SC and gauging the writing development (Verspoor et al., 2017). Also, there are differences in the SC development patterns between less proficient learners and advanced learners, which could find explanations from a DST perspective of language development. To be more specific, due to the dynamic development of their syntactic competence, learners at different proficiency levels would rely on different SC devices. Thus, various measures are necessary to assess the writing proficiency of learners at different stages.

7. Conclusion

Employing large-grained and fine-grained measures, we investigated the syntactic complexity in a dependency treebank of 410 narrative writings by Chinese high school students across four writing proficiency levels. The results suggest that three large-grained measures: mean length of T-unit, mean length of sentence, and dependent clauses per clause, are significantly different among compositions at different proficiency levels and demonstrate higher effect sizes, serving as better measures of the writing proficiency of beginner and intermediate L2 English learners. The results also indicate that some fine-grained measures concerning subordinate clauses and noun modifiers have the same functions: Students with higher proficiency, compared with those of lower proficiency, tend to use significantly more complement clauses, adverbial clauses, relative clauses, prepositional phrases, and adjectival relative clauses.

This study is, to our knowledge, the first one to integrate the four recent trends in syntactic complexity studies, to exclusively focus on young beginner and intermediate EFL learners, and to devote itself, on the basis of a dependency treebank, to the dynamic changes in syntactic complexity measures at different granularity levels. Such a study may reveal a rather complete and detailed picture of syntactic complexity development in L2 writing. In addition, this study also shows the considerable value of studying the writings by young beginner and intermediate EFL learners, especially for the research on linguistic complexity development, which has largely focused on advanced L2 learners. It is also advisable for future studies to incorporate more fine-grained measures to examine the developmental properties in other dimensions such as adverbial modifiers. Lastly, since syntactic complexity development is a dynamic rather than a static process, there is abundant room for future research to explore the dynamic development pathways of large-grained and fine-grained syntactic complexity measures more systematically in L2 writing, especially in a longitudinal fashion, which might provide more insights into the developmental process.

Methodologically, we have built an annotated and manually-checked EFL learner dependency treebank, which could be adopted by researchers in their future studies whose learners are of other language backgrounds. As can be seen from this study, dependency syntactically-annotated corpora make data extraction convenient, especially for the fine-grained measures. Pedagogically, language teachers can utilize the findings of our study to improve students' writing proficiency. For example, relative clauses, which are affirmed by our study as an important means of subordinate clauses and noun modifiers, may deserve extra attention in language teaching.

However, our study has two main limitations which are noteworthy. All the essays in our study are narratives, which may cast some doubt on the generalizations of the findings, since genre is generally regarded as having some influence on syntactic complexity evaluation (Qin & Uccelli, 2016). In addition, all the students in our study are native speakers of Chinese, which renders it impossible to find out whether the findings are universal patterns or merely unique to Chinese learners of English.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study is supported by the National Social Science Foundation of China (Grant No. 17AYY021), the Fundamental Research Funds for the Central Universities (Zhejiang University), Zhejiang University Liberal Arts Teachers' Teaching and Research Development Special Funding, and the MOE Project of the Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies.

References

Ai, H., & Lu, X. (2013). A corpus-based comparison of syntactic complexity in NNS and NS university students writing. In A. Díaz-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.

Ansarifar, A., Shahriari, H., & Pishghadam, R. (2018). Phrasal complexity in academic writing: A comparison of abstracts written by graduate students and expert writers in applied linguistics. *Journal of English for Academic Purposes, 31*, 58–71.

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

Beers, S. F., & Nagy, W. E. (2009). Syntactic complexity as a predictor of adolescent writing quality: Which measures? Which genre? *Reading and Writing, 22*(2), 185–200.

Biber, D. (1988). *Variation across speech and writing.* Cambridge: Cambridge University Press.

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.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English.* Harlow: Pearson Education Limited.

Bulté, B., & Housen, A. (2012). Defining and operationalising 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.

Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing, 26*, 42–65.

Bulté, B., & Housen, A. (2018). Syntactic complexity in L2 writing: Individual pathways and emerging group trends. *International Journal of Applied Linguistics, 28*(1), 147–164.

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.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

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

Gablasova, D., Brezina, V., McEnery, T., & Boyd, E. (2017). Epistemic stance in spoken L2 English: The effect of task and speaker style. *Applied Linguistics, 38*(5), 613–637.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy and fluency in second language acquisition. *Applied Linguistics, 30*, 461–473.

Housen, A., & Simoens, H. (2016). Introduction: Cognitive perspectives on difficulty and complexity in L2 acquisition. *Studies in Second Language Acquisition, 38*(2), 163–175.

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

Hunt, K. W. (1965). *Grammatical structures written at three grade levels (research report no. 3)*Champaign, IL: National Council of Teachers of English.

Hunt, K. W. (1970). Do sentences in the second language grow like those in the first? *TESOL Quarterly, 4*, 195–202.

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

Jiang, J., & Liu, H. (2018). *Quantitative analysis of dependency structures.* Berlin: De Gruyter Mouton.

Jiang, J., Ouyang, J., & Liu, H. (2019). Interlanguage: A perspective of quantitative linguistic typology. *Language Sciences, 74*, 85–97.

Jiang, W. (2013). Measurements of development in L2 written production: The case of L2 Chinese. *Applied Linguistics, 34*(1), 1–24.

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. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *Modern Language Journal, 102*(2), 333–349.

Lahuerta Martínez, A. C. (2018). Analysis of syntactic complexity in secondary education EFL writers at different proficiency levels. *Assessing Writing, 35*, 1–11.

Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics, 27*(4), 590–619.

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

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.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*, 474–496.

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

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.

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.

Meurers, D., & Dickinson, M. (2017). Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics. *Language Learning, 67*(S1), 66–95.

Michel, M., Kormos, J., Brunfaut, T., & Ratajczak, M. (2019). The role of working memory in young second language learners' written performances. *Journal of Second Language Writing, 45*, 31–45.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*(4), 555–578.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics, 24*(4), 492–518.

Ortega, L. (2015). Syntactic complexity in L2 writing: Progress and expansion. *Journal of Second Language Writing, 29*, 82–94.

Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research, 31*(1), 117–134.

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.

Qin, W., & Uccelli, P. (2016). Same language, different functions: A cross-genre analysis of Chinese EFL learners' writing performance. *Journal of Second Language Writing, 33*, 3–17.

Ryshina-Pankova, M. (2015). A meaning-based approach to the study of complexity in L2 writing: The case of grammatical metaphor. *Journal of Second Language Writing, 29*, 51–63.

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.

Schmid, M., Verspoor, M., & MacWhinney, B. (2011). Coding and extracting data. In M. Verspoor, K. de Bot, & W. Lowie (Eds.). *A dynamic approach to second language development: Methods and techniques* (pp. 39–54). Amsterdam: John Benjamins.

Spoelman, M., & Verspoor, M. (2010). Dynamic patterns in development of accuracy and complexity: A longitudinal case study in the acquisition of Finnish. *Applied Linguistics, 31*(4), 532–553.

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.

Tesnière, L. (1959). *Eléments de la syntaxe structurale.* 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 en didactique des langues et des cultures, 14*, 1–27.

Verspoor, M., Lowie, W., & van Dijk, M. (2008). Variability in second language development from a dynamic systems perspective. *Modern Language Journal, 92*(2), 214–231.

Verspoor, M., Schmid, M. S., & Xu, X. (2012). A dynamic usage based perspective on L2 writing. *Journal of Second Language Writing, 21*, 239–263.

Vyatkina, N., Hirschmann, H., & Golcher, F. (2015). Syntactic modification at early stages of L2 German writing development: A longitudinal learner corpus study. *Journal of Second Language Writing, 29*, 28–50.

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.

Yoon, H. J. (2017). Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct multidimensionality. *System, 66*, 130–141.

Yoon, H. J. (2018). The development of ESL writing quality and lexical proficiency: Suggestions for assessing writing achievement. *Language Assessment Quarterly, 15*, 387–405.

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 30 research papers in influential journals and is the author of two monographs about linguistics and applied linguistics.

Peng Bi is a PhD candidate in linguistics and applied linguistics from Zhejiang University. His research interests include applied linguistics and quantitative linguistics. He has published several papers about applied linguistics and quantitative linguistics.

Haitao Liu is a Qiushi Distinguished Professor of Linguistics and Applied Linguistics at Zhejiang University, Distinguished Visiting Professor at Beijing Language and Culture University, Yunshan Leading Professor at Guangdong University of Foreign Studies. His research interests include applied linguistics, dependency grammar and language complex system. He is the author of more than 190 scientific publications about linguistics and applied linguistics.