DOI: 10.1111/modl.12907

The Modern Language Journal

ORIGINAL ARTICLE

L2 English speaking syntactic complexity: Data preprocessing issues, reliability of automated analysis, and the effects of proficiency, L1 background, and topic

Minjin Kim | **Xiaofei Lu**

Department of Applied Linguistics, The Pennsylvania State University, University Park, Pennsylvania, USA

Correspondence
Xiaofei Lu, The Pennsylvania State University, Department of Applied Linguistics, 234 Sparks Building, University Park, PA 16802, USA.
Email: xxl13@psu.edu

Abstract

The effects of learner- and task-related variables on second language (L2) writing syntactic complexity (SC) have been extensively investigated. However, previous research has rarely assessed the reliability of computational tools for analyzing the SC of L2 spoken production, and we know less about the effects of such variables on L2 speaking SC. Using data from the International Corpus Network of Asian Learners of English, this study explores data preprocessing issues for preparing L2 English speech samples for automated SC analysis, evaluates the reliability of L2 Syntactic Complexity Analyzer on preprocessed L2 English speech samples, and examines the effects of proficiency, first language (L1) background, and topic on L2 speaking SC. Our manual analysis of 30 random speech samples identified several issues that can be addressed through preprocessing to improve the accuracy of automated SC analysis. Results from multiple linear mixed-effects models revealed significant effects of proficiency, L1 background, and topic on the mean length of clause, the number of complex AS-units per AS-unit, and the number of dependent clauses and complex nominals per clause in L2 learners' spoken production. Our findings have useful implications for L2 speaking pedagogy and assessment as well as future L2 speaking SC research.

KEYWORDS

L1 background, L2 proficiency, L2 speaking, reliability, syntactic complexity, topic

The effects of important learner- and task-related variables on the syntactic complexity (SC) of second language (L2) learners' written production have received much attention in L2 writing research (e.g., Atak & Saricaoglu, 2021; Biber et al., 2016; Polio & Yoon, 2018; Yang et al., 2015). Commonly employing computational tools for automating SC analysis to analyze large-scale corpora of L2 writing, this body of research has shed useful light on the ways in which L2 writing SC may vary as a function of such learner-related variables as L2 proficiency and first language (L1) background and such task-related variables as task or prompt type, genre, and topic, often also offering insights into the reliability of computational tools for L2 writing SC analysis (e.g., Ai & Lu, 2013; Lu, 2011; Yoon & Polio, 2017). Meanwhile, similar research in the domain of L2 speaking has lagged behind, partially due to the more limited availability of large-scale corpora of L2 speaking that encode rich learner- and task-related variables as well as the greater difficulty in analyzing such corpora using computational tools (Brezina et al., 2022; Chen & Zechner, 2011; Chen et al., 2018). Consequently, we know much less about the effects of such variables on L2 speaking SC or the reliability of computational tools for L2 speaking SC analysis. Knowledge of such issues, however, will have clear implications for understanding variation in L2 use, for L2 speaking pedagogy, and for SC research methodology.

In light of these research gaps, the current study aims to achieve two goals using data sampled from the monologue subcorpus of the International Corpus Network of Asian Learners of English (ICNALE; Ishikawa, 2014). First, we assess the reliability of the L2 Syntactic Complexity Analyzer (L2SCA; Lu, 2010), a tool originally designed for automating the analysis of SC of L2 written production, on the SC of L2 spoken production. This is done by comparing our manual analysis and the analysis by the tool of a random sample of 30 transcripts from the dataset, following a procedure that identified multiple issues of the speech samples that can be systematically addressed through preprocessing to enhance the accuracy of the tool. Second, we construct multiple linear mixed-effects (LME) models to examine the effects of three learner- and task-related variables—namely, L2 proficiency, L1 background, and topic—on the SC of L2 English spoken production, with two different operationalizations of L2 proficiency. Our findings are discussed in terms of their implications for L2 speaking SC research and L2 speaking pedagogy and assessment.

SYNTACTIC COMPLEXITY: CONCEPTUALIZATION, OPERATIONALIZATION, AND AUTOMATED TOOLS

Several prior investigations into L2 complexity have distinguished between absolute and relative linguistic complexity, with the former referring to the inherent complexity of linguistic units that can be objectively quantified with respect to the number, variety, and elaborateness of their structural elements and the latter referring to the relative difficulty (e.g., for comprehension or learning) of linguistic units based on their usage features (e.g., Bulté & Housen, 2012, 2018; Kyle, 2016). The current study is concerned with SC as an absolute, objective, and primarily quantitative characteristic of linguistic units, defined as the degree to which the syntactic structures used in language production are diverse, sophisticated, and elaborate (Bulté & Housen, 2012, 2018; Housen & Kuiken, 2009; Ortega, 2003; Pallotti, 2009, 2015). There is good consensus in the L2 acquisition and development literature that SC should be conceptualized as a construct with multiple dimensions, including global complexity, complexity by subordination, complexity by coordination, and phrasal complexity (e.g., Norris & Ortega, 2009; Ortega, 2015). Over the past few decades, numerous holistic and fine-grained measures have been proposed to operationalize different dimensions of this construct. Holistic measures assess the degree of complexity of broad categories of structural units without differentiating specific structural subtypes within those categories (e.g., Ortega, 2003; Wolfe-Quintero et al., 1998). For example, the number of dependent clauses per clause, a holistic measure of complexity by subordination, does not differentiate different subtypes of dependent clauses. Fine-grained measures

TABLE 1 Holistic measures in L2 Syntactic Complexity Analyzer (Lu, 2010).

Type	Syntactic complexity measure	Code
Length of production unit	Mean length of clause	MLC
	Mean length of sentence	MLS
	Mean length of T-unit	MLT
Sentence complexity	Clauses per sentence	C/S
Subordination	Clauses per T-unit	C/T
	Complex T-units per T-unit	CT/T
	Dependent clauses per clause	DC/C
	Dependent clauses per T-unit	DC/T
Coordination	Coordinate phrases per clause	CP/C
	Coordinate phrases per T-unit	CP/T
	T-units per sentence	T/S
Particular structures	Complex nominals per clause	CN/C
	Complex nominals per T-unit	CN/T
	Verb phrases per T-unit	VP/T

tap into specific subtypes of clausal and phrasal structures (e.g., Biber et al., 2016; Kyle, 2016; Kyle & Crossley, 2018). For example, the number of adverbial clauses per clause, a fine-grained measure of complexity by subordination, focuses on the use of a particular type of dependent clause. These two types of measures usefully complement each other, allowing researchers to focus on the degree of granularity that best matches their analytical purposes.

The large number of measures proposed to operationalize SC has found their way into several computational tools for automating SC analysis. For example, the L2SCA (Lu, 2010) provides 14 holistic measures of five categories, as summarized in Table 1. The Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC; Kyle, 2016) provides a large set of fine-grained measures of clausal (e.g., clause complements per clause) and phrasal complexity (e.g., adjectival modifiers per object of the preposition). Coh-Metrix (McNamara et al., 2014) offers a set of SC measures gauging embeddedness (e.g., word counts before the main verb), edit distance (e.g., number of changes necessary to make one sentence identical to its next one), syntax similarity (i.e., the ratio of intersecting nodes between two parse trees), and syntactic pattern density (e.g., number of prepositional phrases per 1,000 words). With their free availability and their ability to efficiently process language samples in large batches, these tools have facilitated much recent L2 SC research.

Meanwhile, Alexopoulou et al. (2021) underscored the importance of recognizing and understanding potential sources that may introduce noise and/or other confounds in the data used for automated language analysis. In their analysis of L2 writing, Huang et al. (2018) found that common learner errors—such as those related to punctuation, argument structure, and preposition usage—often cause parsing errors and suggested error correction as an effective preprocessing step. Due to its more spontaneous nature, spoken data is likely to contain additional sources of noise—particularly disfluency features such as false starts and repetitions—that may affect the accuracy of automated SC analysis, making careful preprocessing especially necessary in preparing spoken data for such analysis. While some efforts have been made to address such issues in constructing disfluency-annotated corpora of L1 English speech (e.g., Godfrey et al., 1992) and researching L1 speech recognition (e.g., Griffis et al., 2016), less attention has been paid to these issues in L2 speaking research.

EFFECTS OF LEARNER- AND TASK-RELATED VARIABLES ON L2 WRITING SYNTACTIC COMPLEXITY

The effects of key learner- and task-related variables on the SC of L2 learners' written production have been a major concern of L2 writing SC research. While many learner-related variables can lead to variation in L2 use—such as L2 proficiency, linguistic and cultural background, and sociolinguistic characteristics (e.g., gender)—learner corpus studies of their effects on L2 writing SC have centered on L2 proficiency and L1 background (Gablasova et al., 2019; Lu, 2023). Proficiency level has been measured using both learner-centered methods based on learner characteristics (e.g., proficiency test scores) and text-centered methods based on text characteristics (e.g., human ratings of writing quality), and there is increasing effort to align proficiency levels with a common proficiency scale such as the Common European Framework of Reference for Languages (CEFR) to facilitate cross-study comparisons (Carlsen, 2012). While increases in SC measures do not inherently signify language development (Pallotti, 2009, 2015), some measures have been found to discriminate between or correlate with proficiency levels. For example, using a corpus of essays written by Chinese English-as-a-foreign-language (EFL) learners and operationalizing proficiency level as school level, Lu (2011) found that 10 holistic indices from the L2SCA differentiated learners at different school levels in different ways. The three length-based measures (i.e., MLS, MLT, and MLC) and two complex nominal measures (i.e., CN/C and CN/T) increased consistently from Level 1 to Level 4; two clausal subordination measures (i.e., DC/C and DC/T) initially increased from Level 1 to Level 2 and then declined from Level 2 to Level 4; two coordinate phrase measures (i.e., CP/C and CP/T) increased from Level 1 to Level 3 and then stayed stable from Level 3 to Level 4; and the sentence complexity measure (i.e., C/S) declined consistently from Level 1 to Level 4. Using a corpus of independent essays on the Test of English as a Foreign Language (TOEFL), Kyle and Crossley (2018) reported that one fine-grained clausal complexity measure (i.e., nominal subjects per clause) and six fine-grained phrasal complexity measures (e.g., adjectival modifiers per object of the preposition) from the TAASSC significantly predicted the holistic scores of those essays.

Previous research has shown that L2 learners' L1 background can affect different aspects of their L2 use in various ways (e.g., MacWhinney, 2012; Murakami & Alexopoulou, 2016), although relatively few studies have directly assessed the effect of L1 background on L2 writing SC. In an earlier study, Crossley and McNamara (2012) reported significant differences in one SC measure (i.e., the number of words before the main verb) among essays written by L2 English learners from four L1 backgrounds. Specifically, L1 German and Spanish learners used significantly more words before the main verb than L1 Finnish learners, who in turn used more than L1 Czech learners. Partially motivated by that study, Lu and Ai (2015) systematically examined differences in the SC of English writing among college-level English writers with eight L1 backgrounds representing four language families—namely, Indo-European (Bulgarian, English, French, German, and Russian), Japonic (Japanese), Niger–Congo (Tswana), and Sino–Tibetan (Chinese). Based on an analysis of 1,600 argumentative essays (200 from each L1 group) using 14 holistic SC indices from the L2SCA, they reported significant differences between the L1 English group and one or more L2 groups in all 14 measures, as well as drastically varied patterns of difference from the L1 English group among the seven L2 groups. They argued that the observed patterns of intergroup variation cannot be accounted for by proficiency alone and noted some influence of the typological similarities and differences between English and the other seven L1s on those patterns.

A sizable body of research has explored the effects of task-related variables on L2 writing SC, such as writing topic (e.g., Atak & Saricaoglu, 2021; Yang et al., 2015; Yoon, 2017), task or prompt type (e.g., Biber et al., 2016; Michel et al., 2019; Shi et al., 2020), and genre (e.g., Lu, 2011; Yoon & Polio, 2017). Given the focus of the current study, we limit our review of empirical findings to the few studies of topic effect in the literature. Yang et al. (2015) analyzed 380 argumentative essays written by 190 English-as-a-second-language (ESL) graduate students on two topics (i.e., whether people overemphasize personal appearance and whether early-life planning helps ensure a good future) using

eight holistic SC indices. The essays on the first topic showed a higher amount of elaboration at the finite clause level, with longer clauses and more coordinate phrases and complex noun phrases per clause. Those on the second topic demonstrated greater overall sentence complexity and a higher amount of subordination, with longer sentences and more dependent clauses per T-unit and more nonfinite elements per clause. Yang et al. (2015) argued that the more frequent use of multipropositional sentences with subordination for the second topic was expected in that it was more likely to elicit causal reasoning, which requires juxtaposition of the relationship between multiple entities or events. Yoon (2017) analyzed 1,198 argumentative essays written by college-level EFL learners on two topics: a part-time job topic ("It is important for college students to have a part-time job") and a smoking topic ("Smoking should be completely banned at all the restaurants in the country"). He found that the essays on the part-time job topic demonstrated significantly higher levels in six holistic measures, including three length measures (i.e., MLS, MLT, and MLC), a subordination measure (i.e., C/T), and two phrasal complexity measures (i.e., CP/C and CN/C). He argued that the part-time job topic was likely more relevant and familiar to college students than the smoking topic. Atak and Saricaoglu (2021) analyzed a set of essays on three topics related to the death penalty, online learning, and cell phones. They found that the death penalty topic elicited significantly more complex structures than the other two topics, even though it was deemed to be the least relevant topic for college students. They argued that the topic effect observed could not be attributed to the relevance of the topic; rather, the death penalty topic was more impersonal and therefore cognitively more demanding.

PREVIOUS RESEARCH ON L2 SPEAKING SYNTACTIC COMPLEXITY

Syntactic complexity in L2 speech has also attracted substantial interest from L2 researchers. Much research in this area focused on the effects of task-related variables such as production mode (i.e., speaking vs. writing) and task complexity. For example, in a longitudinal study of SC development in two 15-year-old identical twin Chinese EFL learners' writing and speaking, Chan et al. (2015) found that MLT, DC/T, and CP/T were all higher in spoken than in written production, but not much change occurred over the 8-month period. They attributed the lower SC in written production to the extra processing cost associated with writing and the learners' greater focus on using the right words than writing complex sentences. In Hwang et al.'s (2020) comparison of spoken and written data produced by young Korean EFL learners (mean age = 11.26 years), CP/T was also significantly higher in spoken production, but DC/T and two other measures (i.e., MLS and VP/T) were significantly higher in written production. They contended that learners can allocate attention and control linguistic forms more effectively while writing than while speaking. Meanwhile, the differences in the L1 backgrounds and age of the young learners between the two studies may have contributed to the inconsistency in their results, and the small sample size of Chan et al.'s (2015) study may have been another factor. Biber et al. (2016) analyzed a corpus of responses to two spoken and two written tasks in the TOEFL iBT (Internet-based test) using 23 fine-grained grammatical complexity features and reported that speech is less linguistically complex than writing. Specifically, while adverbs and finite adverbial clauses occurred significantly more frequently in spoken production, many other structures, particularly those related to noun phrase modification, occurred significantly more frequently in writing than in speech. The study further revealed that integrated tasks in the TOEFL iBT elicited more complex structures than independent tasks. Many task-based language teaching (TBLT) studies also investigated SC in L2 speech along with accuracy and fluency. The majority of these studies have used a relatively small number of representative measures for each construct to examine the effects of task complexity (e.g., Levkina & Gilabert, 2012; Révész, 2011) or the interaction effects of task complexity and production mode (Vasylets et al., 2017). While such studies have offered useful insights for enhancing TBLT classroom design, the effects of multiple important learner-related (e.g., proficiency and L1 background) and task-related (e.g., topic) variables on L2 speaking SC remain underexplored.

A small group of studies investigated which SC features could distinguish proficiency levels, predict quality ratings or communicative adequacy, or show longitudinal growth in L2 spoken production. For example, Iwashita et al. (2008) analyzed 250 spoken responses to 10 different tasks in the TOEFL iBT, grouped into five assessed proficiency levels, using four SC indices—namely, C/T, DC/C, VP/T, and mean length of utterance. They reported that the last two indices significantly discriminated between proficiency levels, with responses rated at the two highest levels demonstrating higher values in these indices. Regarding the relationship between SC and L2 speaking quality, Chen and Zechner (2011) reported that a regression model with 17 SC features, including several features from L2SCA (i.e., MLS, MLT, DC/C, and CT/T), achieved an overall correlation of 0.49 with human ratings of L2 English speech. In their later work on building an automated speech scoring system, Chen et al. (2018) further confirmed the usefulness of some SC features in such a system. Révész et al. (2014) reported that ESL learners' speech samples with higher communicative adequacy ratings tended to have higher SC (measured using clauses per AS-unit, words per AS-unit, and conjoined clauses per 100 words). Finally, in a longitudinal study, Mostafa et al. (2021) analyzed speech samples produced by 76 ESL learners over a 15-week period using multiple linguistic features and reported that the number of coordinate clauses per AS-unit (i.e., an utterance with an independent clause or subclausal unit along with its associated subordinate clauses) increased significantly over time.

Chen and Zechner (2011) and Chen et al. (2018) noted that measuring L2 speaking SC is challenging due to the prevalence of errors and conversational features. This observation echoes that of Meurers and Dickinson (2017), who outlined specific challenges learner data may pose to automated analysis. Such challenges may have contributed to the relative paucity of large-scale L2 speaking SC research. They also call for research into ways to mitigate the effects of such errors and conversational features and improve the accuracy of automated analysis of L2 speaking SC. With respect to the effects of learner- and task-related variables on L2 speaking SC, the studies reviewed in this section suggest significant effects of proficiency (measured using text-centered methods) and task type. However, much remains to be explored, including the effects of proficiency measured using learner-centered methods or mapped onto CEFR levels, L1 background, and topic. These factors have been consistently shown to affect L2 writing SC, and an understanding of their effects on L2 speaking SC will have tangible implications for L2 speaking pedagogy and assessment. These research gaps have motivated the two goals of the current study. Our first goal is to identify issues in speech samples produced by L2 English learners that can be addressed to enhance the accuracy of an automated SC analysis tool—that is, L2SCA—and to evaluate the reliability of the tool following the preprocessing step. Our second goal is to examine the effects of three learner- and task-related variables on L2 speaking SC—namely, proficiency, L1 background, and topic. The specific research questions addressed are:

RQ1. What features of L2 English speech samples may affect the accuracy of the L2 Syntactic Complexity Analyzer (Lu, 2010), and how reliable is the tool for analyzing L2 English speech samples after those issues are systematically resolved?

RQ2. What are the effects of proficiency, L1 background, and topic on the syntactic complexity of L2 English speaking?

METHODOLOGY

Data

The data used in the current study were sampled from the spoken monologue subcorpus of the ICNALE (Ishikawa, 2014). The ICNALE consists of over 10,000 essays and speeches produced by college-level L2 English learners in 10 Asian countries or regions, as well as L1 English speakers,

TABLE 2 Distribution of the speech samples by Common European Framework of Reference for Languages level and L1 background.

L1 background	A2	B1	B2	Total
Chinese	60	152	56	268
Japanese	60	248	116	424
Korean	0	116	120	236
Total	120	516	292	928

organized into four components: spoken monologues, spoken dialogues, written essays, and edited essays. In addition to its availability and scale, the ICNALE has two desirable features that make it especially appropriate for our study. First, the corpus compilers have carefully controlled task-related variables such as topic and time constraints, allowing us to avoid potential effects of such variables on learner production. Second, the corpus provides information on several learner-related variables worthy of investigation, including each learner's proficiency level, L1 background, gender, and college major or occupation. The spoken monologue subcorpus of the ICNALE consists of 4,400 speech samples (60 seconds each) produced by 1,100 participants (four samples per participant). Each participant was asked to state and argue for their opinion regarding two topics: (a) having a part-time job in college, and (b) smoking in restaurants. For each topic, each participant produced two speech samples with different preparation time: 20 seconds for the first speech and 10 seconds for the second. For each speech, the participants were encouraged to speak as much as possible in 60 seconds, and both the resulting audio file and its corresponding manual transcript were included in the corpus. The participants were required to take a vocabulary size test (VST) covering the top 5,000-word levels (Nation & Beglar, 2007) and asked to submit their Test of English for International Communication (TOEIC) or TOEFL scores. They were then classified into one of the four proficiency bands linked to the CEFR, that is, A2 (waystage), B1_1 (threshold: lower), B1_2 (threshold: upper), and B2+ (vantage or higher), based on their scores on the TOEIC, TOEFL, or VST using a mapping scheme developed in 2010 in accordance to the then-current official mapping guidelines offered by administrators of TOEIC and TOEFL (the VST scores were converted to TOEIC scores first; see Ishikawa, 2014, for further details).

Given our analytical focus on the effects of proficiency, L1 background, and topic on the SC of L2 English speaking, we selected 928 speech samples produced by 232 L1 Chinese, Japanese, and Korean EFL learners (four samples per learner) whose TOEIC scores and CEFR levels were both available in the corpus. Table 2 summarizes the distribution of the speech samples by CEFR level (with B1_1 and B1_2 combined into B1) and L1 background. The mean of these learners' TOEIC scores was 708.125 ($SD = 130.010$). We also used the L1-English-speaker component of the spoken monologue subcorpus as a reference corpus, which contains 600 speech samples produced by L1 English speakers under the same task conditions.

Syntactic complexity analysis

While in our view both holistic and fine-grained SC measures are worthy of investigation, we decided to focus on a set of holistic measures in the current study and leave the examination of fine-grained measures to a future study. This decision was based on the need to manually assess the reliability of automated tools on L2 speech samples and informed by previous findings that the accuracy of such tools tends to be lower for measures tapping into more specific structures even on L2 writing data (e.g., Lu, 2010). Specifically, the speech samples were automatically analyzed using L2SCA with its full set of 14 indices that capture different dimensions of SC (see Table 1).

Originally developed for analyzing the SC of L2 English writing, L2SCA takes a language sample in plain text format as input and generates 14 holistic SC indices for the sample. To calculate these indices, L2SCA uses the Stanford Parser (Klein & Manning, 2003) to produce parse trees for the sentences in the sample and the Tregex tool (Levy & Andrew, 2006) to identify and count relevant structural units (i.e., sentences, T-units, complex T-units, clauses, dependent clauses, coordinate phrases, complex nominals, and verb phrases) in the parse trees. Lu (2010) manually annotated 20 essays written by advanced college-level Chinese EFL learners and reported F-scores of 0.83 or more for the structural units identified and correlations of 0.834 or more between the SC scores computed by the tool and human annotators. Several other studies have reported comparable levels of reliability of the tool on essays produced by high-intermediate-level (Polio & Yoon, 2018; Yoon & Polio, 2017) and beginner- and intermediate-level L2 English learners (Jiang et al., 2019).

A few studies have employed the tool to analyze L2 English spoken data (e.g., Chan et al., 2015; Hwang et al., 2020), but systematic evaluations of the reliability of the tool on such data remain rare. Given the differences between spoken and written production (e.g., repetitions, fillers, and self-corrections) as well as the potential variation in transcription conventions (e.g., regarding punctuation and capitalization), such evaluations would appear methodologically necessary, as the reliability of the automated analysis would be consequential for the reliability of the results of the study. To assess the performance of L2SCA on our dataset—and, at the same time, to identify and potentially resolve issues in the dataset that may negatively affect its performance—we employed stratified random sampling to select 30 speech samples representing the proportional distribution of different proficiency groups (i.e., CEFR levels) in the dataset and analyzed them in the following steps: First, we parsed 10 of the 30 samples using the version of the Stanford Parser (i.e., version 4.2.0) included in L2SCA, queried the parse trees using the Tregex patterns in L2SCA, and manually examined the results to identify features of L2 spoken production that could have contributed to structural unit identification errors. Second, to assess the accuracy of word recognition and sentence boundary marking in the manual transcripts of the 30 speech samples from the ICNALE, we manually transcribed these speech samples following the criteria established by Foster et al. (2000) and Hwang et al. (2020) for speech transcription and also used the Sonix Automated Speech Recognition (ASR) software (https://sonix.ai/) to automatically transcribe them as well. We then compared the words and sentence boundaries among the three versions of transcripts of those samples. As the specific ASR system was not sufficiently accurate, we used the ICNALE transcripts for further analysis. Third, we manually annotated disfluency features in the 30 transcripts from the ICNALE. Based on an analysis of the problematic features identified in the first three steps, we wrote a preprocessing script in Python 3 to batch modify each sample to address those issues and assessed the accuracy of the preprocessing script. Finally, we manually coded the 30 samples for all relevant structural units following Lu's (2010) definitions and Polio and Yoon's (2018) guidelines for manual coding and evaluated the accuracy of L2SCA's identification of the structural units as well as the Pearson correlations between human- and tool-computed indices. The accuracy of L2SCA's structural unit identification was assessed using the common measures of precision, recall, and F-score (e.g., Lu, 2010), all of which range from 0 to 1. For each type of structural unit, precision is calculated as the number of identical units between the human and L2SCA coding divided by the number of units in the L2SCA coding, with a higher value indicating fewer false positives in the L2SCA coding; recall is calculated as the number of identical units between the human and L2SCA coding divided by the number of units in the human coding, with a higher value indicating fewer missed units in the L2SCA coding; and F-score is calculated as (2 × Precision × Recall) / (Precision + Recall), with a higher value indicating a higher level of overall agreement between the human and L2SCA coding. Following this evaluation procedure, all remaining samples in the dataset were preprocessed and then automatically analyzed using L2SCA. Figure 1 visually presents the step-by-step procedure of data preprocessing and reliability assessment, and more detailed explanations of some of these steps are provided in the Results section.

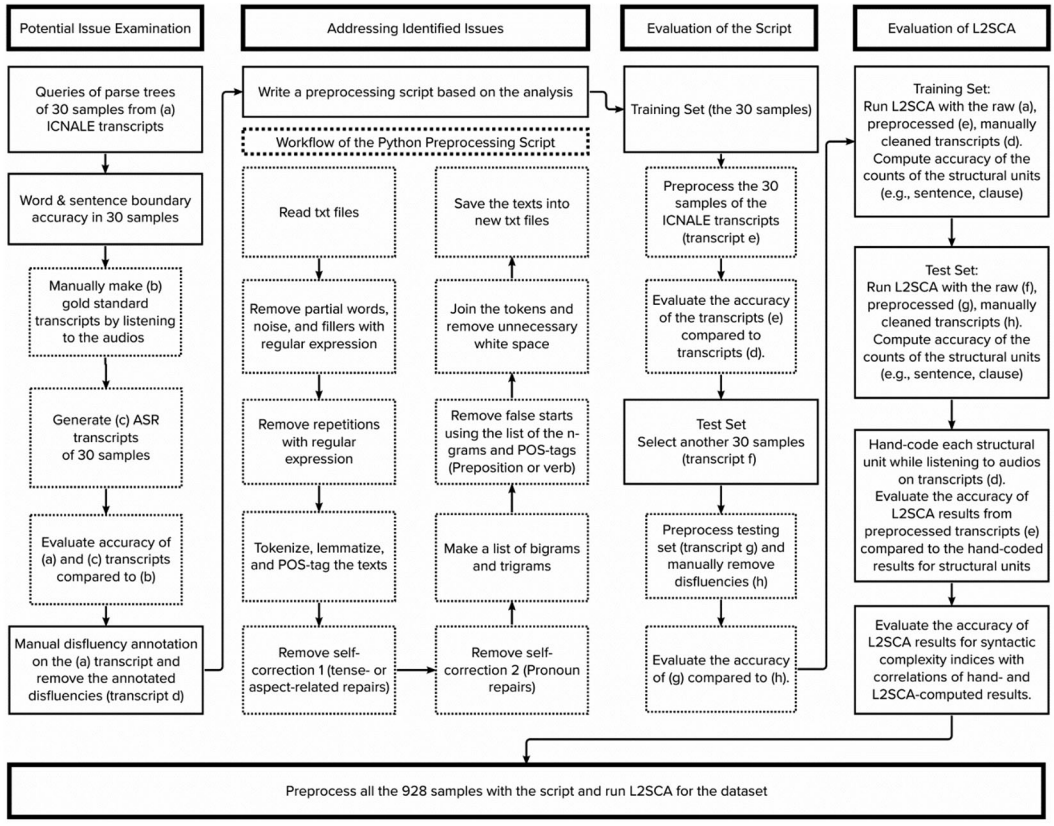

FIGURE 1 Procedure of data preprocessing and L2SCA reliability assessment. ICNALE, International Corpus Network of Asian Learners of English; L2SCA, L2 Syntactic Complexity Analyzer.

Statistical analysis of proficiency, L1 background, and topic effects

To assess the potential effects of proficiency, L1 background, and topic on the SC of L2 English learners' oral production, we performed LME analyses in order to also account for the random variance associated with subject, as each participant contributed four samples (two on each topic). Specifically, we built two models for each SC index, both with the index as the dependent variable; proficiency (operationalized as CEFR level in one model and as TOEIC score in the other), L1 background, topic, and task order (first and second speech) as fixed effects; and participant as a random effect. The random effect structure included by-participant intercepts and by-participant slopes for topic, allowing for individual learners to exhibit random variation in both their intercept and the impact of topic. These choices align with the recommendations by Barr et al. (2013). We conducted the LME analyses using R (R Core Team, 2021) and the lme4 package (Bates, 2010) and reported the coefficients of the predictors (with an alpha value of .05 for statistical significance) and their standard errors, t values, and p values following the recommendations by Mirman (2017). Furthermore, we used the MuMIn package (Nakagawa & Schielzeth, 2013) to calculate two effect size measures for each model—namely, marginal R^2, which is the variance accounted for by fixed effects, and conditional R^2, which is the variance explained by the fixed and random effects combined. The residuals of all models were plotted using the qqnorm(), qqline(), and hist() functions to ensure that there was no heteroscedasticity. All plots were generated using the effects package (Fox, 2003). Additional multiple comparisons were conducted using the emmeans package (Lenth, 2018) with Tukey correction to avoid Type I errors.

RESULTS

Data preprocessing and the reliability of the L2 Syntactic Complexity Analyzer on L2 spoken data

Parse tree queries

As mentioned earlier, prior to evaluating the reliability of L2SCA on L2 spoken data, we performed three types of analysis of 30 speech samples randomly selected from our dataset to identify potential issues of L2 spoken data for L2SCA that could be addressed in preprocessing steps. In the first analysis, we queried the parse trees of the samples using the Tregex patterns in L2SCA to identify issues with structural unit identification. The most prominent problem that we observed in this analysis had to do with clause and dependent clause counts. Specifically, the speech samples included many repetitions of or restarts with a subject and a verb that inflated such counts by L2SCA. In Example 1, because the subject and verb were repeated in "if they want if they want," L2SCA counted three clauses and two dependent clauses in the sentence. Such repetitions were the most frequent features that consistently affected the counts of clauses, dependent clauses, verb phrases, coordinate phrases, and complex nominals by L2SCA. Notably, fragments with no overt verbs (e.g., "okay," "the more, the better"), which are highly frequent in speech, have been usually excluded in traditional definitions of T-units, for which reason researchers have suggested the use of other units of analysis for speech such as the AS-unit (Foster et al., 2000) and C-unit (Pica et al., 1989), which include such pragmatically meaningful fragments. L2SCA attempts to bypass this problem by counting punctuated sentence fragments as T-units, thus including both phrasal C-units and verbless fragments such as those in Examples 2 and 3. The frequency of phrasal C-units was low, likely because the speech samples came from monologues instead of dialogues, in which such C-units are more frequent. Given that the T-units defined by L2SCA were largely equivalent to AS-units, we treated the T-units it identified as AS-units and evaluated its accuracy for AS-unit identification against our manual coding of AS-units of the speech samples.

EXAMPLE 1
(L1 Japanese, B2)
So if—if they want—if they want to do something they must earn money by themselves.

EXAMPLE 2
(L1 Chinese, B1)
However, earning pocket money on his own account that not only can—not only can teach college students how to handle his finance but also can help (…)

EXAMPLE 3
(L1 Korean, A2)
Not duty.

Word recognition and sentence boundary accuracy of ICNALE transcripts

To assess the accuracy of word recognition and sentence boundary marking in the ICNALE transcripts, we manually segmented the 30 transcripts into words and sentences following the criteria established by Foster et al. (2000) and Hwang et al. (2020) for speech transcription while listening to the accompanying audio files. Foster et al. (2000) defined an AS-unit as a single utterance composed of an independent clause or subclausal unit and any related subordinate clauses, with the inclusion

TABLE 3 Word recognition and sentence boundary accuracy of the International Corpus Network of Asian Learners of English and automated speech recognition-generated transcripts of the 30 speech samples.

	Word recognition			Sentence boundary			Word recognition and sentence boundary		
	Precision	**Recall**	**F-score**	**Precision**	**Recall**	**F-score**	**Precision**	**Recall**	**F-score**
ICNALE	0.993	0.992	0.992	1.000	0.872	0.924	0.993	0.984	0.989
ASR	0.887	0.894	0.890	0.600	0.937	0.690	0.850	0.896	0.870

of independent subclausal units that are typical of speech being the main difference from the definition of T-units. They also discussed how tricky cases of disfluency features (e.g., repetitions) can be handled in identifying AS-units using intonation and pauses. For instance, they normally considered a coordinate verb phrase as part of the same AS-unit but would treat the first phrase marked by falling or rising intonation followed by a pause of 500 milliseconds or longer as a separate AS-unit. Hwang et al. (2020) used similar criteria for marking sentence boundaries, operationalizing a sentence as an utterance unit clearly distinguished from the previous unit by a relatively long pause (around 500 milliseconds) along with a clear shift in content. They also recommended using the pitch contour to disambiguate cases where two finite clauses are connected by a common conjunction (e.g., "and," "or," and "so"). Although their participants were young English learners, their criteria are also applicable to speech produced by beginning-level adult English learners in our dataset, as their speech exhibited many similar properties, such as frequent pauses exceeding 500 milliseconds with or without content shift. In comparison, advanced learners tend to exhibit clearer utterance boundaries that are discernible through pauses and pitch contour. Following Hwang et al.'s (2020) recommendations, we considered pause length, content shift, and pitch contour together in determining sentence boundaries in the transcripts. For example, if the word immediately preceding a conjunction had a falling pitch followed by a pause of 500 milliseconds or longer, and if there was also a clear content shift in the second clause, the two clauses were then treated as two separate sentences. In addition to our manually segmented transcripts, we also automatically generated transcripts from the 30 audio files accompanying the 30 speech samples using an ASR system (i.e., Sonix) to investigate potential issues in employing fully automated analyses of speech data.

The accuracy of the original ICNALE transcripts was fairly high for both word recognition (F-score = 0.992) and sentence boundary (F-score = 0.924), as shown in Table 3. Meanwhile, as indicated by the recall for word boundaries (0.872), the ICNALE transcripts treated multiple sentences as a single sentence in a small number of cases. In Example 4, the four clauses were treated as one sentence in the original ICNALE transcript but as three sentences by the first researcher, whose decision was based on the falling pitch contour after each sentence along with a pause longer than 500 milliseconds.

EXAMPLE 4
(L1 Korean, B1)
Original: So, if they only—if they only try to study their university study or their university major, they don't have a chance to experience real society, so they experience the real society from the part-time job, so I think part-time job is very important, and part-time job is good experience for us.
Manually punctuated: So, if they only—if they only try to study their university study or their university major, they don't have a chance to experience real society. So they experience the real society from the part-time job. So I think part-time job is very important, and part-time job is good experience for us.

Aside from the few problematic cases such as the one in Example 4, the overall level of agreement in sentence boundary punctuation between the ICNALE transcripts and our manually segmented

transcripts wss satisfactory, and we thus decided not to make any changes to the sentence boundaries in the ICNALE transcripts.

The ASR-generated transcripts showed a high word recognition accuracy (F-score = 0.890) but a low sentence boundary accuracy (F-score = 0.690), as shown in Table 3. In particular, the sentence boundary precision was low (0.600), largely because the ASR system separated some individual sentences into multiple sentences. In Example 5, the two sentences in a manually segmented transcript were separated into four sentences by the ASR system, along with some transcription errors. The participant's speech included two long pauses (2 seconds between "can't" and "make" and 2.5 seconds between "can't" and "keep"). Such pauses are not uncommon in L2 speech. When not accompanied by a content shift, they alone do not necessarily mark sentence boundaries. In light of the low sentence boundary accuracy of the ASR system, we employed the ICNALE transcripts in our subsequent analysis.

> EXAMPLE 5
> (L1 Japanese, A2)
> **ASR-transcribed**: We can't. Make decent difficulty for smokers. (…) That's why they can't. Keep sale conditions in the smoking.
> **Manually transcribed**: We can't make difficulty for smokers. (…) That's why they can't keep their conditions in smoking.

Manual disfluency annotation

To better understand the types and distribution of disfluency features in the dataset, we annotated all disfluency features in the 30 samples based on Foster et al.'s (2000) definitions and illustrations of different types of disfluency features, including repetitions (e.g., "I think I think"), self-corrections ("I have—I had"), false starts (e.g., "he is I think he wants"), partial words (e.g., "ha–" and "uni–"), and fillers (e.g., "uh" and "um"). A total of 425 such features were found in the 30 transcripts ($M = 14.167$), among which repetitions were the most frequent (41.2%), followed by self-corrections (23.6%), fillers (21.6%), and false starts (13.6%). The only meaningful repetitions found in the transcripts were those of adverbs (e.g., "very very" and "really really") used to emphasize the following adjectives. Based on prosodic information such as pauses and intonation, all other types of repetitions were found to be akin to speech disfluencies. After annotating the 30 transcripts, we manually removed all repetitions that were not found to be meaningful. In addition to these speaker-produced disfluency features, Foster et al. (2000) also noted textual noise (e.g., dashes and superfluous punctuation marks in the transcripts) that arose from the transcription process. We identified such noise in the 30 transcripts as well.

Writing a preprocessing script

Informed by the results of the three analyses presented as well as the practice of Chan et al. (2015)—who manually deleted filled pauses, disfluencies, utterances with no linguistic meaning, and textual noises in the transcripts before analyzing their speech samples with L2SCA—we decided to write a preprocessing script in Python 3 to remove disfluencies from the transcripts. The workflow of the script is presented in Figure 1. The script first reads a transcript as a text file and removes partial words, hyphens, dashes, and other textual noise with regular expression patterns. It then removes repeated words or multiword expressions (ignoring case), with the exception of repetitions of "very," "so," and "really." Finally, it identifies and removes self-corrections and false starts related to verbs, pronouns, and prepositions, which were found to be the most frequent in our dataset. More details about the workflow of the script are provided in the Appendix, along with examples for each step.

The performance of the preprocessing script was evaluated on the 30 random samples (the training set) as well as 30 new samples randomly selected from the dataset representing the proportional

distribution of different proficiency groups (i.e., CEFR levels; the test set). On the training set, the script achieved an F-score of 0.984 (precision = 0.997 and recall = 0.972), leaving 2.8% of disfluency features uneliminated and with only 0.3% of its removals being erroneous. On the test set, the script achieved an F-score of 0.968 (precision = 0.999 and recall = 0.944), leaving 5.6% of disfluency features uneliminated and with only 0.1% of its removals being erroneous. These results show that the script effectively removed the overwhelming majority of disfluency features with extremely rare false positives, as repetitions and patterned repairs that could be reliably identified constituted the most frequent disfluency features in our dataset. The small proportion of uneliminated disfluency features were primarily false starts with no clear pattern, as the preprocessing script primarily addressed patterned self-corrections and false starts.

Evaluation of L2SCA

Following the evaluation of the script, we ran L2SCA on the 30 original (i.e., ICNALE transcripts), preprocessed (i.e., ICNALE transcripts preprocessed by the python script), and manually cleaned transcripts (in the training set mentioned above) separately. Table 4 shows the average counts of each structural unit per transcript in each version, and Table 5 shows the precision, recall, and F-score of the structural units identified by L2SCA from the original and preprocessed transcripts in comparison to those identified manually from the manually cleaned transcripts. These results shed light on the potential differences among the structural units identified by L2SCA from different versions of transcripts.

Finally, the first researcher manually coded all structural units in the 30 manually cleaned transcripts in the training set by listening to the audio files, following Lu's (2010) definitions of the structural units and Polio and Yoon's (2018) guidelines for manual coding for SC. We then evaluated the accuracy of the structural units identified by L2SCA in the preprocessed transcripts against those identified in the manually cleaned transcripts. We further calculated the correlations between the SC indices computed by L2SCA and those computed manually. Tables 6 and 7 show the accuracy of L2SCA's structural unit identification and the correlation coefficients between human and L2SCA-computed SC indices, respectively.

Overall, L2SCA demonstrated a high level of accuracy for structural unit identification on the preprocessed transcripts, with F-scores ranging from 0.906 (for complex T-units) to 0.981 (for sentences and clauses). The T-units it identified matched the AS-units manually identified well (F-score = 0.941), suggesting that the T-units as defined in and extracted by L2SCA could adequately approximate AS-units in spoken data. We therefore treated the T-units counted by L2SCA as AS-units in the subsequent analysis. The correlations between human- and L2SCA-calculated SC indices were all larger than 0.7, with the exception of the correlation for DC/C, which was 0.617. An examination of the transcripts showed that this discrepancy arose largely from issues with overcounting dependent clauses in the preprocessed transcripts. The DC/C index computed by L2SCA for spoken data should thus be treated with more caution, and it may be useful to consider enhancing its accuracy with additional manual preprocessing. All remaining samples in the dataset were subsequently preprocessed and analyzed by L2SCA following this evaluation.

Effects of proficiency, L1 background, and topic on syntactic complexity in L2 speaking models with proficiency operationalized as TOEIC score

Table 8 presents the descriptive statistics of the SC indices by CEFR level. A separate LME model was constructed for each of the 14 SC indices with the index as a dependent variable, TOEIC score, L1 background, topic, and task order as fixed factors, and by-participant random intercepts and by-participant random slopes for topic. The four models fitted with MLC, complex AS-units per AS-unit

TABLE 4 Average structural unit counts in the original, preprocessed, and manually cleaned transcripts.

Set	Type	Words	Sentences	Verb phrases	Clauses	T-units	Dependent clauses	Complex T-units	Coordinate phrases	Complex nominals
Training	Original	90.233	3.800	14.467	11.100	5.400	4.700	3.000	1.233	6.867
	Preprocessed	79.800	3.733	13.500	10.367	4.500	5.033	2.767	1.133	6.300
	Manual	78.033	3.667	13.033	10.000	4.867	4.367	2.800	1.067	5.867
Test	Original	85.867	4.433	14.133	11.733	6.433	4.600	2.933	1.467	7.633
	Preprocessed	76.333	4.367	13.067	10.800	5.867	4.167	3.200	1.167	6.500
	Manual	73.200	4.400	12.200	10.200	5.900	3.967	3.167	1.133	6.233

TABLE 5 Precision, recall, and F-score of structural units identified by L2 Syntactic Complexity Analyzer in the original and preprocessed transcripts in comparison to those identified in the manually cleaned transcripts.

		Training set		Test set	
Unit	**Value**	**Original**	**Preprocessed**	**Original**	**Preprocessed**
Sentences	Precision	0.950	0.984	0.989	1.000
	Recall	0.993	1.000	1.000	0.997
	F-score	0.971	0.992	0.995	0.999
Clauses	Precision	0.914	0.968	0.882	0.959
	Recall	0.993	0.997	0.993	1.000
	F-score	0.952	0.982	0.934	0.979
T-units	Precision	0.838	0.960	0.934	0.979
	Recall	0.946	0.936	0.984	0.959
	F-score	0.888	0.947	0.958	0.968
Dependent clauses	Precision	0.917	0.927	0.843	0.928
	Recall	0.951	0.994	0.966	0.999
	F-score	0.933	0.959	0.900	0.962
Verb phrases	Precision	0.924	0.980	0.883	0.944
	Recall	0.985	0.989	0.999	1.000
	F-score	0.953	0.985	0.937	0.971
Complex T-units	Precision	0.879	0.928	0.947	0.963
	Recall	0.921	0.924	0.861	0.983
	F-score	0.899	0.926	0.900	0.973
Coordinate phrases	Precision	0.846	0.932	0.747	0.934
	Recall	0.972	0.995	1.000	0.920
	F-score	0.904	0.962	0.855	0.925
Complex nominals	Precision	0.901	0.940	0.824	0.942
	Recall	0.984	0.945	1.000	0.980
	F-score	0.941	0.943	0.903	0.960
Mean (F-score)		0.930	0.962	0.923	0.967
SD (F-score)		0.030	0.023	0.043	0.021

(CAS/AS), DC/C, and CN/C as a dependent variable, respectively, were retained, as the residuals of the other models were not approximately normal. In these models, Chinese, the part-time job topic, and the first speech serve as the baselines for L1 background, topic, and task order, respectively. To cross-validate the results of these models, we also constructed additional models with proficiency operationalized as CEFR level with all other elements kept intact. The results of these additional models were very similar to those of the models fitted with TOEIC score.[1]

Table 9 summarizes the output of the four models for MLC (Model 1), CAS/AS (Model 2), DC/C (Model 3), and CN/C (Model 4); Figure 2 plots the fixed effects in each model. Overall, Model 1 achieved a marginal R^2 of 0.041 and a conditional R^2 of 0.255, Model 2 achieved a marginal R^2 of 0.053 and a conditional R^2 of 0.272, Model 3 achieved a marginal R^2 of 0.057 and a conditional R^2 of 0.440, and Model 4 achieved a marginal R^2 of 0.020 and a conditional R^2 of 0.358.

Models 1 (MLC), 2 (CAS/AS), 3 (DC/C), and 4 (CN/C) all showed main effects of TOEIC score, with participants with higher TOEIC scores producing longer clauses ($t = 2.070$, $p = .040$), more

TABLE 6 Accuracy of the structural units identified by L2 Syntactic Complexity Analyzer in the preprocessed transcripts.

Value	Sentences	Verb phrases	Clauses	T-units	Dependent clauses	Complex T-units	Coordinate phrases	Complex nominals
Precision	0.964	0.963	0.966	0.975	0.902	0.905	0.905	0.884
Recall	1.000	0.991	0.996	0.909	1.000	0.908	0.995	0.966
F-score	0.981	0.976	0.981	0.941	0.948	0.906	0.947	0.923

TABLE 7 Correlations between human- and L2 Syntactic Complexity Analyzer-computed syntactic complexity indices.

Index	*r*	95% CI	*p*
Mean length of sentence	0.950	[0.897, 0.976]	<.001
Mean length of T-unit	0.733	[0.506, 0.865]	<.001
Mean length of clause	0.980	[0.973, 0.994]	<.001
Clauses per sentence	0.965	[0.935, 0.985]	<.001
Verb phrases per T-unit	0.730	[0.501, 0.863]	<.001
Clauses per T-unit	0.766	[0.571, 0.886]	<.001
Dependent clauses per clause	0.617	[0.331, 0.800]	<.001
Dependent clauses per T-unit	0.766	[0.560, 0.883]	<.001
T-units per sentence	0.882	[0.765, 0.943]	<.001
Complex T-units per T-unit	0.803	[0.623, 0.902]	<.001
Coordinate phrases per T-unit	0.943	[0.883, 0.973]	<.001
Coordinate phrases per clause	0.979	[0.950, 0.989]	<.001
Complex nominals per T-unit	0.722	[0.489, 0.859]	<.001
Complex nominals per clause	0.913	[0.835, 0.961]	<.001

TABLE 8 Descriptive statistics of the syntactic complexity indices by Common European Framework of Reference for Languages level.

	A2		B1		B2	
Index	Mean	*SD*	Mean	*SD*	Mean	*SD*
Mean length of clause	7.685	2.269	7.964	2.773	8.184	2.268
Mean length of sentence	23.308	14.504	25.928	17.866	30.044	18.496
Mean length of AS-unit	16.692	13.839	16.803	11.556	19.862	13.770
Clauses per sentence	3.137	1.991	3.362	2.316	3.758	2.266
Clauses per AS-unit	2.228	1.747	2.149	1.384	2.449	1.631
Complex AS-units per AS-unit	0.487	0.265	0.516	0.280	0.566	0.253
Dependent clauses per clause	0.377	0.170	0.372	0.167	0.415	0.164
Dependent clauses per AS-unit	0.975	1.271	0.907	0.902	1.127	1.025
Coordinate phrases per clause	0.149	0.171	0.136	0.163	0.147	0.185
Coordinate phrases per AS-unit	0.311	0.386	0.279	0.432	0.364	0.533
AS-units per sentence	1.491	0.706	1.669	1.069	1.652	0.907
Complex nominals per clause	0.608	0.318	0.691	0.384	0.746	0.383
Complex nominals per AS-unit	1.361	1.564	1.487	1.227	1.803	1.337
Verb phrases per AS-unit	2.876	2.502	2.710	1.824	3.166	2.292

TABLE 9 Fixed effects of models with four different dependent variables.

Fixed effect	Estimate	*SE*	*t*	*p*
Model 1, with mean length of clause as a dependent variable				
(Intercept)	7.266	0.575	12.632	<.001
TOEIC score	0.002	0.001	2.070	.040
L1 Japanese	−0.149	0.228	−0.652	.515
L1 Korean	0.091	0.279	0.326	.745
Smoking topic	−0.899	0.166	−5.406	<.001
Second speech	0.115	0.146	0.788	.431
Model 2, with complex AS-unit per AS-unit as a dependent variable				
(Intercept)	0.388	0.063	6.201	<.001
TOEIC score	0.001	0.001	3.532	<.001
L1 Japanese	−0.117	0.025	−4.651	<.001
L1 Korean	−0.127	0.031	−4.128	<.001
Smoking topic	0.031	0.017	1.804	.073
Second speech	−0.019	0.015	−1.268	.205
Model 3, with dependent clauses per clause as a dependent variable				
(Intercept)	0.276	0.041	6.777	<.001
TOEIC score	0.001	0.001	3.453	.001
L1 Japanese	−0.067	0.016	−4.087	<.001
L1 Korean	−0.070	0.020	−3.524	.001
Smoking topic	0.037	0.012	3.147	.002
Second speech	0.001	0.008	0.105	.916
Model 4, with complex nominals per clause as a dependent variable				
(Intercept)	0.400	0.093	4.326	<.001
TOEIC score	0.001	0.001	3.020	.003
L1 Japanese	0.009	0.037	0.248	.804
L1 Korean	0.006	0.045	0.124	.902
Smoking topic	−0.004	0.026	−0.159	.873
Second speech	0.033	0.020	1.653	.099

Abbreviation: TOEIC, Test of English for International Communication.
Note. The L1 baseline was Chinese; the topic baseline was the part-time job topic; the order baseline was first speech.

complex AS-units per AS-unit ($t = 3.532$, $p < .001$), and more dependent clauses ($t = 3.453$, $p = .001$), and complex nominals ($t = 3.020$, $p = .003$) per clause. Models 2 (CAS/AS) and 3 (DC/C) showed main effects of L1 background. Compared to the baseline (L1 Chinese learners), L1 Japanese and Korean learners produced significantly fewer complex AS-units per AS-unit ($t = -4.651$, $p < .001$ for L1 Japanese; $t = -4.128$, $p < .001$ for L1 Korean) and fewer dependent clauses per clause ($t = -4.087$, $p < .001$ for L1 Japanese; $t = -3.524$, $p = .001$ for L1 Korean). Additional multiple comparison tests revealed no significant difference between L1 Japanese and L1 Korean learners. Models 1 (MLC) and 3 (DC/C) also showed main effects of topic. Compared to the baseline (the part-time job topic), the smoking topic elicited significantly shorter clauses ($t = -5.406$, $p < .001$) but significantly more dependent clauses per clause ($t = 3.147$, $p = .002$). No main effect was found for task order.

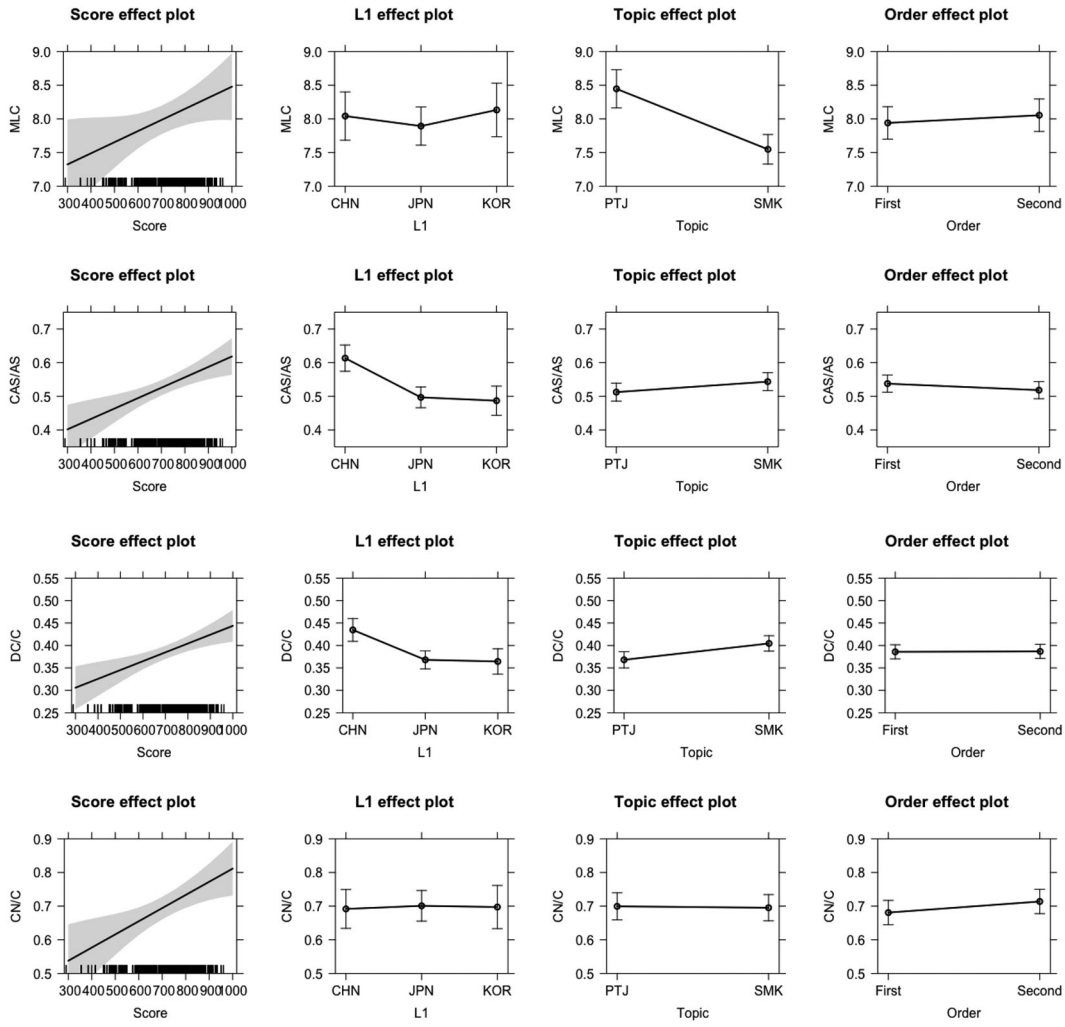

FIGURE 2 Main effects in Models 1 (mean length of clause), 2 (complex AS-units per AS-unit), 3 (dependent clauses per clause), and 4 (complex nominals per clause). CHN, Chinese; JPN, Japanese; KOR, Korean; PTJ, part-time job; SMK, smoking.

Models with L1 speaker data added

To assess the differences in SC between L2 learners and L1 English speakers, we constructed eight LME models with the L1 speaker data added. In Models 5 (MLC), 6 (CAS/AS), 7 (DC/C), and 8 (CN/C), we treated L1 English status as a proficiency level. These models thus had the SC measure as a dependent variable; proficiency (A2, B1, B2, and L1 English status), topic, and task order as fixed effects; and by-participant random intercepts and by-participant random slopes for topic. The effects of proficiency in these models are plotted in Figure 3. Overall, Model 5 achieved a marginal R^2 of 0.044 and a conditional R^2 of 0.294, Model 6 achieved a marginal R^2 of 0.101 and a conditional R^2 of 0.346, Model 7 achieved a marginal R^2 of 0.097 and a conditional R^2 of 0.470, and Model 8 achieved a marginal R^2 of 0.017 and a conditional R^2 of 0.397.

Models 6 (CAS/AS), 7 (DC/C), and 8 (CN/C) showed main effects of proficiency operationalized as CEFR level and L1 English status, while Model 5 (MLC) did not with no significant difference

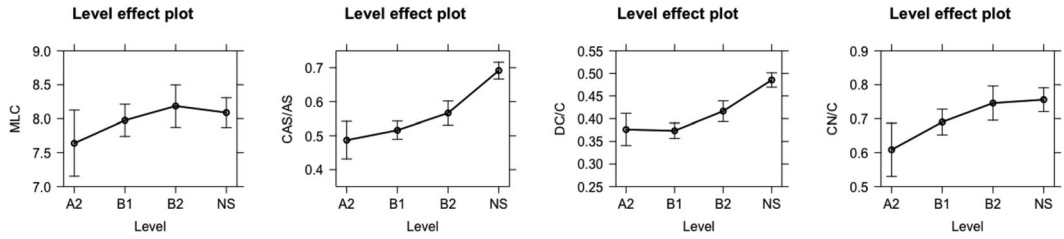

FIGURE 3 Effects of proficiency in Models 5 (mean length of clause), 6 (complex AS-units per AS-unit), 7 (dependent clauses per clause), and 8 (complex nominals per clause). NS, L1 English status.

Note. A2, B1, and B2 are levels in the Common European Framework of Reference for Languages.

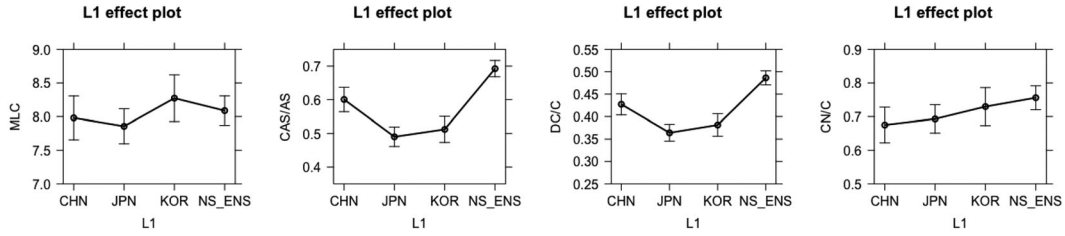

FIGURE 4 Effects of L1 background in Models 9 (mean length of clause), 10 (complex AS-unit per AS-unit), 11 (dependent clauses per clause), and 12 (complex nominals per clause). CHN, Chinese; JPN, Japanese; KOR, Korean; NS_ENS, L1 English.

found in multiple comparisons either. Compared to the baseline (A2 learners), L1 English speakers produced significantly more complex AS-units per AS-unit ($t = 6.619$, $p < .001$) as well as more dependent clauses ($t = 5.523$, $p < .001$) and complex nominals ($t = 3.357$, $p < .001$) per clause. Multiple comparisons further revealed that B1 and B2 learners produced significantly fewer complex AS-units per AS-unit ($t = -9.427$, $p < .001$ for B1; $t = -5.662$, $p < .001$ for B2) and fewer dependent clauses per clause ($t = -9.404$, $p < .001$ for B1; $t = -4.873$, $p < .001$ for B2) than L1 English speakers.

In Models 9 (MLC), 10 (CAS/AS), 11 (DC/C), and 12 (CN/C), we treated L1 English status as an L1 background. These models thus had the SC measure as a dependent variable and L1 background (L1 Chinese, L1 Japanese, L1 Korean, and L1 English), topic, and task order as fixed effects, with by-participant random intercepts and by-participant random slopes for topic. The effects of L1 background in these models are presented in Figure 4. Overall, Model 9 achieved a marginal R^2 of 0.044 and a conditional R^2 of 0.293, Model 10 achieved a marginal R^2 of 0.116 and a conditional R^2 of 0.347, Model 11 achieved a marginal R^2 of 0.107 and a conditional R^2 of 0.471, and Model 12 achieved a marginal R^2 of 0.010 and a conditional R^2 of 0.397.

Models 10 (CAS/AS), 11 (DC/C), and 12 (CN/C) showed main effects of L1 background, while Model 9 (MLC) did not, with no significant difference found in multiple comparisons either. Compared to the baseline group (L1 Chinese learners), L1 English speakers produced significantly more complex AS-units per AS-unit ($t = 4.121$, $p < .001$) as well as more dependent clauses ($t = 4.110$, $p < .001$) and complex nominals ($t = 2.486$, $p = .013$) per clause. Multiple comparisons further revealed that L1 Japanese and Korean learners both produced significantly fewer complex AS-units per AS-unit ($t = -10.489$, $p < .001$ for L1 Japanese; $t = -7.722$, $p < .001$ for L1 Korean) and fewer dependent clauses per clause ($t = -9.831$, $p < .001$ for L1 Japanese; $t = -6.965$, $p < .001$ for L1 Korean) than L1 English speakers.

DISCUSSION

The results of our first RQ revealed a number of issues in L2 English speech samples that may affect the accuracy of automated SC analysis and that can be systematically addressed using a preprocessing script. First, the original transcripts in the dataset and the ASR-generated transcripts indicated some challenges posed by L2 speech on human transcribers and ASR systems. Factors such as extended pauses, unintentional pitch falls, and accented utterances typical in L2 speech could lead to inaccuracies in word recognition and sentence boundary marking. In our dataset, the ASR system in some cases separated a single sentence into multiple ones, a problem that would affect subsequent SC analysis. Recent advances in ASR research have shown much progress in tackling the recognition of accented speech (e.g., Jain et al., 2018). However, there remains a dearth of research on addressing two particular challenges associated with L2 speech—namely, the identification and proper handling of unusually long pauses within a single meaningful unit and the accurate interpretation of falling pitches that do not necessarily signal the end of a sentence. Further efforts are thus necessary to improve the capability of ASR systems to handle such features of L2 speech. Second, removing or normalizing disfluency features in L2 speech samples decreases parsing errors and helps avoid counting repeated structures multiple times. For the eight structural units used to calculate the 14 SC indices, L2SCA achieved a high level of reliability on the 30 preprocessed L2 speech samples, as demonstrated in the accuracy of the structural units identified by L2SCA and the correlations between human- and L2SCA-calculated SC indices. As Polio and Yoon (2018) and Alexopoulou et al. (2021) indicated, it is good practice to assess the reliability of automated linguistic analysis tools on the researchers' dataset at hand; this is useful not only for tools whose developers have not reported reliability but also for those whose developers have reported reliability on datasets that may differ from the dataset at hand in substantial ways. Our study constitutes an early attempt to evaluate and improve the reliability of automated SC analysis tools on L2 speech data. Meanwhile, it should be noted that our findings cannot yet be taken as evidence that L2SCA will perform with comparable levels of reliability on all types of L2 speaking data. Prior research that performed SC analysis of L2 spoken data has largely relied on manual analysis (e.g., Hwang et al., 2020; Mostafa et al., 2021). Chan et al. (2015) analyzed their L2 speech samples using L2SCA after manually removing disfluencies, but their sample size was relatively small. A major reason for the limited use of automated tools for L2 speaking SC analysis may well be that existing tools have not systematically integrated capabilities for principled treatment of disfluency features in ways suggested by Foster et al. (2000). By confirming the reliability of L2SCA on L2 speech samples following an automated preprocessing procedure, we were able to apply the tool to a much larger sample.

In addition to our methodological contribution, the results of our second RQ shed useful light on the effects of two learner-related variables (i.e., proficiency and L1 background) and one task-related variable (i.e., topic) on L2 speaking SC. Our results revealed significant cross-proficiency differences in four measures for which we built LME models—namely, MLC, CAS/AS, DC/C, and CN/C—with all four increasing significantly with TOEIC scores. It should be noted that an increase in a SC feature does not always point to progress or development but may instead indicate the expansion of the repertoire of syntactic structures that learners can use to convey their ideas (Michel et al., 2019; Ortega, 2003; Pallotti, 2009). Meanwhile, MLC, a production length measure indicative of clausal elaboration, and CN/C, a measure of phrasal complexity, have both been shown in L2 writing research to increase with proficiency (e.g., Lu, 2011). Our findings thus suggest that a similar co-growing relationship between these aspects of SC and proficiency may exist in L2 speaking and L2 writing. Finite subordination, measured by DC/C and CAS/AS here, has been shown by some researchers to be primarily reflective of intermediate proficiency in L2 writing and to be subject to a trade-off effect with phrasal complexity at advanced proficiency levels (e.g., Bardovi-Harlig, 1992; Lu, 2011; Norris & Ortega, 2009; Paquot, 2019). Others have also claimed that finite subordination is more characteristic of speaking than of writing (e.g., Biber et al., 2011). Our findings indicated significant increases in

finite subordination in L2 speaking. No trade-off among the four aspects of SC emerged in our analysis. Although we did not include advanced learners, our analysis of L1 speaking data showed that L1 English speakers employed significantly more finite subordination than upper intermediate learners. Vercellotti's (2019) analysis of the longitudinal development of SC in adult instructed ESL learners from low-intermediate to low-advanced levels similarly did not reveal any trade-off effect between finite subordination and other aspects of SC, and she called for research with less proficient and more advanced students to better understand the pattern of SC development in L2 speaking. Our inclusion of high-beginning learners and L1 English speakers usefully complements her analysis. Taken together, our and her findings suggest that the increase in finite subordination may span over a much larger proficiency band in L2 speaking than reported in L2 writing research. It will be useful for future research to include advanced learners and to explore the specific types of dependent clauses that discriminate different proficiency levels in L2 speaking.

Regarding the effect of L1 background, L1 Chinese speakers produced significantly more complex AS-units per AS-unit and finite dependent clauses per clause than L1 Japanese and L1 Korean learners. Furthermore, L1 English speakers produced more complex AS-units per AS-unit and finite dependent clauses per clause than all three L2 groups. Lu and Ai (2015) did not include L1 Korean writers in their analysis, and our results pertaining to L1 Chinese, Japanese, and English speakers mostly align with their results. Similar to what we found here, they reported that L1 Japanese writers demonstrated the lowest level of SC (including MLC, CN/C, and DC/C) among the eight groups they considered. Their L1 Chinese writers also produced significantly fewer finite dependent clauses per clause and a comparable number of complex nominals per clause as L1 English writers, and there was no significant difference in mean length of clause between L1 Chinese and L1 English writers. Taken together, these results suggest three possibilities: First, for beginning and intermediate L1 Japanese EFL learners, SC tends to be low in both spoken and written production in comparison to L1 English speakers and learners from other L1 backgrounds. Second, the syntactic similarity between Japanese and Korean, both of which are SOV (subject-object-verb) languages, might have contributed to the comparable levels of SC between L1 Japanese and L1 Korean speakers, but further research is needed to verify this connection. Finally, L1 background does not appear to affect the clause length or the use of complex nominals as much as it affects other aspects of SC in L2 production, as the only significant difference observed in Lu and Ai's study was between L1 Japanese/Tswana writers and L1 English writers. However, given that the models fitted with L1 background that included the L1 English group did not control for learners' proficiency level and that the L1 Korean group lacked data from A2 learners, we should exercise caution in drawing definitive conclusions.

The two LME models built with MLC and DC/C as a dependent variable demonstrated a significant topic effect on L2 speaking SC, in line with results from L2 writing studies (e.g., Atak & Saricaoglu, 2021; Yang et al., 2015; Yoon, 2017). Specifically, the part-time job topic elicited significantly longer clauses but significantly fewer dependent clauses per clause than the smoking topic. A close examination of the 30 samples in the training set showed that the smoking topic elicited more conditional "if" clauses than the part-time job topic—as writers tended to imagine third-person smokers in restaurants but position themselves as part-time job seekers or workers—while the two topics elicited comparable numbers of other types of dependent clauses (e.g., "because," "so," and "as" clauses). In his analysis of the effect of the same two topics on L2 writing SC, Yoon (2017) reported similar results for MLC, but he also reported significantly higher finite subordination, measured using clauses per T-unit, for the part-time job topic. He argued that the higher SC for the part-time job topic was likely due to its greater familiarity and relevance to the learners than the smoking topic. In light of the partial discrepancy between our results and Yoon's (2017) results and previous claims regarding the more prominent role of finite subordination in speaking than in writing, we posit an interaction between topic and production mode. A more detailed investigation of this interaction using both L2 speaking and writing data would be desirable. Meanwhile, as mentioned earlier, Atak and Saricaoglu (2021) argued that topics that are more impersonal rather than more familiar or relevant elicit greater L2 writing SC. As they used a different set of topics to arrive at this conclusion, it would appear useful to

systematically examine the interaction between different aspects of topics using both L2 writing and speaking data.

Our findings have useful implications for L2 speaking research, assessment, and pedagogy. First, methodologically, the results of our reliability evaluation confirmed the possibility of adopting L2SCA for computing holistic SC indices on transcribed L2 English speech samples. To enhance the accuracy of the analysis, however, it is necessary to prepare the samples carefully. Specifically, disfluency features whose removal does not affect the integrity of SC analysis should be removed either manually (e.g., saved as a separate version during the transcription process) or using a preprocessing script. Second, the capacity to automate preprocessing and SC analysis of corpora of L2 spoken data could substantially expand the scope and scale of L2 speaking SC research, as illustrated by our larger scale analysis of the effects of proficiency, L1 background, and topic on L2 speaking SC than previous analyses. Our specific findings on the effects of L1 background and topic on L2 speaking SC call for careful consideration of such effects in future L2 speaking SC research. More broadly, our analysis might also inspire new areas of inquiry in L2 speaking research such as the analysis of disfluency patterns across L2 proficiency levels and automated sentence boundary detection for L2 speech, to name but two examples. Third, our findings could inform L2 speaking assessment in several ways. The specific measures found to co-increase with proficiency scores (i.e., MLC, CAS/AS, DC/C, and CN/C) in our dataset could constitute useful candidate features for assessing L2 English speaking proficiency and for developing automated speech scoring systems (e.g., Yoon et al., 2020). Additionally, our findings on the effects of proficiency, L1 background, and topic on L2 speaking SC could inform the design of L2 speaking test tasks (e.g., in terms of topic selection) and the integration of SC features in nuanced analytical rubrics (e.g., in terms of the level of SC expected for different proficiency levels). Furthermore, the capacity to automate L2 speaking SC analysis could also be leveraged to efficiently and objectively diagnose L2 speakers' areas of strength or weakness in SC in comparison to other speakers. Finally, our findings can also inform L2 speaking pedagogy in a few ways. For one thing, the results on the ways in which the SC measures vary among L2 English speakers with different proficiency could suggest specific areas of SC for L2 English speaking teachers to pay attention to as they help their students expand their repertories of syntactic structures in their spoken production and improve their speaking proficiency in general. Specifically, higher proficiency L2 speakers and L1 speakers were both found to produce more dependent clauses, complex nominals, and complex AS-units, and higher proficiency L2 speakers were also found to produce longer clauses. Based on these quantitative trends, teachers could find ways to help lower proficiency speakers use different types of dependent clauses and complex nominals to convey their ideas more effectively. For another thing, our findings on the effects of L1 background and topic on L2 speaking SC confirm the importance for L2 English speaking teachers to offer opportunities for their learners to engage with diverse topics and tasks and to take their learners' L1 backgrounds into consideration in monitoring their progress, gauging their areas of strength and weakness, and personalizing instruction.

The current study has several limitations that can be addressed in future research. First, we only evaluated one ASR system in the current study. Future research endeavors could provide a more comprehensive assessment of the accuracy of state-of-the-art ASR systems for L2 speech transcription and identify viable options for L2 speaking research. Relatedly, interdisciplinary collaboration between SLA researchers and computational linguists would hold great promise in the development of ASR systems tailored for L2 speech analysis (Meurers & Dickinson, 2017), as indicated by recent success in the realm of syntactic parsing resulting from such collaboration (e.g., Berzak et al., 2016; Kyle et al., 2022). Second, the preprocessing script adopted a pattern-finding approach and left a small proportion of disfluency features unaddressed. This limitation may have contributed to L2SCA's relatively low accuracy in computing the DC/C index from the preprocessing data. Future research could adopt a data mining approach to handle unpatterned self-corrections and false starts in transcripts, particularly if disfluency-annotated L2 speech corpora similar to the Switchboard corpus of L1 speech become available. It would then be desirable to reevaluate the accuracy of the DC/C index computed and

obtain more conclusive results on this index. Additionally, it would also be useful to explore whether new parsing models based on large language models (e.g., spaCy's transformer models) that have been shown to outperform the Stanford Parser (especially on out of domain text types) could be exploited to further improve the overall accuracy of L2SCA. Third, our dataset consisted solely of monologue samples and thus may not represent natural conversational contexts. The speech features of other types of spoken production such as naturally occurring dialogues with more diverse, authentic conversational features can be examined to enhance the preprocessing process and further improve the accuracy of automated analysis. Fourth, the dataset represented a small range of L1 backgrounds (i.e., Chinese, Japanese, and Korean) and topics (i.e., a part-time job topic and a smoking topic), with no data points for A2 level L1 Korean speakers. These constraints could affect the generalizability of our findings. Future research can investigate the generalizability of our findings by analyzing speech samples produced on more diverse topics by learners with more diverse L1 backgrounds and better balance in their proficiency levels (including advanced levels). Fifth, we limited our analysis to holistic SC indices in this early attempt to examine the reliability of automated analysis of L2 speaking SC. Future research can expand the analysis to more fine-grained indices. Finally, while some studies have examined the effects of the interaction between production mode and task complexity on L2 SC (e.g., Vasylets et al., 2017), future research can further investigate the effects of the interaction between production mode and learner-related variables.

CONCLUSION

This study evaluated the reliability of L2SCA for analyzing the SC of transcribed L2 English speech samples and examined the effects of three learner- and task-related variables—namely, proficiency, L1 background, and topic—on L2 speaking SC. Our analyses identified several issues in the speech samples that can be addressed in a preprocessing step to improve the accuracy of L2SCA and yielded good reliability of L2SCA for computing several holistic indices on preprocessed L2 speech samples. Our findings further revealed significant cross-proficiency differences in four holistic SC measures—namely, MLC, CAS/AS, DC/C, and CN/C—that co-grow with proficiency without any trade-off effect. L1 background and topic were both found to significantly affect L2 speaking SC. Our study serves as an initial effort to facilitate the automatic assessment of SC of large-scale L2 English speech samples, and our findings pertaining to the ways in which proficiency, L1 background, and topic may affect L2 speaking SC have useful implications for future research into L2 speaking SC and for L2 speaking assessment and pedagogy.

ORCID

Minjin Kim https://orcid.org/0000-0001-9935-7867
Xiaofei Lu https://orcid.org/0000-0003-2365-2581

ENDNOTE

[1] The models fitted with proficiency operationalized as CEFR level yielded largely similar results as the models fitted with TOEIC score. The same effects of L1 background, topic, and task order were identified. However, CEFR level had main effects only in the models with CAS/AS, DC/C, and CN/C as dependent variables (but not MLC). Compared to the baseline (A2 learners), B2 learners produced more complex AS-units per AS-unit ($t = 3.159$, $p = .002$) and more dependent clauses ($t = 2.446$, $p = .015$) and complex nominals ($t = 2.375$, p = 0.018) per clause. Multiple comparisons showed that B2 learners also produced more complex AS-units per AS-unit ($t = –2.577$, p = .029) and more dependent clauses per clause ($t = –3.114$, $p = .006$) than B1 learners.

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.), *Studies in corpus linguistics* (pp. 249–264). John Benjamins. https://doi.org/10.1075/scl.59.15ai

Alexopoulou, T., Meurers, D., & Murakami, A. (2021). Big data in SLA: Advances in methodology and analysis. In N. Ziegler & M. González-Lloret (Eds.), *The Routledge handbook of second language acquisition and technology* (pp. 92–106). Routledge. https://doi.org/10.4324/9781351117586-9

Atak, N., & Saricaoglu, A. (2021). Syntactic complexity in L2 learners' argumentative writing: Developmental stages and the within-genre topic effect. *Assessing Writing*, *47*, 100506. https://doi.org/10.1016/j.asw.2020.100506

Bardovi-Harlig, K. (1992). A second look at T-unit analysis: Reconsidering the sentence. *TESOL Quarterly*, *26*, 390–395. https://doi.org/10.2307/3587016

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D. M. (2010). *lme4: Mixed-effects modeling with R*. Springer.

Berzak, Y., Kenney, J., Spadine, C., Wang, J. X., Lam, L., Mori, K. S., Garza, S., & Katz, B. (2016). Universal dependencies for learner English. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 737–746). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1070

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

Biber, D., Gray, B., & Staples, S. (2016). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics*, *37*, 639–668. https://doi.org/10.1093/applin/amu059

Brezina, V., Gablasova, D., & McEnery, T. (2022). Corpus-based approaches to spoken L2 production: Evidence from the Trinity Lancaster Corpus. *International Journal of Learner Corpus Research*, *5*, 119–125. https://doi.org/10.1075/ijlcr.00008.int

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). John Benjamins. https://doi.org/10.1075/lllt.32.02bul

Bulté, B., & Housen, A. (2018). Syntactic complexity in L2 writing: Individual pathways and emerging group trends. *International Journal of Applied Linguistics*, *28*, 147–164. https://doi.org/10.1111/ijal.12196

Carlsen, C. (2012). Proficiency level: A fuzzy variable in computer learner corpora. *Applied Linguistics*, *33*, 161–183. https://doi.org/10.1093/applin/amr047

Chan, H., Verspoor, M., & Vahtrick, L. (2015). Dynamic development in speaking versus writing in identical twins: Dynamic development in speaking versus writing. *Language Learning*, *65*, 298–325. https://doi.org/10.1111/lang.12107

Chen, L., Zechner, K., Yoon, S., Evanini, K., Wang, X., Loukina, A., Tao, J., Davis, L., Lee, C. M., Ma, M., Mundkowsky, R., Lu, C., Leong, C. W., & Gyawali, B. (2018). Automated scoring of nonnative speech using the *SpeechRater*^{SM} v. 5.0 engine. *ETS Research Report Series*, *2018*, 1–31. https://doi.org/10.1002/ets2.12198

Chen, M., & Zechner, K. (2011). Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech. In D. Lin, Y. Matsumoto, & R. Mihalcea (Eds.). *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 722–731). Association for Computational Linguistics. https://aclanthology.org/P11-1073

Crossley, S. A., & McNamara, D. S. (2012). Detecting the first language of second language writers using automated indices of cohesion, lexical sophistication, syntactic complexity and conceptual knowledge. In S. Jarvis & S. A. Crossley (Eds.), *Approaching language transfer through text classification* (pp. 106–126). Multilingual Matters. https://doi.org/10.21832/9781847696991-005

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*, 354–375. https://doi.org/10.1093/applin/21.3.354

Fox, J. (2003). Effect displays in R for generalised linear models. *Journal of Statistical Software*, *8*, 1–27. https://doi.org/10.18637/jss.v008.i15

Gablasova, D., Brezina, V., & McEnery, T. (2019). The Trinity Lancaster Corpus: Development, description, and application. *International Journal of Learner Corpus Research*, *5*, 126–158. https://doi.org/10.1075/ijlcr.19001.gab

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of Institute of Electrical and Electronics Engineers Computer Society* (pp. 517–520). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/ICASSP.1992.225858

Griffis, D., Shivade, C., Fosler-Lussier, E., & Lai, A. M. (2016). A quantitative and qualitative evaluation of sentence boundary detection for the clinical domain. In *American Medical Informatics Association Summits on Translational Science Proceedings*, *2016* (pp. 88–97). American Medical Informatics Association.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*, 461–473. https://doi.org/10.1093/applin/amp048

Huang, Y., Murakami, A., Alexopoulou, T., & Korhonen, A. (2018). Dependency parsing of learner English. *International Journal of Corpus Linguistics*, *23*, 28–54. https://doi.org/10.1075/ijcl.16080.hua

Hwang, H., Jung, H., & Kim, H. (2020). Effects of written versus spoken production modalities on syntactic complexity measures in beginning-level child EFL learners. *Modern Language Journal*, *104*, 267–283. https://doi.org/10.1111/modl.12626

Ishikawa, S. (2014). Design of the ICNALE-spoken: A new database for multi-modal contrastive interlanguage analysis. *Learner Corpus Studies in Asia and the World*, *2*, 63–75. https://doi.org/10.24546/81006690

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, *29*, 24–49. https://doi.org/10.1093/applin/amm017

Jain, A., Upreti, M., & Jyothi, P. (2018). Improved accented speech recognition using accent embeddings and multi-task learning. In *Proceedings of Interspeech* (pp. 2454–2458). International Speech and Communication Association. https://doi.org/10.21437/Interspeech.2018-1864

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

Klein, D., & Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (pp. 423–430). Association for Computational Linguistics. https://doi.org/10.3115/1075096.1075150

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication*. [Unpublished doctoral dissertation]. Georgia State University. https://doi.org/10.57709/8501051

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

Kyle, K., Eguchi, M., Miller, A., & Sither, T. (2022). A dependency treebank of spoken second language English. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, & T. Zesch (Eds.), *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 39–45). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.bea-1.7

Lenth, R. (2018). *emmeans: Estimated marginal means, aka least-squares means*. R package version 1.4.7. https://CRAN.R-project.org/package=emmeans

Levkina, M., & Gilabert, R. (2012). The effects of cognitive task complexity on L2 oral production. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 171–198). John Benjamins.

Levy, R., & Andrew, G. (2006). Tregex and Tsurgeon: Tools for querying and manipulating tree data structures. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, & D. Tapias (Eds.), *Proceedings of the Fifth International Conference on Language Resources and Evaluation* (pp. 2231–2234). European Language Resources Association.

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. (2023). *Corpus linguistics and second language acquisition: Perspectives, issues, and findings*. Routledge.

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

MacWhinney, B. (2012). The logic of the unified model. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 211–227). Routledge.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

Meurers, D., & Dickinson, M. (2017). Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics. *Language Learning*, *67*, 66–95. https://doi.org/10.1111/lang.12233

Michel, M., Murakami, A., Alexopoulou, T., & Meurers, D. (2019). Effects of task type on morphosyntactic complexity across proficiency: Evidence from a large learner corpus of A1 to C2 writings. *Instructed Second Language Acquisition*, *3*, 124–152. https://doi.org/10.1558/isla.38248

Mirman, D. (2017). *Growth curve analysis and visualization using R*. Taylor & Francis. https://doi.org/10.1201/9781315373218

Mostafa, T., Crossley, S. A., & Kim, Y. (2021). Predictors of English as second language learners' oral proficiency development in a classroom context. *International Journal of Applied Linguistics*, *31*, 526–548. https://doi.org/10.1111/ijal.12358

Murakami, A., & Alexopoulou, T. (2016). L1 influence on the acquisition order of English grammatical morphemes. *Studies in Second Language Acquisition*, *38*, 365–401. https://doi.org/10.1017/S0272263115000352

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*, 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Nation, P., & Beglar, D. (2007). A vocabulary size test. *Language Teacher*, *31*, 9–13.

Norris, J. M., & 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. https://doi.org/10.1093/applin/24.4.492

Ortega, L. (2015). Syntactic complexity in L2 writing: Progress and expansion. *Journal of Second Language Writing*, *29*, 82–94. https://doi.org/10.1016/j.jslw.2015.06.008

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, *30*, 590–601. https://doi.org/10.1093/applin/amp045

Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research*, *31*, 117–134. https://doi.org/10.1177/0267658314536435
Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, *35*, 121–145. https://doi.org/10.1177/0267658317694221
Pica, T., Holliday, L., Lewis, N., & Morgenthaler, L. (1989). Comprehensible output as an outcome of linguistic demands on the learner. *Studies in Second Language Acquisition*, *11*, 63–90. https://doi.org/10.1017/S027226310000783X
Polio, C., & Yoon, H. (2018). The reliability and validity of automated tools for examining variation in syntactic complexity across genres. *International Journal of Applied Linguistics*, *28*, 165–188. https://doi.org/10.1111/ijal.12200
R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/
Révész, A. (2011). Task complexity, focus on L2 constructions, and individual differences: A classroom-based study. *Modern Language Journal*, *95*, 162–181. https://doi.org/10.1111/j.1540-4781.2011.01241.x
Révész, A., Ekiert, M., & Torgersen, E. N. (2014). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics*, *37*, 828–848. https://doi.org/10.1093/applin/amu069
Shi, B., Huang, L., & Lu, X. (2020). Effect of prompt type on test-takers' writing performance and writing strategy use in the continuation task. *Language Testing*, *37*, 361–388. https://doi.org/10.1177/0265532220911626
Vasylets, O., Gilabert, R., & Manchón, R. M. (2017). The effects of mode and task complexity on second language production. *Language Learning*, *67*, 394–430. https://doi.org/10.1111/lang.12228
Vercellotti, M. L. (2019). Finding variation: Assessing the development of syntactic complexity in ESL Speech. *International Journal of Applied Linguistics*, *29*, 233–247. https://doi.org/10.1111/ijal.12225
Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, & complexity*. Second Language Teaching & Curriculum Center, University of Hawaii at Manoa.
Yang, W., Lu, X., & Weigle, S. C. (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
Yoon, H.-J. (2017). Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct multidimensionality. *System*, *66*, 130–141. https://doi.org/10.1016/j.system.2017.03.007
Yoon, H.-J., & Polio, C. (2017). The linguistic development of students of English as a second language in two written genres. *TESOL Quarterly*, *51*, 275–301. https://doi.org/10.1002/tesq.296
Yoon, S.-Y., Lu, X., & Zechner, K. (2020). Features measuring vocabulary and grammar. In K. Zechner & K. Evanini (Eds.), *Automated speaking assessment: Using language technologies to score spontaneous speech* (pp. 123–137). Routledge.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Kim, M., & Lu, X. (2024). L2 English speaking syntactic complexity: Data preprocessing issues, reliability of automated analysis, and the effects of proficiency, L1 background, and topic. *Modern Language Journal*, *108*, 270–296. https://doi.org/10.1111/modl.12907

APPENDIX

Preprocessing of disfluencies in transcripts

Actions	Execution of codes
Partial words, noise, and fillers	Remove words that end with hyphens and end dashes.
Remove partial words, noise, fillers with regular expression	Remove hyphens (-) and end dashes (–) only when they are surrounded by whitespaces. Remove interjections found in disfluency annotations. Remove "[***]" for unheard audio and "…" for incomplete clauses.

(Continues)

Actions	Execution of codes
Example	**Original sentence**: People **ha-** have— **uh—**have have the right to refuse—refusing others do something that may harm themselves [***]…
	Preprocessed sentence: People have have have the right to refuse refusing others do something that may harm themselves.
Repetitions	Find repeated words or phrases (ignoring case) within the boundary of a sentence using regular expressions.
Remove repeated words or phrases with a list and regular expression	The regular expressions exclude repetitions of "very," "really," and "so," which are frequently used to emphasize the following adjectives.
Example	**Original sentence**: People have **have** have the right to refuse refusing others do something that may harm themselves.
	Preprocessed sentence: People have the right to refuse refusing others do something that may harm themselves.
Self-correction 1	Tokenize, lemmatize, and part-of-speech tag the transcript.
Remove reparandum related to tense or aspect repairs	When two consecutive verbs are inflectional forms of the same verb, remove the first verb.
Example	**Original sentence**: People have the right to **refuse** refusing others do something that may harm themselves.
	Preprocessed sentence: People have the right to refusing others do something that may harm themselves.
Self-correction 2	If two consecutive pronouns are found, remove the first pronoun.
Remove reparandum related to pronoun repair	If the first and third words are pronouns and the second and fourth words are identical, remove the first two words.
Example	**Original sentence**: If you are smoking, **you** **can** it can really make and ruin their experience of having a good dinner or a lunch.
	Preprocessed sentence: If you are smoking, it can really make and ruin their experience of having a good dinner or a lunch.
False starts	Penn Tree bank was customized to distinguish subordinating conjunctions and prepositions.
Remove reparandum related to false starts	Make a list of bigrams and trigrams in the transcript. If a bigram or trigram is repeated in a near distance within a sentence, remove the first occurrence.
Example	**Original sentence**: …but **it's** I think it's difficult to ban smoking in the restaurant…
	Preprocessed sentence: …but I think it's difficult to ban smoking in the restaurant…