Research Article

Yuko Goto Butler* and Yeting Liu

Developmental trajectories of discourse features by age and learning environment

https://doi.org/10.1515/iral-2024-0199
Received June 30, 2024; accepted December 10, 2024; published online January 21, 2025

Abstract: This study investigated the development of discourse features in young learners of a foreign language (YLLs), focusing on their complexity, accuracy, fluency (CAF), and vocabulary. The study also examined the relationship between CAF and communicative adequacy, and the influence of YLLs' socio-economic status (SES) on discourse development. The participants were Grades 5, 8, and 10 learners of English in China (32 students in each grade level) who were selected through a stratified random sampling from a larger project and 15 advanced adult learners of English as a comparative group. They engaged in a story-telling task using a wordless picture book. The participants' communicative adequacy was operationalized as the narrative structure based on story grammar, and 17 discourse features representing CAF were examined across grade levels and SES groups. A series of ANOVA and correlational analyses found that CAF measures generally showed significant differences by grade with some varied patterns reflecting the multidimensional natures of CAF constructs. SES effects appeared in secondary school levels. CAF measures were not interrelated significantly in Grade 5 but showed greater interrelatedness within and across dimensions among students in higher grades. Fluency contributed most to the communicative adequacy measured by story grammar, followed by vocabulary.

Keywords: young learners; complexity; accuracy; fluency; vocabulary; communicative adequacy

1 Introduction

Children learning additional languages in instructional settings (referred to as young language learners, YLLs, hereafter) have grown in number in recent years. Many

***Corresponding author: Yuko Goto Butler**, University of Pennsylvania, 3700 Walnut Street, Philadelphia, PA 19104-6216, USA, E-mail: ybutler@gse.upenn.edu
Yeting Liu, Faculty of Humanities and Social Sciences, City University of Macau, Macau, China

curricula of YLLs emphasize oral language development, especially at the early stages of their language learning, and various oral activities, including narrative activities such as storytelling, have been implemented in YLLs' classrooms (Butler 2025). Narrative activities are popular because narrative skills—an ability to describe events and actions to tell a story—not only help to develop children's oral proficiency in general but also relate to children's later literacy development (Stadler and Ward 2005). Therefore, as Berman (2009) stated, narratives "provide an advantageous site for tracing the long developmental route from emergence to mastery in language acquisition" (p. 355). Despite the popularity of various oral activities in YLLs' classrooms, there is insufficient understanding of how YLLs develop their discourse features in oral language, even though one can expect that information on YLLs' discourse development potentially has profound implications for curriculum and assessment designs. For example, information on children's discourse development can be useful in deciding which types of communicative tasks should be incorporated into learning materials. It can also be used to set benchmarks for oral assessments. The present study, therefore, aims to better understand YLLs' development of discourse features through a storytelling task.

While there are many studies concerning children's oral narratives, both in first language (L1) and early second language (L2) development research (e.g., Bamberg 1997; Berman and Slobin 1994; Hickmann 2003 for L1-learning children: Verhoeven and Strömqvist 2001 for multilingual children in immersion contexts), most of them examined children up to sometime around the age of 10. However, judging by the distinctive differences between adults' and children's narratives (e.g., adults' greater use of the historical present, as reported by Hickmann 2003), one can speculate that children develop discourse skills throughout primary school and beyond (Viberg 2001). Thus, using a cross-sectional design, the present study examined YLLs' narratives throughout the school years (primary, middle, and high school years). Accordingly, YLLs are defined very broadly in this study, including children from primary school to young adolescents in high school.[1] The study also analyzed the discourse features of adult learners to understand developmental trajectories in L2, ranging from primary school students to adults. In order to avoid falling into the *native-speaker fallacy* (Phillipson 1992) in L2 research, very advanced adult L2 learners with the same linguistic background as the YLLs, rather than native speakers, were recruited. These advanced learners served as a model for the L2 learners.

1 The definition of "young language learners" can vary, and it is not uncommon to include adolescents. For example, *Language Teaching for Young Learners*, a well-established academic journal in applied linguistics, includes adolescents as well as children in its definition of "young learners" (https://benjamins.com/catalog/ltyl).

In second language acquisition (SLA) research, in addition to vocabulary knowledge, three aspects of learners' language production—complexity, accuracy and fluency (CAF)—have been extensively examined as a proficiency indicator (Michel 2017). Complexity, accuracy and fluency can be generally understood as "how elaborate a learner's language is," "the extent to which a learner follows the rule system of the target language," and "smoothness and ease of expression" respectively (Bui and Skehan 2018, p. 1). While how to operationalize these concepts varies across studies, the CAF measures have been employed extensively in SLA research both as an indicator of learners' developmental trajectories as well as their task performance (Bui and Skehan 2018). Using a discourse analytic approach (Hsieh and Wang 2019), therefore, this study explored how YLLs develop discourse features, relying on select CAF measures. In response to the call for a better understanding of the relationship between CAF measures and *communicative adequacy* (Pallotti 2009; Révész et al. 2016), the present study also examines how communicative adequacy relates to the CAF measures, using story grammar as a measure. A robust analysis of the CAF measures and communicative adequacy among YLLs is important not only theoretically but also practically. In particular, given the fact that AI-based assessments, which often rely on structure features such as CAF, are increasingly used even among YLLs (e.g., Evanini et al. 2017), it is critical to examine the interrelationships among these variables in order to determine their adequacy (or lack thereof) as indicators for YLLs' L2 development.

Furthermore, given the diversity of learning environments and resource availability among YLLs, this study aims to fill the gap in understanding the effect of YLLs' socio-economic status (SES) on their discourse development, in addition to age. While SES, or social class, has been identified as a significant factor influencing children's academic performance in general, it remains relatively underexplored in SLA research (Block 2014). SES is considered critical, especially among young learners due to its expected long-term influence on their learning (Butler et al. 2018). One can expect that examining L2 development across SES groups using CAF and communicative adequacy may also inform the development of materials and assessments tailored to the needs of YLLs from diverse backgrounds.

2 Background

2.1 Complexity, accuracy, and fluency in SLA

Upon an agreement that L2 learners' proficiency and performance are multi-compositional, SLA researchers studying L2 development and performance have employed CAF measures as "major research variables" (Housen and Kuiken 2009,

Table 1: CAF measures employed in the present study.

	Measures		Definitions & Illustrations
	Measures	**Other terms used**	
Vocabulary	PPVT		Receptive vocabulary knowledge test (note that this measure is not usually part of the vocabulary dimension in CAF previous studies but was included in this study as a global vocabulary measure)
	Total word count (word tokens)	Lexical complexity [lexical diversity] (Michel 2017)	Total word count of the narrative excluding sound fragments (e.g. *chi-chi-chicken* is counted as one word) or help request for the interviewer
	Word types	Lexical complexity [lexical diversity] (Michel 2017)	Unique words in a text
	Adjusted type token ratio	Lexical complexity [lexical diversity] (Larsen-Freeman 2006)	Word types/square root (2*word token)
Syntactic complexity	Number of AS-units	Production volume (Michel 2017)	AS-unit: "A single speaker's utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clause(s) associated with either" (Foster et al. 2000, p. 365).
	Number of clause	Production volume (Michel 2017)	Clauses counted in this study include the main clause, coordinated or subordinate clauses in an AS-unit, and any independent sub-clausal units that may appear un-grammatical but could be expanded into a clause based on the discoursal context (Foster et al. 2000). In our dataset, it appears often in lower grader's speech. (e.g.) it night = it is nighttime.
	Mean length of AS-unit	Overall syntactic complexity (Norris and Ortega 2009)	Total word count/the number of AS units
	Mean length of clause	Sub-clausal complexity (Norris and Ortega 2009)	Total word count/clause number
	Coordinated clauses/AS-unit	Syntactic complexity via coordination (Norris and Ortega 2009)	The mean number of coordinated clauses per AS-unit
	Subordinate clauses/AS-unit	Syntactic complexity via subordination (Norris and Ortega 2009)	The mean number of subordinate clauses per AS-unit

Table 1: (continued)

	Measures		Definitions & Illustrations
	Measures	**Other terms used**	
Accuracy	Error-free clause percentage	Global measures (Michel 2017)	Percentage of clauses without any forms of errors
	Correct verb form percentage	Specific measures (Michel 2017)	Percentage of verbs free from errors in tense, aspect, modality and subject-verb agreement
Fluency	Speech rate	Overall fluency	Total word count/Total speech time
	Pruned speech rate	Speed fluency (Tavakoli and Skehan 2005)	Pruned word count/Total speech time Pruned word count: total word count minus interviewer's intervention and dis-fluency features such as false start, repetition, self-correction and filled pause
	False start	Breakdown fluency (Tavakoli and Skehan 2005)	A false start is "an utterance which is begun and then either abandoned altogether or reformulated in some way" (Foster et al. 2000, p. 368) (e.g.) the animals {is so(.)} is very surprised.
	Repetition	Breakdown fluency (Tavakoli and Skehan 2005)	Repetition is where the speaker repeats previously produced speech (Foster et al. 2000, p. 368)
	Self-correction	Repair fluency (Tavakoli and Skehan 2005)	A self-correction is an effort to make a "structural change" (Foster et al. 2000, p. 368) after the speaker realizes an error is produced. (e.g.) the fox and the chicken hide in the cave, um, {inside a} inside the hill. In this study, the speaker's correction initiated by the interviewer is also counted.
	Filled pause	Breakdown fluency (Tavakoli and Skehan 2005)	A filled pause is a pause filled with interjections (e.g., *oh*, *uh*, *well*, *hmm*)

p. 461). Despite the popularity of CAF, however, construct definitions of CAF lack clear theoretical justifications. In practice, various measures have been used in each of the dimensions (i.e., complexity, accuracy and fluency), which in turn led to heated discussions concerning what exactly is captured by these various measurements. Certain aspects of vocabulary knowledge, referred to as *lexical complexity*, such as word types, word tokens, and type-token ratios, are often considered part of complexity. Norris and Ortega (2009) stated that "complexity, accuracy, and fluency are each quite complex subsystems with multiple parts, and trying to get a good look at all of the elements that constitute any one of these constructs is a major measurement endeavor" (p. 556). Some measurements are more general, while others are more specific and developmentally more sensitive (see Table 1 for examples of CAF measures), and the interrelations among these measurements are not entirely clear (Housen and Kuiken 2009). Furthermore, learners' CAF are developmental in nature. It is also subject to change by task characteristics and implementation conditions (Ellis 2009; Skehan 2009) as well as learning contexts (Norris and Ortega 2009). Learners' CAF in L2 are by no means universal or stable constructs that can apply to any learning context.

When it comes to the development of CAF, one might assume that displaying higher complexity, accuracy, and fluency is a sign of more advanced proficiency. However, the picture is more complicated than one might think. Research has shown that these dimensions do not necessarily develop linearly or in tandem (Michel 2017). Skehan's (1998) Limited Attentional Capacity Model predicts a trade-off effect between complexity and accuracy when the task requires higher cognitive demands due to the competition for available cognitive resources. In contrast, Robinson's (2001) Cognition Hypothesis argues that both complexity and accuracy can improve simultaneously if certain conditions are met in task design because increased cognitive demands necessitate more focused attention to language. Existing empirical studies have shown inconclusive results (Jackson and Suethanapornkul 2013; Skehan and Foster 2012), partially due to the varied measurements and research designs employed in previous studies.

Which CAF measures should be used depends on the learners' characteristics and research purposes. Reviewing previous studies on CAF for oral production, Hasnain and Hilder (2024) argued that certain measures might be more suitable for beginners while others would be more appropriate for advanced students. For example, when it comes to complexity, assessing one's use of coordinates is a good predictor for beginners' proficiency, while phrasal complexity works better for intermediate and advanced students (Norris and Ortega 2009). With respect to fluency, articulation rates appear to capture beginners' performance well, whereas repair fluency such as false starts and repetitions seems to be a more appropriate indicator for advanced learners (De Jong et al. 2012). Based on their review, Hasnain

and Hilder (2024) concluded that "there is no fixed way of analyzing second language production using CAF, though undeniably, it remains a 'scientifically valid and informative' (Pallotti 2009) way of measuring language production" (pp. 154). They also suggested that researchers should have more than one measure for a given dimension, both general and specific measures, to obtain a balanced picture of one's proficiency.

Curiously, although communicativeness has been emphasized in language instruction and many existing studies using CAF were conducted during communicative tasks, the relationships between learners' performance assessed by CAF measures and communicative adequacy have been little examined (Pallotti 2009). Pallotti defined communicative adequacy as "the degree to which a learner's performance is more or less successful in achieving the task's goals efficiently" (p. 596). However, SLA researchers do not seem to have an agreed-upon conceptualization for communicative adequacy (Révész et al. 2016).

For studies concerning tasks, Pallotti (2009) suggested that communicative adequacy can be treated as a separate dimension from CAF or as a means to help interpret the CAF results. Rare exceptions of CAF studies on oral production that incorporate communicative adequacy include De Jong et al. (2012), Révész et al. (2016), Ogawa (2022), and Koizumi and In'nami (2024). Focusing on oral fluency among adults, De Jong et al. (2012) found that task complexity influenced fluency measures (breakdown fluency, speed fluency, and repair fluency, following the three types of fluency proposed by Tavakoli and Skehan 2005) and communicative adequacy differently, and the effects varied between L2 and L1 speakers. Révész et al. (2016) found that fluency, especially breakdown fluency, turned out to be the most critical predictor of communicative adequacy among adult L2 learners, while other CAF measures had significant but weaker effects on communicative adequacy. In addition, repair fluency was the only variable that showed different impacts according to learners' proficiency level. Similarly, both Ogawa (2022) and Koizumi and In'nami (2024) found that fluency measures strongly predicted human ratings of Japanese college students' opinion-based monologues and picture-description tasks, respectively, while complexity and accuracy measures played minor roles.

2.2 Developmental trajectory based on the CAF measures among young learners

As discussed above, existing studies employing the CAF measures were predominantly conducted among *adult* L2 learners, and limited information is available for young learners. A handful of studies compared performance between young L2 learners and L1-speaking children, while focusing on different dimensions. Perhaps

because these studies dealt with younger students (pre-school to early primary school students), they often found little or no significant differences between L2 learners and L1 speakers. For example, Vermeer (2000) focused on vocabulary: the breadth and depth of vocabulary knowledge. Composed of a set of studies—one concerned kindergarten children and the other concerned primary school children aged 4 and 7 in the Netherlands—, the study found that, while L1 speakers (monolingual children) received higher scores in both measures than L2 learners (bilingual immigrant children), the relations between the breadth and depth of vocabulary knowledge showed no difference between the two groups. Significant relationships were also found between the possibility of knowing a given word and the frequency of language input in primary school education for *both* monolingual and bilingual children, suggesting the critical role of instruction in vocabulary development regardless of children's linguistic background. From an assessment point of view, Wolf et al. (2017) focused on accuracy and compared linguistic features in assessment tasks between L2 learners and L1 speakers at the K-2 grade levels (5–8 years of age) in the United States. The results of their error analysis indicated that both L2 learners and L1 English speakers made similar types of syntactic, discourse, and content-related errors in all four of the oral production tasks they employed. The results indicated that, at least among these young age groups, bilingual and monolingual students are similarly in the process of developing their English skills.

In task-based research, Sample and Michel (2015) examined the effect of task repetition among six 9-year-old English-as-a-foreign language (EFL) learners in Hong Kong, using select CAF measures. The study found that the students' performance (task completion) improved as they repeated the tasks. Fluency showed consistent improvement across task repetitions, while complexity and accuracy indicated a "mixed picture" (p. 43). Complexity values decreased but with the exception of the number of clauses. No notable trend was observed concerning accuracy. Their correlational analysis supported Skehan's (1998) trade-off hypothesis between complexity and accuracy (and fluency) during the first two task sessions, but the effect disappeared in the third session. As the authors acknowledged, the study was based on a single trial and the sample size was small. García Mayo et al. (2018) also examined the effects of task repetition with a larger number of Spanish learners of English (8–9 and 9–10 year-olds, 120 in total) using CAF measures. The study found positive effects on fluency among the 8–9-year-old group and accuracy among the 9-10-year-old group, while no effect was found in terms of complexity. A complex picture emerged in the effect of task repetition on CAF, interacting with children's age.

The improvement in fluency was also reported in a longitudinal study by Mihaljević Djigunović (2016). The participants were 24 EFL students in Croatia, followed from Grade 5 to Grade 8 (ages 11 to 14) over four years. The students' task

achievements (a picture description task with different content by grade, followed by an individual interview) remained high throughout the four years. While this study's focus was on the changes in motivation and self-concepts on task achievement, vocabulary, accuracy, and fluency were also examined. Note, however, that a single holistic measure was used for each dimension. The measures all showed somewhat non-linear developmental patterns, while fluency had a relatively steady improvement during the four years.

Hsieh and Wang's (2019) study, conducted from an assessment point of view, appears to be the most relevant to the present study. Based on 179 students' scores in the speaking section of the TOEFL Junior Comprehensive Test, which consisted of a picture description and an integrated listening and speaking task, the participants were divided into four proficiency levels. The participants' age range was not specified, except that the test itself was designed for learners older than 11 years of age. Twenty-one measures concerning fluency, grammar (covering both accuracy and complexity), vocabulary, and content were employed. The results indicated that the majority of these measures significantly differentiated students with different proficiency levels; in particular, all the fluency measures showed significant differences among proficiency levels. Concerning complexity, the number of words per clause did not predict the learners' proficiency. Vocabulary measures generally showed higher scores for more proficient students; however, lexical sophistication, measured by word frequency, did not. Based on their large effect sizes, content measures, which were treated separately from the ACF measures, were critical indicators of students' proficiency. Finally, the study found that the task types had significant impacts on grammar, vocabulary, and content, but curiously, not on fluency.

2.3 Story grammar

Considering that task types may influence YLLs' performance measured by CAF (Hsieh and Wang 2019), if communicative adequacy were to be included, one could argue that the measure should be designed specifically for the goal of the given task. Since the present study used a storytelling task based on a series of pictures, *story grammar*, a popular content structural measure for YLLs' narratives, would be a promising candidate.

Researchers have approached story grammar differently, but they all rest on the assumption that narratives have common underlying components. Among various story grammar frameworks, the present study employs the framework developed by Stein and her colleagues (e.g., Stein 1988; Stein and Albro 1997) because it is considered "some of the most careful, systematic, conceptually self-conscious, and

broadly influential investigations of narrative coherence in developmental research" (Nicolopoulou 2008, p. 301). Stein's framework considers that "a good story" is not a simple connection of episodes that are arranged in temporal sequence. The story should be centered around the main protagonists' goal-oriented actions, and it should contain a series of elements in sequence. Such elements include descriptive sequence, action sequence (with temporal relations), reactive sequence (with causal relations): goal-based actions, obstacles, and ending (Stein and Albro 1997, p. 9). Based on the inclusion or exclusion of these elements, the model categorizes stories into eight developmental levels, each presumed to represent an increase in cognitive complexity.

While Stein's story grammar framework has been used widely in developmental studies, it is not free from criticism. One of the major criticisms concerns the fundamental assumption of story grammar frameworks, namely, there are common underlying elements in coherent stories. For example, Nicolopoulou (2008), while acknowledging that the model "certainly captures *some* aspects of children's narrative activities and development" (emphasis in original), it is "simply too restricted" and therefore, "it misses a good deal of what is interesting and complex about children's stories" (p. 305). Other researchers also argued for linguistic and cultural non-uniformity in conceptualizing what constitutes "good stories" (e.g., Lee et al. 2011; Wang and Leichtman 2000). Lee et al. (2011), for example, compared narrative coherence between American and Korean preschool children using Stein's story grammar framework and found that the former outperformed the latter. The authors attributed the result to the differences in teachers' attitudes and practices of narratives in classrooms. Differences were also found in narratives created by children with different socioeconomic or social class backgrounds even within the 'same' linguistic communities (Aksu-Koç 1996; Butler and Zeng 2014). Butler and Zeng (2014) found that mothers' education significantly influenced Chinese EFL students' narratives both in Chinese (L1) and English (L2) at the Grade 4 level, but the effects were not found among Grade 6 and Grade 8 students. Furthermore, the 8-level progression proposed by Stain's framework has been controversial. For example, Level 6 and Level 7 in the framework indicate "goal-based episodes with ending but no obstacle" and "goal-based episodes with obstacles but no ending" (Stein and Albro 1997, p. 9); however, this progression has not been sufficiently validated either theoretically or empirically.

3 Research questions

Based on the discussions above, this exploratory study attempted to answer the following research questions:

RQ1: How do YLLs' features of complexity, accuracy, fluency (CAF), and vocabulary in their oral narratives in the target language (i.e., English in this study) differ across grade levels and learning environments (two distinctively different SES groups)?

RQ2: How do the features of CAF and vocabulary differ between YLLs and advanced L2 adult learners?

RQ3: How do YLLs and advanced L2 adult learners achieve "communicative adequacy" measured by story grammar?

RQ4: What are the relations among all these CAF measures and communicative adequacy?

4 Methods

4.1 Participants

The participants were Grade 5, 8, and 10 EFL students (ages 10–11, 13–14, and 15–16 respectively) in a medium-sized coastal city in China. There were 32 students in each grade level, resulting in 96 students in total. The present study is part of a larger longitudinal project composed of three cohorts of students (primary, middle school, and high school cohorts, $N = 572$). The students in the original project came from two sets of schools located in two distinctively different socioeconomic areas. Although each cohort of students was followed for three years in the original study, the present study employed a cross-sectional research design; each grade level was composed of different students.

The participants of the present study were selected from the larger project as focused group students, based on stratified random sampling while controlling for gender, socioeconomic status (SES) and general proficiency levels. In other words, at each grade level, the groups were composed of an equal number of female and male students, students from schools located in lower and higher SES areas, and students with lower, lower-middle, upper-middle, and higher proficiency in English. The four proficiency levels were determined relative to each grade using multiple measures, including a standardized general proficiency test (from the Cambridge English Language Assessment series), a locally administered standardized achievement test, teacher-made classroom assessments, and teachers' holistic judgments.

All the participants had received English lessons at school since Grade 3 based on the uniformed curricula. However, students in higher SES schools used additional materials along with designated textbooks and engaged in some extra English activities. Narrative activities in L1 were popular in language art classes at the pre-primary and lower primary school levels, and various oral activities including picture-based narrative tasks were occasionally used in English classes across grade levels.

In addition to the Grade 5, 8, and 10 students, 15 advanced adult L2 learners of English participated in the study as a comparative group. The advanced learners were recruited from a Teaching English to Speakers of Other Languages (TESOL) program in a graduate school of education at a university in the United States. They were all international students from China, with Chinese as their L1. They were all in their mid-twenties. They were advanced L2 learners; they all had scores of 110 or higher on the TOEFL iBT test or an average of 7.5 or higher in the four skills on the IELTS when they entered the TESOL program. As mentioned, this group's performance can be considered a goal for the participating YLLs in our study. They are referred to as "TESOL students" hereafter.

4.2 Materials

The students were asked to tell a story based on a wordless picture book "The Chicken Thief" (Rodriguez 2005). In this story, while the animals were having lunch, a fox stole one of the chickens. The chicken's friends chased the fox to get the chicken back. After facing a few obstacles, the animals reached the fox's house. To their surprise, they found that the fox was not a villain and the chicken wanted to be with him. The animals decided to leave them behind in the end. None of the students had seen the book before participating in the study.

Although "Frog, Where Are You?" (Mayer 1969) has been used frequently in L1 narrative studies, in consultation with the participating students' teachers, we decided to use "The Chicken Thief" (Rodriguez 2005) instead, for the following reasons: (1) a pilot study indicated that it was too challenging for the youngest students (Grade 4) in the original larger project to tell a story in English based on "Frog, Where Are You?"; (2) "The Chicken Thief" has very colorful and cute illustrations and was generally well-received by the students in the pilot study; (3) "The Chicken Thief" has a clear storyline with obstacles and an ending, aligning well with Stein's story grammar framework; and (4) its follow-up story, "Fox and Hen Together," was useful to implement in subsequent years in the original longitudinal project.

4.3 Procedures

All the YLLs were asked to tell a story based on "The Chicken Thief," which is composed of 10 pictures (scenes), first in their L1 (Chinese) and then in their L2 (English). Note, however, that the present study only concerned their English narratives (see Butler and Zeng 2014, for a comparison of narratives between YLLs' L1 and L2). The students were allowed to look at all the pictures in advance and take as much time as they wanted before starting to tell a story to a researcher who had spent some time with the students before the study. The storytelling task itself was not timed either. If necessary, the students were allowed to ask the researcher for help with vocabulary. Some Grade 5 students asked the researchers for help but hardly any Grade 8 and 10 students did (The words provided by the researcher were not included in the vocabulary analysis below). After telling a story in both L1 and L2, the students were asked a series of comprehension questions, followed by semi-structured interviews concerning their perception of the narrative task as well as their English study in general. The students' narratives and the successive interviews were all audio-taped and transcribed. In addition, between telling stories in L1 and L2, the students took the Peabody Picture Vocabulary Test (PPVT) in English, a standardized receptive vocabulary test. Although there was much variability across students (and across time within individuals), it took a student approximately 30–40 min on average to complete the entire procedure.

The students' SES-related information, including parental education and cultural capital, was obtained through questionnaires distributed to the students and their parents in the original larger project. The present study utilized this information to ensure that the students in the two sets of schools were indeed distinctively different in SES.

The adult TESOL students were also asked to individually tell a story based on "The Chicken Thief" both in Chinese and English. They also filled out a background survey. They did not, however, take the PPVT due to a logistical difficulty with their schedule.

4.4 Analyses

Consulting previous studies (Hsieh and Wang 2019; Larsen-Freemen 2006; Yuan and Ellis 2003), we used several CAF measures as well as PPVT scores for the analysis. The measures that we used in the study and their brief descriptions are summarized in Table 1.

The students' communicative adequacy in the storytelling task was measured by story grammar. As mentioned, Stein's story grammar framework was composed of

Table 2: Story grammar implemented in this study.

Elements	Scoring	Examples
Sequence (0–2)	0 for descriptive statements only; 1 for stories including action statements, and 2 for stories including reactive statements.	
Goal (0–2)	0 for no goal statement; 1 for an implicit goal statement; 2 for an explicit goal statement.	In the present study, the goal statement is considered explicit if the student employed linguistic devices such as mental state verbs (e.g. *decide*, *want*), causal connectives, temporal connectives, and two-place predicate to signal the relationship between the characters (e.g. *catch*, *rescue*). A simple description of animals running after each other is coded as implicit. e.g. Mr. Bear turned around and found out the chicken was stolen. **So** he hollered "hurry, **let's catch** the chicken thief." (explicit goal) Bear, rabbit are having breakfast, but the fox is take to the kitchen. Fox is running, bear, rabbit and cock are (:) run after him. (implicit goal)
Obstacle (0–2)	0 for mentioning 0–2 obstacles; 1 for 3–4 obstacles; 2 for 5–8 obstacles.	Eight obstacles that correspond to the plot in different pictures (e.g. animals weathered a storm at sea) were pre-identified by researchers.
Ending (0–2)	0 for no ending; 1 for a partial ending; 2 for a full ending.	If the student can rationalize the end of the story with a detailed explanation, the narrative will be assigned a full ending. A simple description without fleshing out a potential cause for the end is coded as a partial ending. e.g. but the(:) ur(:) but the white chicken(:) ur(.)interrupt them, um(:) ur(:) ur she he said that the fox ur is a friend, is a friend, is a friend who he miss (.)missed long long ago. Ur(:) (..) ur now he(:)/f/now he found he found (..)her friend fox, so he, he was very exciting, ur he was very exciting and(.) he(:) and he want to(:) stay(:) stayed with the fox forever. (full ending) The fox and the white chicken stay/wɪs/together (.)/ænə//zə/bear the rabbit and one of the chicken go home. (partial ending)

four elements: sequence, goal, obstacles, and ending (e.g., Stein 1988; Stein and Albro 1997). We gave the students scores based on the framework, with some modifications to fit the specific story implemented in this study, as shown in Table 2.

The coding of the discourse features via CAF measures and the story grammar was conducted separately by different groups of research assistants (six and four assistants, respectively). As explained in Table 2, with respect to story grammar, each narrative was coded for the four elements (i.e., sequence, goal, obstacles, ending) according to Stein's story grammar framework (1988).

Both the CAF and story grammar analyses involved two steps: (1) initial training followed by trial coding on four randomly selected narratives, and (2) independent coding for the rest of the narratives. The initial interrater reliabilities during the trial for the CAF measures and story grammar ranged from 76.92 to 93.23 % and 75 to 91.7 %, respectively. Additional clarifications were made after the training to help the coders deepen their understanding of the coding scheme. During the independent coding stage, each narrative was coded simultaneously by two coders with different coder combinations. Any discrepancies were then discussed and resolved until the coders reached 100 % agreement for all the narratives.

A series of ANOVAs and correlational analyses were performed to identify differences in students' performance measured by CAF and communicative adequacy (i.e., story grammar) by grade levels and SES, as well as the interrelations among these dimensions.

5 Results

5.1 RQ1: CAF and vocabulary by grade levels and SES

First, descriptive statistics are presented for each dimension, as one can see in Table 3. Note that Table 3 includes the results of TESOL students which will be discussed later (RQ 2). Next, a series of two-way ANOVAs (three grade levels and two SES groups) were performed after ensuring the relevant statistical assumptions were met (Table 4).

While there are some variabilities across different variables, generally speaking, the oral performance examined by CAF measures among Grades 8 and 10 was significantly higher than that among Grade 5, while significant differences were not often found between Grade 8 and Grade 10. The exceptions to this trend included AU-unit, and three disfluency measures (false starts, repetition, and filled-pause). The analyses also indicated that significant effects of SES began in Grade 8 and even widened in Grade 10 in most CAF and vocabulary measures.

Table 3: Descriptive statistics.

	Grade 5			Grade 8			Grade 10			TESOL
	All	H-SES	L-SES	All	H-SES	L-SES	All	H-SES	L-SES	All
Vocabulary										
PPVT	35.34	37.94	32.75	48.03	56.50	39.56	58.06	75.69	40.44	n.a.
	(10.11)	(10.06)	(9.78)	(15.51)	(11.47)	(14.59)	(23.71)	(19.60)	(10.74)	
Total word count	195.59	203.31	187.88	276.34	358.63	194.06	252.81	330.5	175.13	429.33
	(74.81)	(72.98)	(78.19)	(163.52)	(178.47)	(94.68)	(146.22)	(167.03)	(58.39)	(201.68)
Word types	49.56	50.81	48.31	75.75	93.63	57.88	74.09	91.88	56.31	129.07
	(13.16)	(11.15)	(15.18)	(30.37)	(29.48)	(18.87)	(36.64)	(44.45)	(11.16)	(43.27)
TTR	2.5	2.54	2.46	3.26	3.56	2.96	3.43	3.82	3.05	4.43
	(0.45)	(0.58)	(0.27)	(0.57)	(0.57)	(0.38)	(0.62)	(0.62)	(0.3)	(0.45)
Complexity										
AS-units	20.53	20	21.06	21.63	26.56	16.69	21.41	27.25	15.56	30.6
	(7.51)	(6.09)	(8.89)	(9.99)	(11.53)	(4.6)	(10.71)	(12.21)	(3.89)	(13.96)
Clauses	23.53	24.06	23	30.5	39.5	21.5	31.56	41	22.13	53.27
	(8.84)	(8.35)	(9.55)	(17.18)	(18.04)	(10.57)	(17.11)	(19.46)	(6.03)	(28.02)
Mean length of AS-unit	9.6	10.18	9.01	12.31	13.35	11.27	11.51	11.82	11.21	14.1
	(1.78)	(1.94)	(1.41)	(2.33)	(2.03)	(2.19)	(2.52)	(2.45)	(2.62)	(1.92)
Mean length of clause	8.39	8.6	8.18	9.65	9.64	9.66	7.91	7.97	7.85	8.25
	(1.61)	(1.99)	(1.12)	(3.82)	(3.94)	(3.84)	(1.01)	(0.77)	(1.22)	(2.01)
Coordinate clauses/AS-unit	0.04	0.07	0.02	0.15	0.21	0.09	0.16	0.17	0.16	0.31
	(0.05)	(0.05)	(0.03)	(0.14)	(0.16)	(0.09)	(0.13)	(0.11)	(0.14)	(0.19)
Subordinate clause/AS-unit	0.12	0.15	0.09	0.29	0.37	0.22	0.29	0.32	0.25	0.51
	(0.1)	(0.12)	(0.08)	(0.18)	(0.17)	(0.16)	(0.15)	(0.19)	(0.11)	(0.2)
Accuracy										
Err-free clause percentage	0.31	0.33	0.3	0.52	0.62	0.43	0.49	0.61	0.37	0.69
	(0.17)	(0.2)	(0.14)	(0.23)	(0.21)	(0.22)	(0.21)	(0.14)	(0.21)	(0.17)

Table 3: (continued)

	Grade 5			Grade 8			Grade 10			TESOL
	All	**H-SES**	**L-SES**	**All**	**H-SES**	**L-SES**	**All**	**H-SES**	**L-SES**	**All**
Correct verb form percentage	0.62	0.67	0.56	0.74	0.8	0.68	0.76	0.86	0.65	0.89
	(0.15)	(0.13)	(0.14)	(0.15)	(0.12)	(0.16)	(0.17)	(0.1)	(0.15)	(0.05)
Fluency										
Speech rate	0.65	0.66	0.64	0.95	1.05	0.85	0.95	1.11	0.79	1.7 (0.3)
	(0.23)	(0.24)	(0.22)	(0.3)	(0.2)	(0.35)	(0.34)	(0.29)	(0.31)	
Pruned speech rate	0.51	0.51	0.5	0.71	0.78	0.65	0.76	0.89	0.64	1.58
	(0.21)	(0.2)	(0.22)	(0.21)	(0.14)	(0.25)	(0.29)	(0.21)	(0.31)	(0.32)
False start	13.28	14.63	11.94	15.53	19.63	11.44	12.94	15.13	10.75	8.6
	(8.54)	(9.48)	(7.55)	(12.2)	(14.35)	(8.12)	(10.29)	(13.35)	(5.51)	(9.18)
Repetition	9.88	11.38	8.38	13.78	15.75	11.81	12.22	16.56	7.88	9.07
	(7.47)	(8.94)	(5.52)	(11.45)	(12.74)	(10.03)	(12.58)	(15.43)	(6.97)	(8.32)
Self-correction	1.47	1.25	1.69	3.63	5.69	1.56	3.25	4.31	2.19	2.47
	(1.72)	(1.61)	(1.85)	(4.01)	(3.75)	(3.18)	(2.72)	(3.01)	(1.97)	(2.45)
Filled pause	13.88	16.56	11.19	18.94	22.31	15.56	19.72	19.44	20	24.4
	(12.07)	(14.06)	(9.38)	(20.9)	(27.28)	(11.59)	(18.23)	(23.86)	(10.84)	(20.69)

Table 4: Two-way ANOVA results.

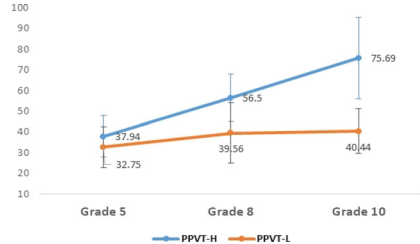
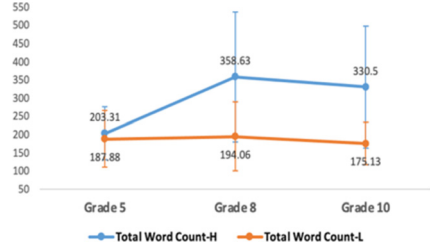
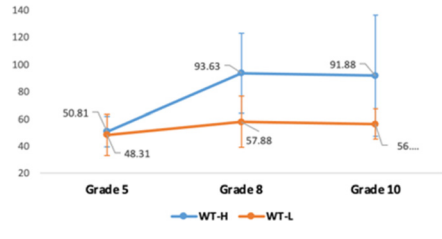
	Main effect of Grade	Main effect of SES	Interaction effect: Grade and SES	Visual presentation
Vocabulary				
PPVT	$F(2, 90) = 23.92$, $p = 0.00$, $\eta^2p = 0.35$	$F(1, 90) = 50.63$, $p = 0.00$, $\eta^2p = 0.36$	$F(2, 90) = 10.59$, $p = 0.00$, $\eta^2p = 0.19$	
Total word count	$F(2, 90) = 3.96$, $p = 0.02$, $\eta^2p = 0.08$	$F(1, 90) = 21.54$, $p = 0.00$, $\eta^2p = 0.19$.	$F(2, 90) = 4.01$, $p = 0.02$, $\eta^2p = 0.08$	
Word types	$F(2, 90) = 11.22$, $p = 0.00$, $\eta^2p = 0.20$	$F(1, 90) = 23.69$, $p = 0.00$, $\eta^2p = 0.21$	$F(2, 90) = 4.78$, $p = 0.01$, $\eta^2p = 0.10$	

Table 4: (continued)

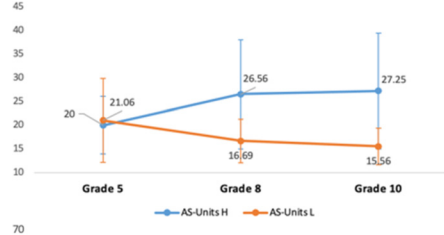
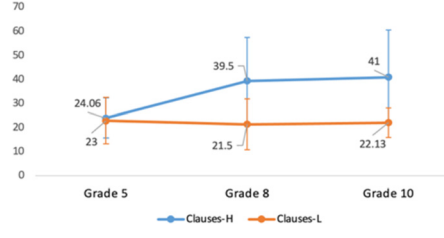
	Main effect of Grade	Main effect of SES	Interaction effect: Grade and SES	Visual presentation
TTR	$F(2, 90) = 34.31$, $p = 0.00$, $\eta^2p = 0.43$	$F(1, 90) = 24.81$, $p = 0.00$, $\eta^2p = 0.22$	$F(2, 90) = 4.58$, $p = 0.01$, $\eta^2p = 0.09$	
Complexity				
AS-units	$F(2, 90) = 0.15$, $p = 0.86$	$F(1, 90) = 15.48$, $p = 0.00$, $\eta^2p = 0.15$	$F(2, 90) = 5.26$, $p = 0.00$, $\eta^2p = 0.11$	
Clauses	$F(2, 90) = 3.61$, $p = 0.03$, $\eta^2p = 0.07$	$F(1, 90) = 22.73$, $p = 0.00$, $\eta^2p = 0.20$	$F(2, 90) = 4.78$, $p = 0.01$, $\eta^2p = 0.10$	

Table 4: (continued)

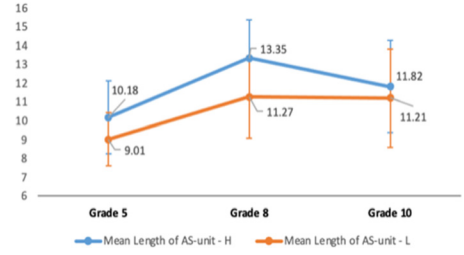
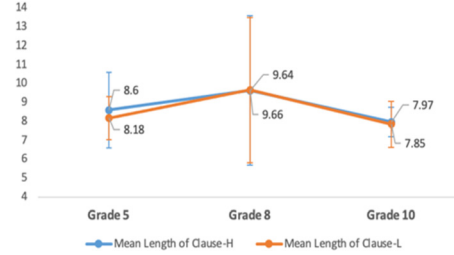
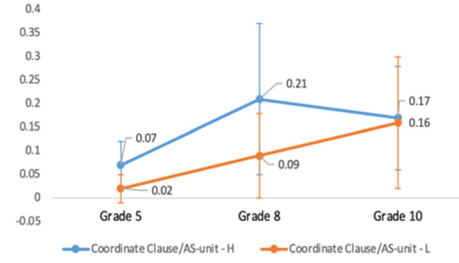
	Main effect of Grade	Main effect of SES	Interaction effect: Grade and SES	Visual presentation
Mean length of AS-unit	$F(2, 90) = 13.50$, $p = 00$, $\eta^2 p = 0.23$	$F(1, 90) = 8.65$, $p = 00$, $\eta^2 p = 0.09$	$F(2, 90) = 0.96$, $p = 0.39$	
Mean length of clause	$F(2, 90) = 4.12$, $p = 0.02$, $\eta^2 p = 0.08$	$F(1, 90) = 0.11$, $p = 0.74$	$F(2, 90) = 0.07$, $p = 0.94$	
Coordinate clauses/ AS-unit	$F(2, 90) = 12.17$, $p = 0.00$, $\eta^2 p = 0.21$	$F(1, 90) = 6.67$, $p = 0.01$, $\eta^2 p = 0.07$	$F(2, 90) = 2.29$, $p = 0.11$	

Table 4: (continued)

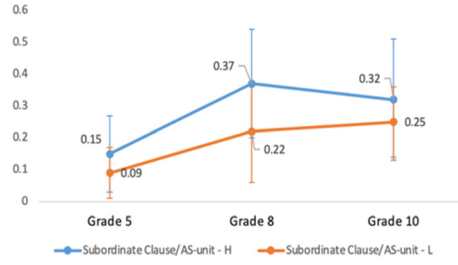
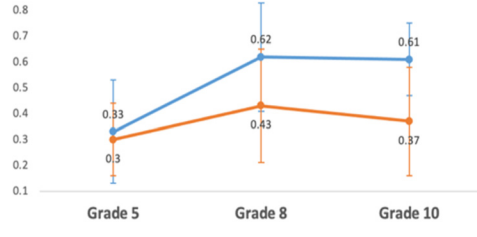
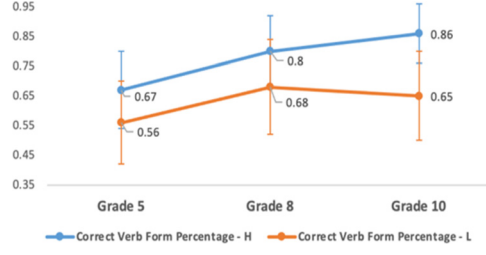
	Main effect of Grade	Main effect of SES	Interaction effect: Grade and SES	Visual presentation
Subordinate clause/ AS-unit	$F(2, 90) = 16.30$, $p = 0.00$, $\eta^2p = 0.27$	$F(1, 90) = 9.68$, $p = 0.00$, $\eta^2p = 0.10$	$F(2, 90) = 1.02$, $p = 0.37$,	
Accuracy				
Err-free clause percentage	$F(2, 90) = 11.34$, $p = 0.00$, $\eta^2p = 0.20$	$F(1, 90) = 14.59$, $p = 0.00$, $\eta^2p = 0.14$	$F(2, 90) = 2.75$, $p = 0.07$	
Correct verb form percentage	$F(2, 90) = 10.24$, $p = 0.00$, $\eta^2p = 0.19$	$F(1, 90) = 27.34$, $p = 0.00$, $\eta^2p = 0.23$	$F(2, 90) = 1.19$, $p = 0.31$	

Table 4: (continued)

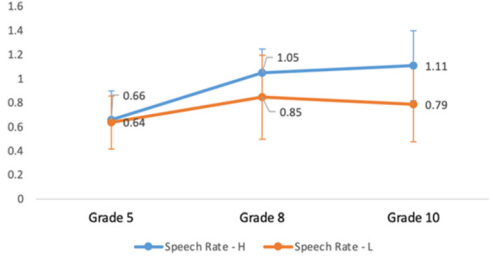
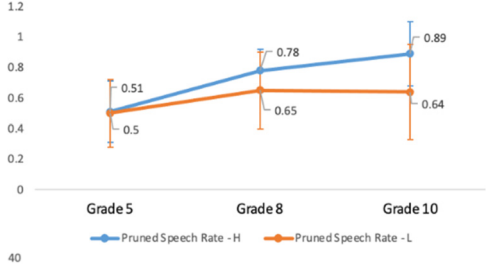
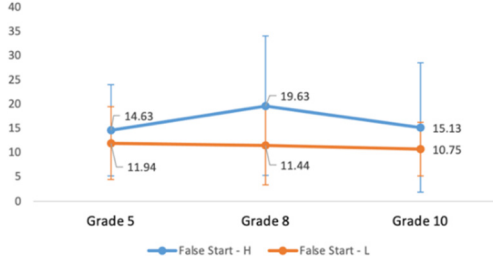
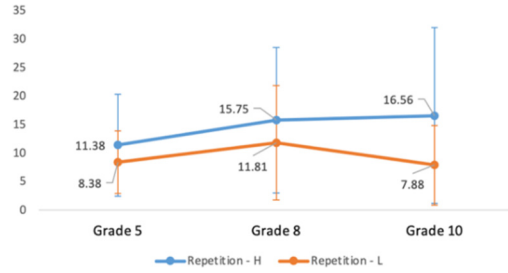
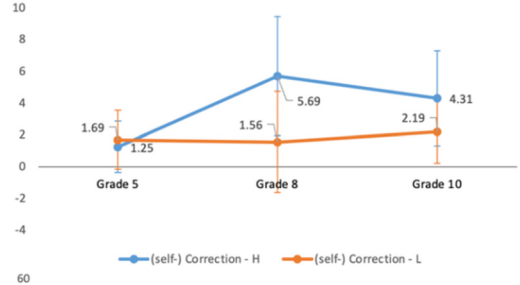
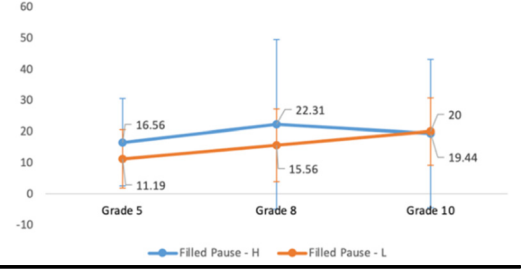
	Main effect of Grade	Main effect of SES	Interaction effect: Grade and SES	Visual presentation
Fluency				
Speech rate	$F(2, 90) = 12.67$, $p = 0.00$, $\eta^2 p = 0.22$	$F(1, 90) = 10.66$, $p = 0.00$, $\eta^2 p = 0.11$	$F(2, 90) = 2.51$, $p = 0.09$	
Pruned speech rate	$F(2, 90) = 11.36$, $p = 0.00$, $\eta^2 p = 0.20$	$F(1, 90) = 7.80$, $p = 0.01$, $\eta^2 p = 0.08$	$F(2, 90) = 2.30$, $p = 0.11$	
False start	$F(2, 90) = 0.61$, $p = 0.66$	$F(1, 90) = 5.93$, $p = 0.02$, $\eta^2 p = 0.06$	$F(2, 90) = 0.61$, $p = 0.55$	

Table 4: (continued)

	Main effect of Grade	Main effect of SES	Interaction effect: Grade and SES	Visual presentation
Repetition	$F(2, 90) = 1.13$, $p = 0.33$	$F(1, 90) = 5.92$, $p = 0.02$, $\eta^2p = 0.06$	$F(2, 90) = 0.68$, $p = 0.51$	
Self-correction	$F(2, 90) = 5.90$, $p = 0.00$, $\eta^2p = 0.12$	$F(1, 90) = 12.52$, $p = 0.00$, $\eta^2p = 0.12$	$F(2, 90) = 1.19$, $p = 0.00$, $\eta^2p = 0.11$	
Filled pause	$F(2, 90) = 1.04$, $p = 0.36$	$F(1, 90) = 1.16$, $p = 0.29$	$F(2, 90) = 0.39$, $p = 0.68$	

Significant effects are highlighted.

5.2 RQ2: Differences in CAF between YLLs and L2 adult learners

The analyses based on CAF measures among the advanced adults (i.e., TESOL students) were added and the results were compared with those among YLLs. Figures 1–4 visually present the results (also see Table 3). In addition, Table 5 shows the results of a series of one-way ANOVAs examining the differences in the mean scores by grade level. Note that the PPVT scores were missing among the TESOL students.

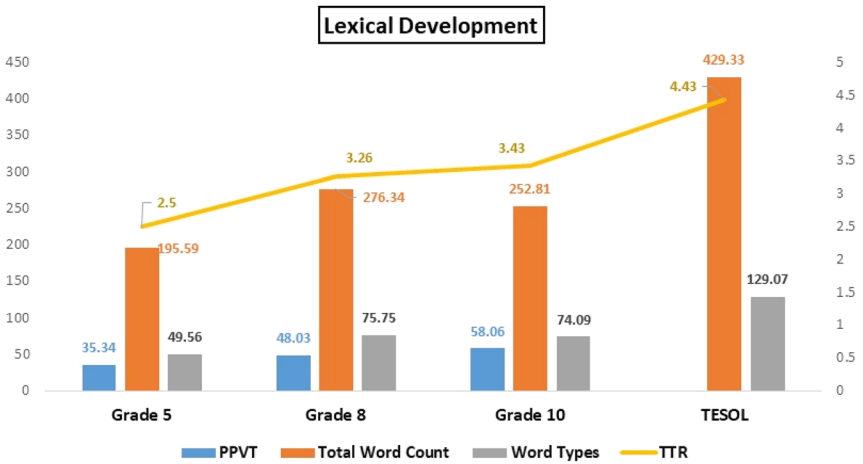


Figure 1: Vocabulary.

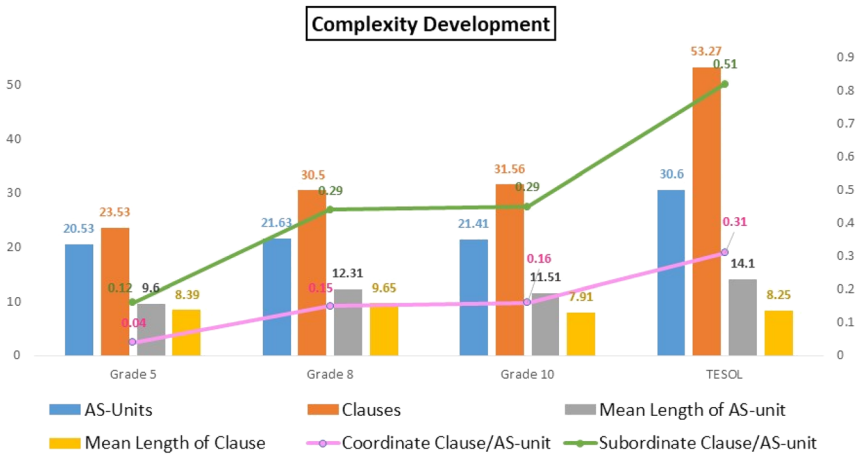


Figure 2: Complexity.

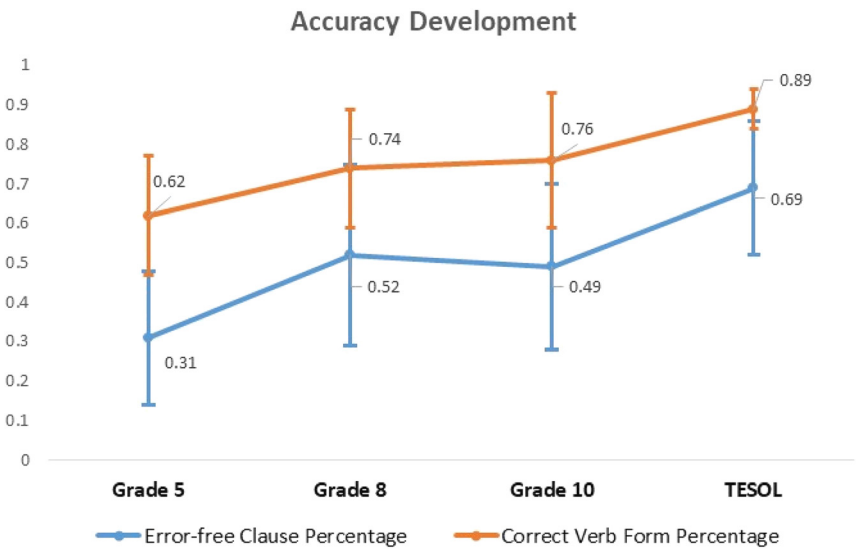


Figure 3: Accuracy.

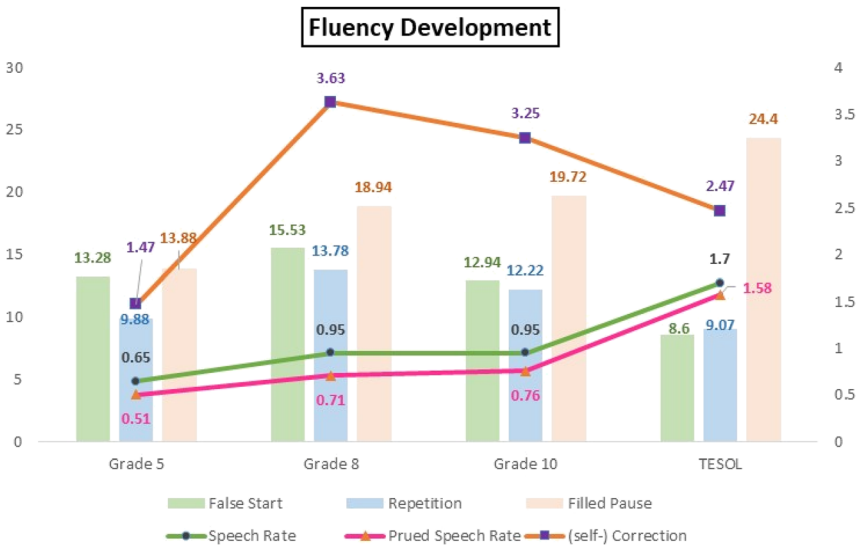


Figure 4: Fluency.

With respect to vocabulary, all the vocabulary-related measures generally showed relatively steady improvements across grade levels. TESOL students produced significantly more words and more variety of words, as expected.

Table 5: Results of One-way ANOVAs (including TESOL students).

	CAF measures	*F*/Welch's	Eta/Epsilon[a] squared	Post Hoc (Tukey/Games Howell)					
				5–8	5–10	5–T.	8–10	8–T.	10–T.
Vocabulary	Total word count	7.97**	0.18[a]			√			√
	Word type	22.84**	0.37[a]	√	√	√		√	√
	TTR	46.02**	0.56	√	√	√		√	√
Complexity	AS-units	3.75*	0.1			√		√	√
	Clauses	7.02**	0.2[a]			√		√	
	Mean length of AS-unit	16.52	0.32	√	√	√		√	√
	Mean length of clause	3.07*	0.08				√		
	Coordinate clauses/ AS-unit	20.84**	0.3[a]	√	√	√		√	
	Subordinate clause/ AS-unit	24.20**	0.36[a]	√	√	√		√	√
Accuracy	Error-free clause percentage	13.34**	0.27	√	√	√		√	√
	Correct verb form percentage	34.13**	0.25[a]	√	√	√		√	√
Fluency	Speech rate	43.93**	0.56	√	√	√		√	√
	Pruned speech rate	64.59**	0.64	√	√	√		√	√
	False start	1.55	–						
	Repetition	1.09	–						
	Self-correction	4.78**	0.06[a]	√	√				
	Filled-pause	1.31	–						

*$p < 0.05$, **$p < 0.01$. [a]equal variance is violated; thus Welch's test, epsilon squared, and Games-Howell were used with the alpha level of 0.05.

Regarding complexity measures, the total numbers of AS-units and clauses showed significant differences between YLLs and TESOL students, reflecting the larger amount of texts produced by TESOL students. The rest of the measures, except the mean lengths of clauses, appeared to distinguish students' performance by grade level, including both within YLLs and between YLLs and TESOL students.

When it comes to accuracy, both accuracy measures (the error-free clause percentage and the correct verb form percentage) showed gradual improvement across grade levels.

Finally, concerning fluency measures, both the speech rate and pruned speech rate indicated relatively steady improvements by grade levels including TESOL students. Disfluency (also referred to as breakdown fluency) measures, namely false starts, repetitions, and filled pauses, failed to show any significant differences across

grade levels, including TESOL students. The self-correction displayed a non-linear pattern; the frequencies were the highest in Grade 8 and Grade 10 but went down among the TESOL students. This result of self-correction may reflect the students' metacognitive development as well as accuracy development (see Section 6).

5.3 RQ3: Communicative adequacy among YLLs and advanced adult L2 learners

Next, the story grammar performance was analyzed. The descriptive statistics are summarized in Table 6. Among TESOL students, all of their stories contained reactive sequences and a full ending. Among YLLs, generally speaking, the element of sequence appeared relatively early, followed by goals and ending.

The means scores of story grammar performance were examined by a two-way ANOVA with repeated measures (Grade x Elements). Since the sphericity assumption was violated, the degrees of freedom were adjusted using the Greenhouse-Geisser Correction. Other assumptions including the independence of observations, normality, and the equality of covariance matrices were met. The results indicated that there was a main effect on grades ($F(3, 107) = 7.00$, $p < 0.01$, Partial $\eta^2 = 0.16$), a main effect on elements ($F(2.68, 286.47) = 47.78$, $p > 0.01$, Partial $\eta^2 = 0.31$). There was an interaction effect as well ($F(8.03, 286.47) = 2.96$, $p < 0.01$, Partial $\eta^2 = 0.08$). Since the developmental level specification made by Stein was somewhat controversial, as mentioned in the literature review section, we used the combined scores of the four elements and used them as story grammar scores for the rest of the analyses.

Table 6: Story grammar performance (means and SD).

	Grade 5	Grade 8	Grade 10	TESOL
Sequence	1.56 (0.50)	1.88 (0.34)	1.91 (0.39)	2.00 (0.00)
Goal	1.31 (0.59)	1.72 (0.46)	1.52 (0.62)	1.80 (0.41)
Obstacle	1.09 (0.69)	0.97 (0.70)	0.78 (0.79)	1.47 (0.52)
Ending	1.34 (0.55)	1.72 (0.52)	1.62 (0.49)	2.00 (0.00)

The maximum score that one could get in each element was 2.

5.4 RQ4: Relationships among the features of CAF, vocabulary and communicative adequacy among YLLs

To answer this question, first, a correlational analysis was performed among CAF and vocabulary measures and the story grammar scores for each grade level. The results for each grade level are shown in Tables 7–9. Note that only significant

Table 7: Correlations of variables among Grade 5.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1. PPVT	_																		
2. Word type	0.359*	_																	
3. Total word count		0.757**	_																
4. TTR		0.533**		_															
5. AS-unit		0.787**	0.865**		_														
6. Clause		0.846**	0.849**		0.950**	_													
7. Mean length. AS-unit			0.378*				_												
8. Mean length. clause			0.398*				0.822**	_											
9. Coordinate clause/AS-unit				0.421*					_										
10. Subordinate clauses/AS-unit				0.384*				0.448*		_									
11.Err-free clause %	0.410*										_								
12. Correct verb form %				0.460**				0.438*				_							
13. Speech rate	0.555**		0.420*		0.438*	0.418*							_						
14. Pruned speech rate	0.555**			0.371*	0.388*	0.361*							0.438*	_					
15. False start			0.669**		0.369*	0.389*	0.575**	0.552**							_				
16. Repetition			0.591**		0.353*	0.438*	0.496**	0.354*							0.640**	_			
17. Self-correction			0.363*		0.396*	0.377*											_		
18. Filled pause			0.379*				0.474**	0.486**							0.490**		0.444*	_	
19. Story grammar	0.379*	0.560**	0.612**		0.593**	0.619**													_

**Correlation is significant at the 0.01 level (2-tailed).
*Correlation is significant at the 0.05 level (2-tailed).

Table 8: Correlations of variables among Grade 8.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1. PPVT	_																		
2. Word type	0.686**	_																	
3. Total word count	0.634**	0.918**	_																
4. TTR	0.503**	0.780**	0.483**	_															
5. AS-unit	0.615**	0.889**	0.957**	0.460**	_														
6. Clause	0.641**	0.918**	0.945**	0.543**	0.929**	_													
7. Mean length. AS-unit	0.551**	0.663**	0.665**	0.448*	0.453**	0.615**	_												
8. Mean length. clause								_											
9. Coordinate clause/AS-unit	0.393*	0.553**		0.731**		0.377*	0.652**		_										
10. Subordinate clauses/AS-unit		0.387*		0.513**		0.351*	0.622**	0.621**		_									
11.Err-free clause %	0.700**	0.683**	0.512**	0.724**	0.499**	0.588**	0.453**	0.486**	0.448*		_								
12. Correct verb form %	0.656**	0.570**	0.420*	0.605**	0.364*	0.432*	0.535**	0.455**		0.730**		_							
13. Speech rate	0.400*	0.616**	0.577**	0.438*	0.518**	0.627**	0.549**	0.356*	0.546**	0.587**			_						
14. Pruned speech rate		0.454**		0.467**		0.443*	0.392*	0.387*	0.621**	0.575**			0.924**	_					
15. False start	0.429*	0.690**	0.827**		0.764**	0.715**	0.528**								_				
16. Repetition		0.618**	0.742**		0.737**	0.679**	0.416*								0.691**	_			
17. Self-correction	0.588**	0.844**	0.855**	0.498**	0.797**	0.849**	0.671**	0.395*	0.363*	0.507**	0.442*		0.686**	0.496**	0.628**	0.649**	_		
18. Filled pause		0.444*	0.588**		0.550**	0.520**									0.634**	0.707**	0.526**	_	
19. Story grammar	0.772**	0.767**	0.699**	0.599**	0.659**	0.707**	0.681**	0.538**	0.449*	0.645**	0.602**		0.514**	0.418*	0.515**	0.404*	0.656**		_

**Correlation is significant at the 0.01 level (2-tailed).
*Correlation is significant at the 0.05 level (2-tailed).

Table 9: Correlations of variables among Grade 10.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1. PPVT	_																		
2. Word type	0.660**	_																	
3. Total word count	0.598**	0.809**	_																
4. TTR	0.783**	0.851**	0.740**	_															
5. AS-unit	0.636**	0.828**	0.960**	0.798**	_														
6. Clause	0.592**	0.793**	0.980**	0.758**	0.967**	_													
7. Mean length. AS-unit			0.489**			0.401*	_												
8. Mean length. clause			0.365*				0.701**	_											
9. Coordinate clause/ AS-unit							0.616**		_										
10. Subordinate clauses/AS-unit			0.460**			0.493**	0.681**			_									
11.Err-free clause %	0.526**			0.511**	0.394*						_								
12. Correct verb form %	0.571**	0.358*		0.508**	0.366*						0.865**	_							
13. Speech rate	0.532**	0.383*	0.511**	0.503**	0.473**	0.548**	0.352*			0.447*	0.385*	0.361*	_						
14. Pruned speech rate	0.485**			0.472**	0.356*	0.398*					0.416*	0.396*	0.941**	_					
15. False start		0.523**	0.707**		0.599**	0.666**	0.518**	0.387*		0.421*					_				
16. Repetition	0.366*	0.567**	0.820**	0.434*	0.742**	0.774**	0.477**	0.397*							0.820**	_			
17. Self-correction		0.465**	0.670**	0.423*	0.619**	0.714**	0.395*			0.476**					0.666**	0.532**	_		
18. Filled pause			0.412*		0.372*	0.420*									0.796**	0.618**	0.511**	_	
19. Story grammar	0.559**	0.685**	0.694**	0.683**	0.705**	0.715**					0.435*	0.399*	0.704**	0.637**	0.446*	0.455**	0.530**		_

**Correlation is significant at the 0.01 level (2-tailed).
*Correlation is significant at the 0.05 level (2-tailed).

correlation coefficients are indicated in those tables; it would be easier to see general tendencies that way. Comparing these tables, one can make the following general observations. First, among Grade 5 students, namely, young learners with only a couple of years of English learning as their foreign language, there were relatively few significant relationships. In contrast, in Grade 8 and Grade 10, one can see many more significant correlations than in Grade 5 both within and across CAF dimensions. The same tendency was observed among correlations between CAF measures (including vocabulary) and story grammar scores. More significant correlations were obtained in Grade 8 and Grade 10, compared with Grade 5. At Grade 5, essentially only the measures related to text amounts and vocabulary showed significant correlations with the story grammar score.

Interestingly, significant correlations were fewer in number again among TESOL students, compared to Grade 8 and Grade 10 students (see Table 10). Most notably, one can see some negative correlations in fluency categories. Within the fluency dimension, such negative correlations were mostly found in disfluency measures such as false starts, repetition, and filled-pauses; negative correlations should be expected theoretically. Concerning the relationships between CAF measures and the story grammar, significant correlations were found only with vocabulary measures and two complexity measures (the numbers of AS-units and clauses) which also reflected the overall amount of production as well as complexity. It should be noted, however, that the variance in story grammar scores among the TESOL students was small; thus, correlational analyses involving these scores among the TESOL students should be interpreted with caution (see Section 6).

In general, interrelations among CAF measures as well as the relationship between CAF measure and communicative adequacy (story grammar) seem to show a non-linear pattern across grade levels.

Next, in order to examine the relative contributions of the CAF measures to the story grammar scores among YLLs, following the procedure taken by Ogawa (2022), a hierarchical regression analysis was performed among YLLs (N = 96). To avoid multicollinearity, some of the CAF measures which were theoretically and empirically highly correlated variables were checked and removed if necessary. The variables entered into the model are shown in Table 11. Based on Ogawa, predictive variables were entered into the model in the following order: fluency, vocabulary, complexity, and accuracy. All the Variance Inflation Factor (VIF) values of the variables were less than 10. As indicated in Table 11, fluency accounted for the strongest predictor of the story grammar (R^2 change = 0.37), followed by vocabulary (R^2 change = 0.12). While accuracy also showed a significant but minor contribution to the story grammar, complexity was not significant.

Table 10: Correlations of variables among TESOL students.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1. Word type	—																	
2. Total word count	0.985**	—																
3. TTR	0.917**	0.838**	—															
4. AS-unit	0.950**	0.964**	0.807**	—														
5. Clause	0.924**	0.889**	0.862**	0.895**	—													
6. Mean length. AS-unit						—												
7. Mean length. clause							—											
8. Coordinate clause/AS-unit								—										
9. Subordinate clauses/AS-unit	0.537*		0.658**						—									
10.Err-free clause %										—								
11. Correct verb form %										0.705**	—							
12. Speech rate		−0.517*		−0.613*				0.595*				—						
13. Pruned speech rate	−0.528*	−0.543*		−0.610*				0.657**				0.973**	—					
14. False start	0.710**	0.697**	0.629*	0.679**	0.765**								−0.526*	—				
15. Repetition	0.766**	0.796**	0.633*	0.803**	0.688**							−0.746**	−0.822**	0.811**	—			
16. Self-correction							0.640*									—		

Table 10: (continued)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
17. Filled pause	0.834**	0.810**	0.789**	0.826**	0.789**							−0.576*	−0.567*	0.616*	0.712**		—	
18. Story grammar	0.780**	0.823**	0.605*	0.746**	0.663**										0.619*			—

**Correlation is significant at the 0.01 level (2-tailed). *Correlation is significant at the 0.05 level (2-tailed).

Table 11: Multiple regression analysis results using analytical CAF as predictors of communicative adequacy.

Predicated variable	Predictor dimension	Predictor variables entered in the model	R^2	R^2 change	F	p
Story grammar	Fluency	False start, pruned speech rate, filled pause, self-correction	0.398	0.371	11.291	$p < 0.000$
	Vocabulary	PPVT, corrected type-token ratio, total word count	0.515	0.116	11.536	$p < 0.000$
	Complexity	Mean length of clause, coordinate clause/AS-unit, subordinate clause/AS-units, AS-unit, mean length of AS-unit	0.547	0.032	8.347	n.s.
	Accuracy	Correct verb %, error free clause %	0.582	0.035	8.048	$p < 0.05$

6 Discussions

Based on a cross-sectional dataset, the present study aimed to understand how YLLs' storytelling performance differed across grade levels and by students' SES in terms of CAF and vocabulary measures. The study also examined such performance differences between YLLs and very advanced adult L2 learners who shared the same L1 with the YLLs, instead of comparing YLLs with 'native speakers' of English. The study further investigated the relationships between communicative adequacy, operationalized as story grammar, and CAF/vocabulary measures, first by examining correlations among the variables and second by performing a hierarchical regression analysis.

In many CAF and vocabulary measures, significant differences were found between Grade 5 and Grade 8, but not between Grade 8 and Grade 10. This lack of difference between Grade 8 and Grade 10 might be attributed to greater variability of performance within the same age group at older grade levels. Indeed, the effects of SES began to manifest from Grade 8, and appeared even widened at Grade 10; however, a longitudinal study is necessary to confirm this observation. Some exceptional variables that did not follow the general tendency above (i.e., variables that did not show significant differences by grade) include the number of AS-unit, false start, repetition, and filled-pause. One can argue that the number of AS-units is somewhat related to the amount of oral text production. Even though the analysis failed to find the main effect on Grade in AS-units, given that it had a significant interaction effect, this variable also reflects a similar tendency with other complexity variables; namely, there was a substantial and widened disparity between high and

low SES groups. The remaining three measures that did not show differences by grade – false start, repetition, and filled-pause – are all disfluency (breakdown fluency) measures. In the present study, these disfluency measures did not show linear patterns; various other factors might have easily influenced them. L1 studies found that children's disfluencies were influenced by tasks but not age. Since narrative tasks tend to induce more disfluency than conversation tasks among school-age children who stutter, narrative tasks such as the one implemented in this study are often used to elicit disfluencies for diagnostic purposes (Byrd et al. 2012). Considering Byrd al.'s study, it is not surprising that the present study failed to find any differences in the disfluency measures by grade. Individual differences might have obscured any set patterns in the present study as well.

As expected, adult TESOL students' performance was significantly higher than YLLs' in most CAF and vocabulary measures. In general, vocabulary and accuracy tended to improve as the grade level increased; however again, a longitudinal investigation is necessary to confirm this observation. Complexity measures also showed significant differences between TESOL students and YLLs, except for the mean length of clauses. This result of the mean length of clauses might have been related to the fact that this measure is "radically different from the other length-based measures" because, it captures a specific type of complexity at the phrase level, not at the clause or sentence level (Norris and Ortega 2009, p. 561). For the particular task implemented in this study, the mean length of clauses might not have been a useful one to distinguish students with different grade levels. Fluency is considered a multidimensional construct and can be categorized into three sub-types: speed fluency, breakdown fluency, and repair fluency (Tavakoli and Skehan 2005). TESOL students' narratives were significantly higher in speed fluency (speech rate and pruned speech rate) than YLLs', but failed to show significant differences in breakdown fluency, as seen already. Repair fluency (i.e., self-correction) indicated a non-linear pattern. It was low in frequency in Grade 5, the highest in Grades 8 and 10, and lower again in TESOL students. Considering that self-correction requires metacognitive abilities (Postma 2000), this variable may reflect the students' metacognitive development. It might also be influenced by accuracy among TESOL students. If students did not make errors, they did not need to self-correct in the first place.

In this study, the significant correlations among CAF/vocabulary measures in Grade 5 were much fewer than those among Grade 8 and Grade 10, both within and across dimensions (complexity, accuracy, fluency, and vocabulary). Different dimensions were not yet mutually connected with each other in Grade 5. Among these young beginners of English learning, essentially, only vocabulary measures and the number of AS-unit and clauses (somewhat related to the amount of oral text production) turned out to show some correlations with story grammar. Among

Grade 8 and Grade 10, CAF/vocabulary measures were much more interrelated significantly. Many more CAF/vocabulary variables were also significantly correlated with story grammar in these secondary school students. Interestingly, among TESOL students, a fewer number of significant correlations were found again. The lack of significant correlations of complexity – both among themselves and with other measures in general – was particularly notable. This result confirmed Pallotti's (2009) statement that greater complexity does not necessarily imply greater advancement. Similar to Grade 5 students, only vocabulary measures and the number of AS-units and clauses were significantly correlated with story grammar in the TESOL group. As mentioned, this lack of significant correlations among many variables in TESOL students is most likely be attributable to their generally high performance with less variability; they may have a ceiling effect.

As far as the school curriculum is concerned, one can assume that the Grade 8 and Grade 10 students in this study were, by and large, at the intermediate level. If this assumption is correct, one can argue that CAF measures were most useful among intermediate proficiency students because they were mostly associated with communicative adequacy. In contrast, for beginners and advanced learners, given the lack of associations of many CAF measures with communicative adequacy at these levels, vocabulary measures, rather than CAF measures, appeared to be more reliable measures. This finding has useful implications for material and assessment development. For example, teachers should pay sufficient attention to their YLLs' proficiency levels and ages when determining which assessment measures to use.

When focusing on YLLs, a hierarchical regression analysis indicated that fluency contributed most to story grammar, followed by vocabulary. Accuracy made a significant but minor contribution to story grammar scores. Complexity was not significant. The largest contribution of fluency to communicative adequacy in oral tasks is consistent with previous studies, both concerning adults (Koizumi and In'nami 2024; Ogawa 2022; Révész et al. 2016) and children (Hsieh and Wang 2019; Mihaljević Djigunović 2016). This consistent results with previous studies are interesting in that, while previous studies use general oral proficiency judged by raters for communicative adequacy, our study used story grammar, which is presumably a multidimensional construct, not only relies on linguistic abilities but also memory, attention and other cognitive and socio-cognitive abilities (Duinmeijer et al. 2012). Since fluency itself is a multidimensional construct (Tavakoli and Skehan 2005), more work is necessary to unpack the complex mechanisms between fluency and communicative adequacy. A significant role of vocabulary in communicative tasks was also consistent with Hsieh and Wang (2019) conducted among young learners. It is worth noting that our vocabulary measures consisted of lexical diversity (in addition to general receptive vocabulary knowledge, PPVT). Hsieh and Wang (2019)

speculated about the possibility of having "a qualitative shift" (p. 43) in vocabulary use as YLLs develop their oral proficiency. While examining the depth of vocabulary knowledge across different age and proficiency levels is challenging methodologically, this topic is important to better understand YLLs' vocabulary development and communicative adequacy.

The present study was not free from limitations. First is its research design: a cross-sectional design. While cross-sectional design is not uncommon in developmental studies, it is limited in that it cannot establish clear temporal and causal relationships among variables. A few observations made by this study need to be confirmed by longitudinal investigations. Second, related to the first limitation, the present study primarily concerned the group mean performance. Considering potentially large individual differences in language development among YLLs, even those learning English under the 'same' curriculum, investigations focusing on individual differences would be important for pedagogy. Third is the validity of story grammar as a measure of communicative adequacy. As discussed in the literature review section, it is not entirely clear if story grammar was an adequate measure for this study. Story grammar may not be appropriate for older learners and learners who did not have English as their L1 because the original framework was developed among monolingual English-speaking children. Finally, this study was conducted among Chinese-speaking learners of English. Studies with different L1 backgrounds and/or those who learn English in various environments would be necessary if one wants to examine the generalizability of the findings in the current study.

7 Conclusions

The present study examined how YLLs' discourse features were developed by focusing on complexity, accuracy, fluency (CAF), and vocabulary – variables that had been used widely in L2 development studies that primarily concerned adult learners. The study examined the development of these variables among YLLs in relation to advanced adult L2 learners. The study also uniquely examined the effect of socioeconomic status (SES) as well as grade levels. Furthermore, the study analyzed the relations between communicative adequacy (measured by story grammar) and CAF/vocabulary measures.

The results indicated that while many CAF measures showed significant differences across grades, significant effects of SES were also found and widened in secondary school. This finding of the SES effect highlights the important role that learning/teaching contexts play when using CAF measures for understanding students' language development. In addition, since all dimensions (i.e., complexity,

accuracy, fluency and vocabulary) are multifaceted, not all the variables within the given dimensions showed similar increasing patterns across grade levels. For example, the mean length of clauses was found to be different from other complexity measures among YLLs, consistent with previous studies (e.g., Hsieh and Wang 2019). Speed fluency, breakdown fluency, and repair fluency all showed different patterns. Speed fluency showed a relatively steady increase as the grade level increased (even including adult TESOL students), while breakdown fluency failed to find significant differences by grade. Repair fluency showed a non-linear pattern; it showed the highest frequency in Grade 8 and Grade 10, but lower in Grade 5 and TESOL students. Some measures, such as repair fluency, can be assumed to be influenced by meta-cognitive development, which correlates with age. Choosing different measures appropriate for age and proficiency levels appears to be important when these measures are used for assessment.

In this study, in Grade 5, CAF measures were not yet interrelated, whereas they were more interrelated within the given dimension and across dimensions in secondary school grades. Fluency measures contributed most to story grammar, followed by vocabulary. These findings are generally consistent with previous studies conducted both among adults and children. Note, however, that previous studies used general oral proficiency judged by raters as a measure of communicative adequacy. More research is necessary to unpack the complex interplay between communicative adequacy and fluency and vocabulary measures, while implementing multiple tasks that require different levels of cognitive maturity and cognitive/socio-cognitive resources. Such efforts are valuable considering that AI-based assessments, which often rely on structural features of learners' L2 performance, are increasingly popular even among YLLs. More thorough examinations are necessary to identify which features are valid and reliable according to learners' age, proficiency level and learning backgrounds, in order to adequately design and use such AI tools for YLLs.

Acknowledgments: The authors are grateful to all the participating students and their teachers for making this study happen.
Research ethics: The local Institutional Review Board approved the study, and the authors obtained assent or consent from the participants and their parents for the underage participants.
Author contributions: The authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Conflict of interest: The authors state no conflict of interest.
Research funding: None declared.
Data availability: The raw data may be obtained on request from the corresponding author.

References

Aksu-Koç, Ayhan. 1996. Frames of mind through narrative discourse. In Dan. I. Sloban, Julie Gerhardt, Amy Kyratzis & Jiansheng Guo (eds.), *Social interaction, social context, and language: Essays in honor of Susan Ervin-Tripp*, 309–328. New York: Lawrence Erlbaum Associates.

Bamberg, Micheal. 1997. *Narrative development: Six approaches*. New York: Lawrence Erlbaum Associates.

Berman, Ruth A. & Dan I. Slobin. 1994. *Relating events in narrative: A crosslinguistic developmental study*. New York: Lawrence Erlbaum Associates.

Berman, Ruth A. 2009. Language development in narrative contexts. In Edith. L. Bavin (ed.), *The Cambridge handbook of child language*, 355–376. Cambridge: Cambridge University Press.

Block, David. 2014. *Social class in applied linguistics*. New York: Routledge.

Bui, Gavin & Peter Skehan. 2018. Complexity, accuracy, and fluency. In John I. Liontas (ed.), *The TESOL encyclopedia of English language teaching*. Hoboken, New Jersey: John Wiley & Sons.

Butler. 2025. *Children's additional language learning in instructional settings: Implications for teaching and future research*. Bristol: Multilingual Matters.

Butler, Yuko G., Peter Sayer & Becky Huang. 2018. Introduction: Social class/socioeconomic status and young learners of English as a global language. *System* 73. 1–3.

Butler, Yuko G. & Wei Zeng. 2014. Young learners' storytelling in their first and foreign languages. In Jeffrey Connor-Linton (ed.), *Measured language: Quantitative approaches to acquisition, assessment, processing, and variation*, 79–94. Georgetown: Georgetown University Press.

Byrd, Courtney T., Kenneth J. Logan & Ronald B. Gillam. 2012. Speech disfluency in School-age children's conventional and narrative discourse. *Language, Speech, and Hearing Sciences in Schools* 43. 153–163.

De Jong, Nivia H., Margarita P. Steinel, Arjen F. Florijn, Rob Schoonen & Jan H. Hulstijn. 2012. The effect of task complexity on functional adequacy, fluency and lexical diversity in speaking performances of native and non-native speakers. In Alex Housen, Folkert Kuiken & Ineke Vedder (eds.), *Dimensions of L2 performance*, 121–142. Amsterdam: Benjamins.

Duinmeijer, Iris, Jan de Jong & Annette Scheper. 2012. Narrative abilities, memory and attention in children with a specific language impairment. *International Journal of Language & Communication Disorders* 47. 542–555.

Ellis, Rod. 2009. The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Applied Linguistics* 30(4). 474–509.

Evanini, Keelan, Mmaurice Cogan Hauck & Hakuta Kenji. 2017. *Approaches to automated scoring of speaking for K-12 English language proficiency assessments*. Policy information report and ETS research report series NO. RR-17-18. Princeton, NJ: Educational Testing Service.

Foster, Pauline, Allan Tonkyn & Gillian Wigglesworth. 2000. Measuring spoken language: A unit for all reasons. *Applied Linguistics* 21(3). 354–375.

Hasnain, Shazia & Santoshi Hilder. 2024. Intricacies of the multifaceted triad-complexity, accuracy, and fluency: A review of studies on measures of oral production. *Journal of Education* 204(1). 145–158.

Hickmann, Maya. 2003. *Children's discourse: Person, space and time across languages*. Cambridge: Cambridge University Press.

Housen, Alex & Folkert Kuiken. 2009. Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics* 30(4). 461–473.

Hsieh, Ching-Ni & Yuan Wang. 2019. Speaking proficiency of young language students: A discourse-analysis study. *Language Testing* 36(1). 27–50.

Jackson, Daniel O. & Sakol Suethanapornkul. 2013. The cognition hypothesis: A synthesis and meta-analysis of research on second language task complexity. *Language Learning* 63(2). 330–367.

Koizumi, Rie & Yo In'nami. 2024. Predicted functional adequacy from complexity, accuracy, and fluency of second-language picture-prompted speaking. *System* 120. https://doi.org/10.1016/j.system.2023.103208.

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

Lee, Young-Ja, Jeehyun Lee, Myae Han & Judith A. Schickedanz. 2011. Comparison of preschoolers' narratives, the classroom book environment, and teacher attitudes towards literacy practices in Korea and the United States. *Early Education and Development* 22(2). 234–255.

Mayer, Mercer. 1969. *Frog, where are You?* New York: Dial Press.

Mayo, García, María del Pilar, Ainara Imaz Agirre & Agurtzane Azkarai. 2018. Task repetition effects on CAF in EFL child task-based interaction. In Mohammad Javad Ahmadian & María del Pilar García Mayo (eds.), *Recent perspectives on task-based Language learning and teaching*, 9–28. Berlin, Boston: Mouton.

Michel, Marije. 2017. Complexity, accuracy, and fluency in L2 production. In Shawn Loewen & Masatoshi Sato (eds.), *The Routledge handbook of instructed second language acquisition*, 50–68. New York: Routledge.

Mihaljević Djigunović, Jelena. 2016. Individual learner differences and young learners' performance on L2 speaking tests. In Marianne Nikolov (ed.), *Assessing learners of English: Global and local perspectives*, 243–261. New York: Springer.

Nicolopoulou, Ageliki. 2008. The elementary forms of narrative coherence in young children's storytelling. *Narrative Inquiry* 18(2). 299–325.

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

Ogawa, Chie. 2022. CAF indices and human ratings of oral performances in an opinion-based monologue task. *Language Testing in Asia* 12(4). https://doi.org/10.1186/s40468-022-00154-9.

Pallotti, Gabriele. 2009. CAF: Defining, refining and differentiating constructs. *Applied Linguistics* 39(4). 590–601.

Phillipson, Robert. 1992. *Linguistic imperialism*. Oxford: Oxford University Press.

Postma, Albert. 2000. Detection of errors during speech production: A review of speech monitoring models. *Cognition* 77(2). 97–132.

Révész, Andrea., Monika Ekiert & Eivind Nessa Torgersen. 2016. The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics* 37(6). 828–848.

Robinson, Peter. 2001. Task complexity, task difficulty and task production: Exploring interactions in a componential framework. *Applied Linguistics* 22(1). 27–57.

Rodriguez, Béatrice. 2005. *The chicken Thief*. Wellington, New Zealand: Gecko Press.

Sample, Evelyn & Marije Michel. 2015. An exploratory study into trade-off effects of complexity, accuracy, and fluency on young learners' oral task repetition. *TESL Canada Journal* 31. 23–46.

Skehan, Peter. & Pauline Foster. 2012. Complexity, accuracy, and fluency and lexis in task/based performance: A synthesis of the ealing research. In Alex Housen, Folkert Kuiken & Ineke Vedder (eds.), *Dimensions of L2 performance and proficiency*, 199–220. Amsterdam: John Benjamins.

Skehan, Peter. 1998. *A cognitive approach to language learning*. Oxford: Oxford University Press.

Skehan, Peter. 2009. Modeling second language performance: Integrate complexity, accuracy, fluency, and lexis. *Applied Linguistics* 39(4). 510–532.

Stadler, Marie A. & Gay Cuming Ward. 2005. Supporting the narrative development of young children. *Early Childhood Educational Journal* 33(2). 73–80.

Stein, Nancy L. & Elizabeth R. Albro. 1997. Building complexity and coherence: Children's use of goal-structured knowledge in telling stories. In Michael Bamberg (ed.), *Narrative development: Six approaches*, 5–44. Hillsdale, New Jersey: Erlbaum.

Stein, Nancy. 1988. The development of children's storytelling skills. In Mayery B. Franklin & Sybil S. Barten (eds.), *Child Language: A reader*, 282–297. Oxford: Oxford University Press.

Tavakoli, Parvaneh. & Peter Skehan. 2005. Strategic planning, task structure and performance testing. In Rod Ellis (ed.), *Planning and task performance in a second language*, 239–277. Amsterdam: John Benjamins.

Verhoeven, Ludo & Sven Strömqvist (eds.), 2001. *Narrative development in a multilingual context*. Amsterdam: John Benjamins.

Vermeer, Anne. 2000. Coming to grips with lexical richness in spontaneous speech data. *Language Testing* 17(1). 65–83.

Viberg, Åke. 2001. Age-related and L2-related features in bilingual narrative development in Sweden. In Ludo Verhoeven & Sven Strömqvist (eds.), *Narrative development in a multilingual context*, 87–128. Amsterdam: John Benjamins.

Wang, Qi. & Michelle D. Leichtman. 2000. Same beginnings, different stories: A comparison of American and Chinese children's narratives. *Child Development* 71(5). 1329–1346.

Wolf, Kim Mikyung, Alexis A. Lopez, Saerhim Oh & Fred S. Tsutagawa. 2017. Comparing the performance of young English language learners and native English speakers on speaking assessment tasks. In Mikyung Kim Wolf & Yuko Goto Butler (eds.), *English language proficiency Assessments for young learners. Innovations in language Learning and Assessment at ETS*, 171–190. New York: Routledge.

Yuan, Fangyuan & Rod Ellis. 2003. The effects of pre-task planning and online planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics* 24(1). 1–27.