Accepted for publication on May 16, 2022

Measurement Properties of Mean Length of Utterance (MLU) in School-Aged Children

Jill R. Potratz[1], Christina Gildersleeve-Neumann[2], and Melissa A. Redford[1]

[1]Department of Linguistics, University of Oregon, Eugene, Oregon

[2]Department of Speech & Hearing Sciences, Portland State University, Portland, Oregon

Author Note

Portions of these finding were presented as a poster at the 2021 Symposium on Research in Child Language Disorders, Madison, WI, United States. We have no conflicts of interest to disclose.

Correspondence concerning this article should be addressed to Jill R. Potratz, Department of Linguistics, 1290 University of Oregon, Eugene, OR 97403. 510-967-8879. Email: jillp@uoregon.edu, *ORCiD:* https://orcid.org/0000-0001-5306-9827

Conflict of Interest

There are no relevant conflicts of interest for any author.

Funding

This work was supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development (NICHD) under grant R01HD087452 (PI: Redford). The content is solely the authors' responsibility and does not necessarily reflect the views of NICHD.

Abstract

Purpose: Mean length of utterance (MLU) is one of the most widely reported measures of syntactic development in the developmental literature, but it's responsiveness in young school-aged children's language has been questioned and it has been shown to correlate with non-syntactic measures. This study tested the extent to which MLU shows measurement properties of responsiveness and construct validity when applied to language elicited from elementary school children.

Methods: Thirty-two typically developing children in two age groups (5 and 8 years old) provided four short language samples each. Language samples were elicited in a question-answer context and a narrative context. MLU was calculated with both morpheme and word counts. Other established measures of syntactic complexity (clausal density, CD; Developmental Level, D-Level; mean length clause, MLC) and lexical diversity (lexical density, LDensity; moving-average type token ratio, MATTR; number of different words, NDW) were also calculated.

Results: Linear mixed-effects analyses revealed that MLU varied systematically with discourse context and children's age group. The syntactic measures, CD and MLC, were found to vary systematically with MLU. None of the lexical diversity measures varied systematically with MLU.

Conclusions: Results suggest that MLU is a responsive and valid measure of children's syntactic development across age and discourse context during the early school-age years.

Introduction

The ability to easily quantify language complexity and its development is vitally important for speech-language pathologists (SLPs) since the evaluation of language skills at a particular developmental stage may lead to a child's initial enrollment in, or continuing eligibility for, special education services (ASHA, 2016). Language sample analysis (LSA) is often used in assessment either as an alternate or supplement to norm-referenced testing (Castilla-Earls et al., 2020). One of the most widely used measures of syntactic complexity derived from LSA is mean length of utterance (MLU), which is seen by clinicians as a robust measure of children's syntactic development (e.g., Bernstein & Tiegerman-Farber, 1997; Miller & Chapman, 1981). MLU has featured prominently in studies of monolingual English language development during the early school-aged years, especially in studies that seek to distinguish children with typical language from those with communication disorders (see e.g., Charest et al., 2020; Fey et al., 2004; Moyle et al., 2011; Rice et al., 2010). But its relevance for capturing syntactic complexity in older, elementary school-aged children's language has been questioned (Bernstein & Tiegerman-Farber, 1997; Blake et al., 1993; Klee & Fitzgerald, 1985; Rondal et al., 1987). The current study therefore asks: Does MLU provide responsive and valid measures of syntactic complexity in young school-aged children's language?

MLU was originally proposed by Brown (1973) to capture the early language development in his longitudinal study of three children. He counted the number of morphemes in each utterance that the children produced and then calculated the mean number of morphemes produced per utterance. Many studies on child language development have followed suit (Heilman et al., 2010; Hewitt et al., 2005; Moyle et al., 2011), but others have used the number of words in an utterance as the unit of count (Charest et al., 2020; Fey et al., 2004; Nippold,

2009). While we might expect that calculating MLU using morphemes rather than words would result in outcomes that provide more developmental information, Parker and Brorson (2005) argue that the unit of count is inconsequential. They found that MLUm (by morphemes) and MLUw (by words) are correlated in 3-year-old English-speaking children's language. Additionally, Rice and colleagues (2010) indicate that MLU in either words or morphemes yields reliable and valid estimates of children's language growth.

Measurement Responsiveness and Validity

Brown (1973) argued that MLU indexed the constructional complexity of children's developing language. Yet, there is disagreement regarding the sufficiency of MLU as a measure of syntactic complexity beyond a certain level of development (Bernstein & Tiegerman-Farber, 1997). For example, the last stage of development that Brown described, Stage V, is characterized by an MLU of between 3.75-4.5 morphemes, which is correlated with use of advanced forms such as third person singular, and contractable auxiliaries and copulas. According to Brown, this stage is typically reached by 41-46 months. After Stage V, Brown found that children produce relative clauses attached to the subject (e.g., *The dog* ***that jumps over the pig*** *bumps into the lion*; Diessel & Tomasello, 2015), as well as embedding and conjoining within the same sentence. Since this type of complexity may not be reflected in an increase in utterance length, Brown suggested using MLU as a measure of syntactic development only up until Stage V.

Subsequent studies have also questioned whether MLU is in fact a valid measure of syntax in older, school-age children. For example, Frizelle, Thompson, McDonald, and Bishop (2018) analyzed narratives obtained from a large cross-sectional sample of speakers ($N = 354$), aged 4 years to adult. They measured syntactic complexity in the samples using both MLU (with

words as the count unit) and clausal density (CD), which was defined by Scott (1988) as the mean number of clauses per utterance. Frizelle and colleagues argued that while both MLU and CD increased with age and were highly inter-correlated, CD provided better evidence of developmentally related change in complexity. They argued that CD provided information about the degree of subordination; utterance length indicates nothing in particular about the syntactic structures in use. Moreover, like Brown (1973), Frizelle and colleagues argued that, at a certain point in development, MLU is likely to be inversely correlated with syntactic ability. Specifically, increases in phrasal complexity entail the use of subordinate clauses, which packs more information into a single sentence using the same number or fewer words.

In studies that focus on identifying language impairment, MLU is paired with lexical measures to provide a more complete description of atypical language (Fey et al., 2004; Miller et al., 1992; Paul et al., 1996). Lexical diversity measures are assumed to provide information about language ability not captured by MLU under the assumption that MLU measures syntax. For example, the number of different words (NDW) in a sample and moving-average type-token ratio (MATTR) are used to capture age- and impairment-related differences in expressive vocabulary (Charest et al., 2020; Dethorne et al., 2005; Fey et al., 2004; Watkins et al., 1995). However, some of these studies also report a correlation between measures of lexical diversity and MLU (Dethorne et al., 2005; Ukrainetz & Blomquist, 2002), which again begs the question of what exactly MLU measures. In the present study, we address this question in younger and older school-aged children by eliciting narrative samples across two discourse contexts.

Current Study

Disagreements exist regarding the upper age at which MLU is useful, and so whether MLU is a responsive measure of language development in young school-age children. Does the

measure capture language differences during these years (i.e., responsiveness) or is it too coarse-grained of a measure to do so, as Frizelle and colleagues (2018) might argue? Additionally, correlations between MLU and measures of lexical diversity, such as NDW, suggest that data generated by MLU, while designed to assess syntax, may instead measure language development more broadly, which raises the question of construct validity. The current study therefore asks two research questions.

Research Question 1 (RQ1): Is MLU a responsive measure of cross-sectional age-related and context-related differences in young school-aged children's language? Responsiveness is typically defined as the ability to detect change over time in the construct being measured (Mokkink et al., 2010). We operationalized measurement responsiveness in the current study as the detection of difference in language between two age groups (i.e., 5- and 8-year-olds) and two discourse contexts: narrative and question-answer contexts.

The two ages of comparison were selected intentionally. Five is the age when children enter kindergarten and may be referred by their teachers for a speech-language evaluation. A comparison group of 8-year-olds was chosen since our targeted expressive language skills will have developed beyond what is seen in 5-year-olds to the extent that group differences are expected. Younger and older school-aged children were expected to differ in their language ability based on typical developmental norms (e.g., Hoff, 2014). This expectation was confirmed in the present study using a norm-referenced language assessment to evaluate language ability in the 5- and 8-year-old child participants.

Two discourse contexts were chosen for the present study because changes in context are known to produce different syntactic outcomes (e.g., Nippold, 2009; Nippold et al., 2015). The different contexts have different task requirements that drive language behavior in certain ways;

in particular, more complex task requirements drive more complex language. For example, Nippold (2009) showed that expository language, elicited by asking school-age children (ages 7;3-15;4) questions about how to play their favorite game, is associated with a greater variety of subordinate clause types compared to conversational samples. MLU has also been reported to be higher for narrative-retelling tasks than for critical-thinking tasks (Nippold et al., 2015), albeit with an older group of children, adolescents. The question-answer context in the current study was expected to elicit greater complexity in language relative to the narrative context. The materials used for language sample elicitation require the children to answer "why" and "how" questions. These types of questions prompt higher-level thinking and reasoning (e.g., making inferences, problem solving, persuasive argument) and word choices that force subordination.

Research Question 2 (RQ2): Is MLU a valid measure of syntactic complexity in young, school-age children, or does it capture other aspects of language complexity? Construct validity is the degree to which a measurement captures the construct in question (Mokkink et al., 2010). Although MLU is assumed to be a measure of syntactic development, it has been correlated with measures of lexical diversity in studies that use both (e.g., Dethorne et al., 2005). If MLU is a valid measure of syntax, then this correlation is likely due to underlying language ability: The child who produces language that is syntactically complex also has a larger expressive vocabulary than the child who produces language that is simple. If this hypothesis is correct, then MLU should correlate more strongly with other known measures of syntax, and less with measures of lexical diversity once the shared variance in language ability is accounted for. To test this prediction, three established measures of syntactic complexity and three established measures of lexical diversity were calculated for each sample (see Method for specific measures

and procedures). Multiple measures were used to robustly characterize "lexical diversity" and "syntax," since we know of no gold standard measures for these constructs.

Methods

Participants

Thirty-two children participated in the study. Participants were sixteen 5-year-old (8 female, 8 male) and sixteen 8-year-old (8 female, 8 male) children with typically developing speech, language, and hearing. The average age in the 5-year-old group was 5;5 years (range = 60 to 70 months); it was 8;5 years in the 8-year-old group (range = 96 to 107 months). Parents reported English as the children's first/native language. The English dialect was Standard American inflected by the back-vowel fronting typical of the West Coast. Five children had one caregiver whose first language was either German (1 in the 5-year-old group and 1 in the 8-year-old group) or Spanish (2 in the 5-year-old group and 1 in the 8-year-old group). As per parent report, the racial and ethnic distribution in the 5-year-old group was 63% White, 13% biracial (Asian American and White), 13% Hispanic, 1% Black, and 1% Native American and in the 8-year-old group the distribution was 88% White and 12% biracial (Asian American and White). Since socioeconomic status is positively correlated with caregiver education (Davis-Kean, 2005), caregiver education information was collected. Children in the current sample had caregivers with mostly high educational attainment: in the 5-year-old group 47% of caregivers had advanced degrees (i.e., masters, PhD, or MD), 28% had a college degree, 13% had some college, 3% had a high-school degree, and 9% had not finished high school and in the in the 8-year-old group 56% of caregivers had advanced degrees (i.e., masters, PhD, or MD), 31% had a college degree, 3% had some college, 3% had a high-school degree, and 6% had not finished high school. Participants were recruited from the Portland, Oregon area using fliers distributed

throughout the community (e.g., libraries, schools) and though word of mouth. IRB approval was granted by the University of Oregon and extended to data collection at Portland State University, which followed all protocol guidelines and used approved consent/assent materials.

Hearing was screened at 1000 Hz, 2000 Hz and 4000 Hz at a threshold of 25 dB SPL. Typical speech and language development was determined using the articulation subtest of the *Diagnostic Evaluation of Articulation and Phonology* (DEAP; Dodd et al., 2002) and the Core Language Score on the *Clinical Evaluation of Language Fundamentals* (CELF-5; Wiig et al., 2013). Children's scaled scores on the DEAP ranged from 7 to 12 (M = 10.53, SD = 1.9). Their Core Language Scores from the CELF-5 ranged from 86 to 136 (M = 116.1, SD = 10.12) without a significant difference between the two age groups (t = 0.019, p = .493, Cohen's d = .007).

Procedure

Each of the 32 participants provided four short language samples for analysis. These were elicited in two contexts: (1) a question-answer context and (2) a storytelling context. The same materials were used in both contexts. These materials included six picture prompts for each of two stories, *Dog Comes Home* and *Bunny Goes to School,* taken from the School-Age Language Assessment Measures (SLAM, Crowley & Baigorri, 2015). For each story, children inspected the picture cards placed before them while answering standard evaluation questions (see Appendix A). The questions required the participant to deduce (e.g., "Why is she in the bathtub with a white dog now?"), infer (e.g., "Why did she come to school?"), problem solve (e.g., "What would you do if a bunny started hopping around your school?"), predict (e.g., "What do you think will happen when the boy goes home?"), and use theory of mind (e.g., "What is the girl thinking here?"). Then, with the pictures still in front of them, the children were asked to tell the story shown in the pictures (see Appendix B for sample narratives). During language sample

elicitation, experimenters provided natural prompts, such as question repetition, encouragement to answer the question (e.g., Experimenter: "What do you think the mother's going to do now?" Child: "I don't know." Experimenter: "Do you have any thoughts? What do you think?"), and encouragement to begin or continue the story narrative (e.g., Experimenter: "Now you tell me the whole story." Child: "I don't want to." Experimenter: "But you already told me so much of it. What's happening here?"). The language samples were audio-visually recorded in a quiet experimental room at Portland State University.

Language Sample Transcription and Coding

Audio files of the language samples were transcribed using Systematic Analysis of Language Transcripts (SALT; Miller & Iglesias, 2019) conventions to identify utterances, words, morphemes, unintelligible segments, and mazes (Miller et al., 2019). Specifically, the samples were first segmented into C-units (i.e., one main clause and any modifiers or subordinate clauses; Loban, 1976) using SALT's C-unit segmentation rules (Miller et al., 2019). Utterances that were less than a C-unit (e.g., "sorry", "bye") were included in the analyses as long as they were not maze behavior (i.e., false starts, repetitions, reformulations, and filled pauses, e.g., um; uh; well). Stereotypic closing (e.g., "the end") and side comments were placed on special lines and were not included as part of the transcription. Lexical verbs and copulas were coded with "[v]," so that the number of clauses (i.e., main and subordinate clauses defined as a statement containing both a subject and predicate) could be tallied automatically by SALT, as per Fey et al. (2004). Infinitives (e.g., "because she want/3s[v] to hide[v] it from the mom") and bare infinitives (e.g., "and she think/3s[v] her mom won't let[v] her keep[v] the dog) were coded as clauses.

Measures

Table 1 summarizes the various measures taken in the present study and how they were

calculated based on the segmentation and transcriptions of the language samples. Several of the measures were automatically calculated using SALT.

Table 1

Measure Calculations

Measure	Calculation
Mean Length of Utterance (MLUm)	Dividing the total number of morphemes by the total number of C-units (SALT-generated)
Mean Length of Utterance (MLUw)	Dividing the total number of words by the total number of C-units (SALT-generated)
Syntactic Complexity	
Clausal Density (CD)	Dividing the total number of clauses in a sample by the total number of C-units
Developmental Level (D-Level)	Manually scoring each utterance on an 8-point complexity scale and calculating the average for each sample
Mean Length of Clause (MLC)	Dividing the number of total words (NTW SALT-generated) in sample by the number of clauses in that sample
Lexical Diversity	
Lexical Density (LDensity)	Dividing the number of lexical items (nouns, verbs, adjectives and adverbs) by the total number of words in each sample*100
Moving-average type-token ratio (MATTR)	Dividing moving-average NDW by moving - average number of total words (SALT-generated)
Number of Different Words (NDW)	Tabulation of the total number of different words in a sample (SALT-generated)

Note. NTW = number of total words; SALT = Systematic Analysis of Language Transcripts

The three syntactic measures were the aforementioned CD measure, a measure known as "Developmental Level" (D-Level; Covington et al., 2006), and mean length of clause (MLC; Kallay & Redford, 2020). CD and D-Level have been widely used in basic and clinical research as measures of syntactic development (Cheung & Kemper, 1992; Covington et al., 2006; Fey et al., 2004; Frizelle et al., 2018; Lu, 2009; Rosenberg & Abbeduto, 1987) and have also been shown to distinguish between samples elicited under different discourse contexts (e.g., conversational versus expository; Nippold et al., 2005; Nippold, 2009). D-Level indexes grammatical complexity with an acquisition-based sentence complexity scale ranging from simple sentences to sentences with more than one level of embedding. Each C-unit was manually scored for its D-Level using the Covington et al. (2006) scale in Table 2. The average D-Level was then calculated for each sample. MLC was calculated by dividing the number of total words (NTW) by the total number of independent and dependent clauses in a sample.

Table 2

D-Level Scale (from Covington et al., 2006)

Score	Description	Examples
0	Simple sentences, including questions; sentences with auxiliaries; simple elliptical sentences	"I'm not taking a bath."
1	Infinitive or *-ing* complement with same subject as main clause	"And the boy is going to give the bunny a carrot."
2	Conjoined noun phrases in subject position; sentences conjoined with a coordinating conjunction; conjoined verbal, adjectival, or adverbial constructions	"The mom came over and grabbed the bunny and went home."
3	Relative clause modifying object of main verb; nominalization in object position; finite clause as object of main verb; subject extraposition; raising	"And (some) some are not happy that the bunny is so sneaky."

4	Non-finite complement with its own understood subject; comparative with object of comparison	"I saw him walking the dog."
5	Sentences joined by a subordinating conjunction; nonfinite clauses in adjust positions	"And ask the teacher if I could feed the bunny some carrots."
6	Relative clause modifying subject of main verb; embedded clause/nominalization serving as subject of main verb	"Because the ones who are afraid are surprised."
7	More than one level of embedding in a single sentence	"It is because the teacher told her that there was a bunny."

Note. Examples are from current study data, except #4 for which there were no data examples.

The three measures of lexical diversity included lexical density (LDensity) and the aforementioned MATTR and NDW. LDensity has been used to analyze the complexity of spoken (Johansson, 2009) and written (Hall-Mills & Apel, 2015) narratives. This measure distinguishes between words with lexical properties (i.e., nouns, verbs, adjectives, and adverbs) and those with grammatical properties (i.e., conjunctions, articles, auxiliary verbs, interjections, determiners, and prepositions), counting only those with lexical properties (Johansson, 2009). LDensity was calculated manually for each sample by first running a Grammatical Categories analysis in SALT (i.e., list of parts of speech generated by the GramCats algorithm which is 95.1% accurate, Channell & Johnson, 1999) to determine the number of lexical items (i.e., content words: nouns, verbs, adjectives, and adverbs) and then dividing by the SALT-generated NTW for the sample. The adverb category included adverbs of manner (e.g., "randomly"), time (e.g., "now"), place (e.g., "outside"), frequency (e.g., "usually"), and purpose (e.g., "so that").

MATTR measures lexical diversity by calculating type token ratios (i.e., TTRs, a ratio of unique lexical items divided by the total number of words in a sample) for successive

nonoverlapping segments of a sample (Covington, & McFall, 2010). This is a measure of unique lexemes but is calculated within a moving window size. Since MATTR uses length-controlled windows, it controls for the known sample length problems of other widely used measures (e.g., TTR; Charest et al., 2020). A MATTR closer to 1.0 indicates a varied vocabulary, and a MATTR closer to 0.0 represents a limited, repetitive vocabulary. MATTR was calculated in SALT with a window size of 20 words.

Lastly, NDW was calculated. SALT calculates NDW based on the production of unique free morphemes, so that *play*, *play/ed*, and *play/ing* would be treated as one word occurring three times (Miller et al., 2019). While NDW is used to calculate MATTR, the two measures provide unique information, and both are widely used in research and clinical settings.

Reliability

Measurement reliability depended on the reliability of the transcriptions, C-unit segmentation, and the manual scoring of D-Level. To assure transcription, segmentation, and coding reliability, we used a consensus procedure (as in Guo & Eisenberg, 2015; Shriberg et al., 1984). Each sample was initially transcribed and segmented by the first author. Then, a trained research assistant reviewed the data while listening to the recorded language samples and reading the initial segmented transcriptions. Transcription, segmentation, or coding disagreements were identified, and then reviewed and discussed until agreement was obtained for all transcripts (e.g., Frizelle, et al., 2018).

Each sample was initially scored for D-Level by the first author. The audio files of 12 language samples (10%) were then randomly selected from the data and a research assistant independently rated and scored these samples for D-Level. The raters' scores had a moderate level of agreement in a Cohen's kappa calculation (K = .70; McHugh, 2012). Despite the relative

difficulty of D-Level judgments, the reliability of these and all other measures was judged to be suitable for their subsequent analyses.

Analyses

Analyses were completed using R Studio (Rstudio Team, 2020), a companion program to R (R Core Team, 2018). R packages were used for data management (tidyr; Wickham & Henry, 2018), analysis (lme4; Bates et al., 2015, sjstats; Lüdecke, 2021), and visualization (ggplot2; Wickham, 2016). For the first research question (RQ1), linear mixed effects modeling was used to test for measurement responsiveness; that is, the effect of the two categorical variables, children's age (between-subjects) and discourse context (within-subjects) on MLUm and MLUw. Participant was entered as a random intercept. Parent-reported sex of the participant was entered as a control variable. Tests for significance were obtained through model comparison using the likelihood ratio tests. Interaction terms were removed from the models when not significant. Partial eta-squared was calculated to determine effect sizes, where .01 = a small effect, .06 = a medium effect, and .14 = a large effect (Field, 2013).

To test whether MLU is a valid measure of syntactic development and complexity (RQ2), the other standard measures of syntax (CD, D-Level, MLC) and lexical diversity (LDensity, MATTR, NDW) were entered as predictors into linear mixed effects models that had MLUm or MLUw as the outcome variable. Participant was entered as a random intercept. In this way, the analyses controlled for shared variance amongst the measures due to language ability, allowing us to test for the predicted unique relationship between specific measures of syntax and MLU.

Results

Table 3 displays means, standard deviations, and ranges of sample lengths in duration (minutes), number of C-units, and NTW by age group for the two types of language samples.

The question-answer sample durations include the time that the experimenter was asking questions. The number of C-units reference only language the children produced. One child declined to tell one of the stories and so the total number of language samples was 127 rather than 128. The total number of utterances evaluated was 1147 C-units.

Table 3

Means, Standard Deviations, and Ranges of Sample Lengths

Sample	Duration (minutes)		C-units		NTW	
	M	Range	*M*(*SD*)	Range	*M*(*SD*)	Range
Question-Answers						
5-year-olds	1.23(.02)	.34-2.30	9.47(2.83)	9-20	54.44(22.87)	21-116
8-year-olds	1.13(.02)	.39-2.36	9.31(2.13)	5-14	68.16(29.69)	23-143
Narratives						
5-year-olds	.41(.01)	.21-1.37	7.81(1.99)	4-12	54.58(17.84)	27-104
8-year-olds	.45(.01)	.17-1.57	9.50(3.10)	5-18	81.59(30.37)	36-175
Overall						
5-year-olds	1.03(.03)	.18-2.30	8.65(2.57)	4-20	54.51(20.38)	21-116
8-year-olds	.59(.01)	.17-2.36	9.41(2.64)	5-18	74.88(30.55)	23-175

Note. NTW = Number of Total Words.

Measurement Responsiveness of MLU (RQ1)

Descriptive data for MLUm and MLUw and complete model results for RQ1 are presented in Table 4. Descriptive data show differences by discourse context and age group for MLUm and MLUw.

The main effects of age group and discourse context (see Table 4) were significant for both MLUm and MLUw, as illustrated by the boxplots in Figure 1. The interaction between Group and Context was not significant. The direction of the effects was as follows: the 8-year-old group produced samples with higher MLUs compared to the 5-year-old group (see Table 4); narratives were produced with higher MLUs compared to the question-answer samples (see

Table 4). Effect sizes in partial eta squared ranged from .05 to .12 for Group and from .12 to .29 for Context.

Table 4

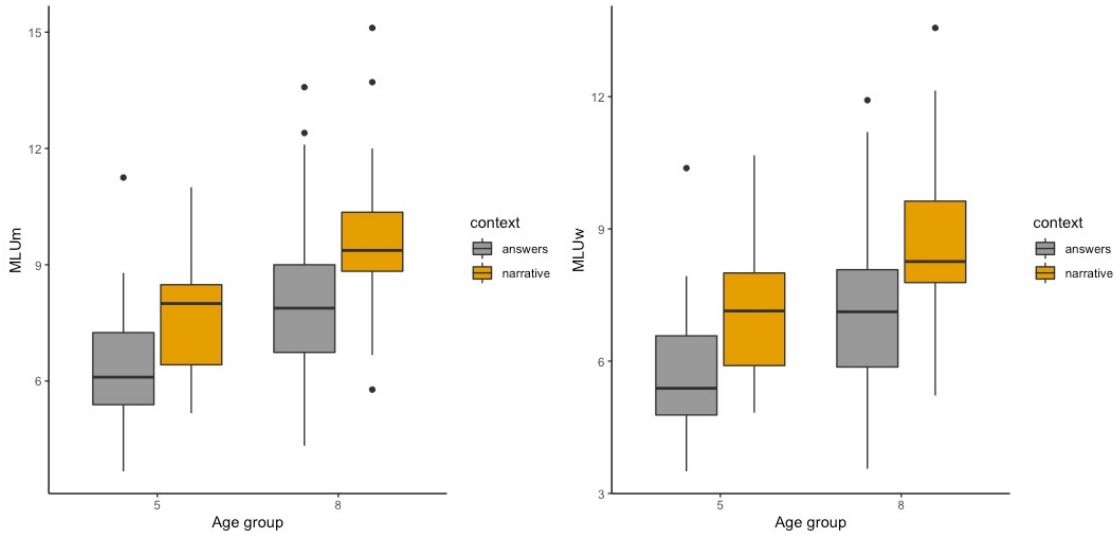
Means, Standard Deviations and Fixed Effects on MLU by Age Group and by Discourse Context

Measure	5-year-olds		8-year-olds						
	M	*SD*	*M*	*SD*	Estimate	SE	χ^2	*p*	ηp^2
MLUm	6.98	1.67	8.73	2.18	1.75	.45	12.05	<.001***	.12
MLUw	6.39	1.57	7.86	1.97	1.47	.42	10.34	.001**	.11
	Question-answers		Narrative						
MLUm	7.11	2.05	8.63	1.93	1.50	.25	31.07	<.001***	.26
MLUw	6.41	1.86	7.87	1.70	1.44	.22	34.75	<.001***	.29

Note. Degree of freedom for all models is 1. "m" indicates morphemes as the unit of measurement; "w" indicates words as the unit of measurement. The interaction between fixed effects is not shown because it was not significant. The reference category for Group was 8-year-olds and for Context was narratives. ** $p \leq .01$, *** $p \leq .001$.

Figure 1

MLU by Age Group and Discourse Context



Note. MLUm (left), MLUw (right).

Construct Validity of MLU (RQ2)

Descriptive data for the three measures of syntactic complexity and three measures of lexical diversity are presented in Table 5 by age group and discourse context. All measure values, except MATTR, were higher in 8-year-olds compared to 5-year-olds as well as higher in the narrative context compared to the question-answer context. Table 6 displays the correlations between the predictor variables. As expected, there is an especially strong correlation between NDW and two of the syntactic complexity measures (i.e., D-Level and CD).

To answer RQ2, the syntactic complexity and lexical diversity measures were used to predict MLUm and MLUw in linear mixed effects models. The results indicated a significant relationship between two of the syntactic measures and MLU: both clausal density (CD) and mean length of clause (MLC) predicted MLUm and MLUw (CD: MLUm, $t = 20.49$, $p < .001$, $\eta p^2 = .79$; MLUw, $t = 26.44$, $p < .001$, $\eta p^2 = .85$; MLC: MLUm, $t = 23.67$, $p < .001$, $\eta p^2 = .83$; MLUw, $t = 33.59$, $p < .001$, $\eta p^2 = .90$). None of the measures of lexical diversity were significantly related to MLU in the overall model, despite a strong bivariate correlation between MLU and NDW: r (32) = .79, $p < .001$. This result shows the importance of controlling for shared variance in the analyses.

Table 5

Means and Standard Deviations for Syntactic Complexity and Lexical Diversity Measures

Measure	5-year-olds		8-year-olds	
	M	*SD*	*M*	*SD*
Syntactic				
Clausal density	1.25	.28	1.50	.32
D-Level	.78	.67	1.39	.85
Mean length clause	5.12	.98	5.24	.81
Lexical				
LDensity	47.45	5.62	45.69	5.06
MATTR	.80	.06	.80	.06
NDW	33.56	10.25	43.23	15.18
	Question-answers		Narrative	
	M	*SD*	*M*	*SD*
Syntactic				
Clausal density	1.36	.36	1.40	.28
D-Level	1.07	.89	1.10	.74
Mean length clause	4.73	.71	5.63	.85
Lexical				
LDensity	45.44	5.82	47.70	4.71
MATTR	.82	.06	.79	.04
NDW	38.36	14.57	38.51	13.10

Table 6

Correlations for Study Variable Means (n = 32)

Variable	1	2	3	4	5	6
1. CD	—					
2. D-Level	.87**	—				
3. MLC	.09	.31	—			
4. LDensity	-.49**	-.54**	-.13	—		
5. MATTR	.40**	.32	.20	-.31	—	
6. NDW	.76**	.76**	.36*	-.34	.62**	—

Note. CD = Clausal Density; D-Level = Developmental Level; MLC = Mean Length Clause; LDensity = Lexical Density; MATTR = Moving Average Type Token Ratio; NDW = Number of Different Words. * $p \leq .05$., ** $p \leq .01$.

Discussion

Overall, the results indicate that MLU is responsive to cross-sectional age-related and context-related differences in the language of young school-aged children. Additionally, the results suggest that MLU is a valid measure of syntactic development in these age groups.

MLUm and MLUw were found to be responsive measures of linguistic complexity. The effect of age group on MLU was in the predicted direction: language samples produced by 8-year-old children had greater MLUs than those produced by 5-year-old children. This result is consistent with the expected development of language; an expectation that was confirmed in the present study with norm-referenced language testing (i.e., raw scores on the CELF-5 were greater in the 8-year-olds). MLU was also different across the narrative and question-answer

discourse contexts, but the direction of this effect ran counter to expectations. Nippold and colleagues (2015) found that MLCU was systematically longer in language samples obtained from children ages 12;10-14;11 during a critical thinking task than in language samples obtained in a conversation or narrative task. This finding led to our prediction that the question-answer context would produce longer MLUs than the narrative context. We expected the question-answer context would require children to think more critically about the picture sequences than they might during a narrative and thus produce longer utterances. However, this prediction did not consider that children in the current study were much younger than those in the Nippold et al. study. It is possible that our prediction would have been born out were we to work with young adolescents who likely have more experience answering critical- thinking questions in a thorough manner. But there are at least two other possible explanations for the unexpected direction of the effect of context on MLU.

First, the systematic difference in MLU by discourse context may reflect the particularly short and formulaic utterances elicited in the question-answer context (e.g., “I don’t know”, “bye”, “sorry”). For example, there were 31 instances of “I don’t know” in the question-answer samples from the 5-year-old group and eight instances in 8-year-old group. Since the questions required higher-level thinking, it is not surprising that the younger children struggled more with answering these types of questions and the result was a perfectly appropriate, but short, “I don’t know” response. In addition to “I don’t know,” there were nine instances of one-word responses in the question-answer context (e.g., “bye” in response to “What is the teacher thinking now?” where the picture shows the teacher waving; “sorry” in response to “What would you say to your mom if you were the girl here?” where the picture shows the mom looking angry). These one-word responses, while short, were nonetheless pragmatically appropriate responses and correct

based on what is shown in the pictures. In a clinical setting, the best practice in this case would be for the tester to prompt for guesses to elicit more language since eliciting high-quality language samples is vital when conducting LSA (Miller et al., 2019). Indeed, elicitation of the best language samples requires the tester to show interest, make natural contributions, and prompt by adding supportive comments (Shipley & McAfee, 2009; Nippold, 2014). In the present study, the different testers did follow up with children, but varied in their level of assertiveness in the follow up. Overall, the question-answer context did elicit different language from the narrative language context, but not in the intended direction because children engaged in pragmatically appropriate behavior by providing just the relevant information in answers to the questions given.

A second possible reason for the unexpected direction of the effect of context on MLU may have been the fixed order in which the tasks were completed. The question-answer task always preceded the narrative elicitation in our study. This order, at variance with SLAM instructions, followed from our previous work. The question-answer context allows the tester to build rapport with the child, helps familiarize the child to the pictures, and helps the child to conceptualize a more coherent and complete narrative than they might otherwise produce (Kallay & Redford, 2020; Redford, 2013). But, by tapping into children's reasoning skills and prompting them to think conceptually prior to producing a narrative, we may have increased the probability that narratives would be produced with more complex syntax, given the relationship between higher-level thinking skills and syntactic complexity (deVilliers & Pyers, 2002; Nippold et al., 2007). This possibility is instructive clinically in that it suggests narrative language samples may provide especially good information about syntactic development if children are encouraged to think through the story they will produce with a question-answer phase preceding narrative

elicitation. Therefore, both tasks of the SLAM should be used with early school age children since they each provide information about the speaker's ability to formulate complex language and comprehend the reasoning and problem-solving type questions.

Whether it is calculated in morphemes or in words, MLU varied systematically with the age of children who produced the language samples under analysis and with the discourse contexts in which the samples were produced. So MLUm and MLUw are responsive measures in typically developing young school-aged children. Since there were no significant differences between MLU results that varied by unit of count (morpheme versus word), we conclude that the unit of count does not matter in these age groups, as other studies have similarly shown (Parker & Brorson, 2005; Rice et al., 2010). However, when an SLP is conducting LSA as part of a language evaluation and suspects issues with morphological development, using MLUm would be beneficial for establishing baselines and identifying errors.

We conducted a linear mixed effects analysis to assess the strength of the relationship between MLU and established measures of syntactic complexity and lexical diversity to determine the construct validity of MLU as a measure of strictly syntactic complexity. Indeed, MLU was found to be a valid measure of syntactic development in young school-aged children: it covaried with measures of syntactic complexity and not with measures of lexical diversity in models that controlled for shared variance between the different measures. Strong correlations between two syntactic complexity measures (CD and DLevel) and two lexical diversity measures (NDW and MATTR) can be interpreted as the measures indexing the same construct. CD and DLevel both relate to the complexity of the utterances, mainly due to subordination, and MATTR and NDW both focus on the variety of vocabulary used by the children. In clinical

practice, these finding suggest the use of MLU to measure syntactic complexity and that additional measures are needed to assess lexical diversity.

Taking a closer look at the relationship between MLU and syntactic complexity, the descriptive data for our measures indicate that as the length of an utterance increases, so too does the number of clauses in the utterance regardless of a child's age. An example, from an 8-year-old's narrative, illustrates the relationship between MLU and clause number: "and then he jumped out 'cause he saw the carrot on the chalkboard" contains just one C-unit, but is composed of 14 morphemes and two clauses. Thus, while developmental increases in syntactic complexity can occur independent of decreasing MLU via subordination (Frizelle et al., 2018), we did not find this to be true in the present study. Children in the older group did engage in more subordination than the children in the younger group, but their utterances were longer due to their use of more adjectives (e.g., "new", "little"), more adverbs (e.g., "so", "really"), and fewer sentence fragments. Overall, the result indicates an important relationship between clausal density and MLU and between mean clause length (MLC) and MLU in these age groups.

The example utterance above suggests that complex syntax, associated with conceptual complexity, is also associated with a conceptual richness reflected in the expressive vocabulary. Indeed, a bivariate correlation between a measure of lexical diversity, NDW, and MLU was strong in the present data. This finding replicates previous work that has reported the correlation between NDW and MLU (Dethorne et al., 2005; Ukrainetz & Blomquist, 2002) – a correlation that suggests a strong association between the developing lexicon and the development of complex syntax. The present study's analyses assessed the hypothesis that this association is due to shared variance – specifically, to overall language ability – rather than to a causal or identity relationship between lexical diversity and syntax. This hypothesis was supported: the full model

results indicated that, once shared variance is accounted for, only measures of syntactic complexity correlate with MLU. We take the resulting correlations between other standard measures of syntax and MLU to indicate an identity relationship, and so again conclude that MLU is a valid measure of syntax in young school-aged children.

Limitations and Future Directions

We find that MLU is responsive to age and discourse context, but we acknowledge the small sample may limit the generalizability of our results to other populations. For example, younger children may require longer samples to show this type of effect since their language production is more variable than older children (Guo & Eisenberg, 2015). This is true for children with delayed speech and language skills as well.

More research is required to validate the use of short samples like those analyzed in our study in a variety of populations, particularly with disordered children, since the primary clinical purpose of LSA is to identify language disorders in children. While the traditional recommendation for LSA is to use samples of 50 utterances or more to assess children's development (Eisenberg et al., 2001; Miller et al., 2019), the results from some studies that investigate the effect of sample size on measurement reliability suggest that shorter samples may also yield good quality assessment information. For example, shorter language samples have been shown to be reliable for certain measures of productivity (i.e., number of total utterances and words per minute), lexical diversity (NDW), and for MLUm (Casby, 2011; Heilmann et al., 2010). It is certainly more practical to elicit short language samples so if these can provide an effective measure of language skills in school-age children, busy clinicians will benefit (Casby, 2011; Heilmann et al., 2010; Heilmann & Malone, 2014; Heilmann et al., 2020).

While the order of the question-answer sample always preceding the narrative samples was intentionally set to build rapport and prompt the children to think about the stories before providing a narrative, the lack of counterbalancing the elicitation of the two language sample types is a limitation. The order utilized in this study may have contributed to the narrative samples being longer than the question-answer samples. Asking the questions after the narrative elicitation task may have produced a different result (i.e., longer answers from the children). Future work requires counterbalancing the language sample types to eliminate these types of effects.

Lastly, it may be of interest to reconsider our conclusion that the unit of count (word versus morpheme) did not make a difference with regards to the findings of responsiveness and construct validity with further research. It is tempting to suggest the use of MLUw with early school-age children since it requires less time and effort than MLUm, however it seems likely that morpheme counts provide important additional information about developmental maturity that is not captured by MLUw alone (e.g., systematic deletion of 3rd person markers).

Acknowledgments

We are grateful to the team at Portland State University, including Jillian Adkins and Briana McColgan, for recruiting participants and collecting the experimentally-controlled language samples analyzed in this study and to Stephanie DeAnda and Kristopher Kyle for extensive comments on a previous version of this manuscript.

Data Availability Statement

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

References

American Speech-Language-Hearing Association. (2016). *Scope of practice in speech-language pathology* [Scope of practice]. https://www.asha.org/policy/

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Bennett, R., & Elfner, E. (2019). The syntax–prosody interface. *Annual Review of Linguistics*, 5, 151-171.

Bernstein, D.K., & Tiegerman-Farber, E. (1997). *Language and communication disorders in children* (4th ed.). Allyn & Bacon.

Bernstein Ratner, N., & Brundage, S. B. (2020). *A clinician's complete guide to CLAN and PRAAT*. https://talkbank.org/manuals/Clin-CLAN.pdf

Blake, J., Quartaro, G., & Onorati, S. (1993). Evaluating quantitative measures of grammatical complexity in spontaneous speech samples. *Journal of Child Language*, *20*(1), 139–152. https://doi.org/10.1017/S0305000900009168

Brown, R. (1973). *A first language: The early stages*. Harvard University Press.

Casby, M. W. (2011). An examination of the relationship of sample size and mean length of utterance for children with developmental language impairment. *Child Language Teaching and Therapy*, *27*(3), 286–293. https://doi.org/10.1177/0265659010394387

Castilla-Earls, A., Bedore, L., Rojas, R., Fabiano-Smith, L., Pruitt-Lord, S., Restrepo, M. A., & Peña, E. (2020). Beyond scores: Using converging evidence to determine speech and language services eligibility for dual language learners. *American Journal of Speech-Language Pathology*, *29*(3), 1116–1132. https://doi.org/10.1044/2020_AJSLP-19-00179

Channell, R. W., & Johnson, B. W. (1999). Automated grammatical tagging of child language samples. *Journal of Speech, Language, and Hearing Research*, *42*(3), 727–734. https://doi.org/10.1044/jslhr.4203.727

Charest, M., Skoczylas, M. J., & Schneider, P. (2020). Properties of lexical diversity in the narratives of children with typical language development and developmental language disorder. *American Journal of Speech-Language Pathology*, 1–17. https://doi.org/10.1044/2020_AJSLP-19-00176

Cheung, H., & Kemper, S. (1992). Competing complexity metrics and adults' production of complex sentences. *Applied Psycholinguistics*, *13*(1), 53–76. https://doi.org/10.1017/S0142716400005427

Covington, M. A., He, C., Brown, C., Naçi, L., & Brown, J. (2006). *How complex is that sentence? A proposed revision of the Rosenberg and Abbeduto D-Level scale.* (CASPR Research Report 2006-01). The University of Georgia, Artificial Intelligence Center.

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, *17*(2), 94–100. https://doi.org/10.1080/09296171003643098

Crookes, G. V., & Rulon, K. (1985). *Incorporation of corrective feedback in native speaker/non-native speaker conversation* (Technical Report No. 3). University of Hawaii: Honolulu.

Crystal, D. C., & Davy, D. (1975). *Advanced Conversational English*. Longman.

Crowley, C. & Baigorri, M. (2015). *School-age language assessment measures*. LEADERSproject. https://www.leadersproject.org/disability-evaluation/school-age-language-assessment-measures-slam/

de Villiers, J. G., & Pyers, J. E. (2002). Complements to cognition: A longitudinal study of the relationship between complex syntax and false-belief-understanding. *Cognitive Development*, *17*(1), 1037–1060. https://doi.org/10.1016/S0885-2014(02)00073-4

Davis-Kean, P. E. (2005). The influence of parent education and family income on child achievement: The indirect role of parental expectations and the home environment. *Journal of Family Psychology*, *19*(2), 294–304. https://doi.org/10.1037/0893-3200.19.2.294

Dethorne, L. S., Johnson, B. W., & Loeb, J. W. (2005). A closer look at MLU: What does it really measure? *Clinical Linguistics & Phonetics*, *19*(8), 635–648. https://doi.org/10.1080/02699200410001716165

Diessel, H., & Tomasello, M. (2001). The development of relative clauses in spontaneous child speech. *Cognitive Linguistics*, *11*(1–2). https://doi.org/10.1515/cogl.2001.006

Dodd, B., Zhu, H., Crosbie, S., Holm, Alison, & Ozanne, A. (2002). *Diagnostic Evaluation of Articulation and Phonology*. Pearson Assessments.

Eisenberg, S. L., Fersko, T. M., & Lundgren, C. (2001). The use of MLU for identifying language impairment in preschool children: A review. *American Journal of Speech-Language Pathology*, *10*(4), 323–342. https://doi.org/10.1044/1058-0360(2001/028)

Fey, M. E., Catts, H. W., Proctor-Williams, K., Tomblin, J. B., & Zhang, X. (2004). Oral and written story composition skills of children with language impairment. *Journal of Speech, Language, and Hearing Research*, *47*(6), 1301–1318. https://doi.org/10.1044/1092-4388(2004/098)

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Sage.

Foster, P. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*(3), 354–375. https://doi.org/10.1093/applin/21.3.354

Frizelle, P., Thompson, P. A., McDonald, D., & Bishop, D. V. M. (2018). Growth in syntactic complexity between four years and adulthood: Evidence from a narrative task. *Journal of Child Language*, *45*(5), 1174–1197. https://doi.org/10.1017/S0305000918000144

Guo, L.-Y., & Eisenberg, S. (2015). Sample length affects the reliability of language sample measures in 3-year-olds: Evidence from parent-elicited conversational samples. *Language, Speech, and Hearing Services in Schools*, *46*(2), 141–153. https://doi.org/10.1044/2015_LSHSS-14-0052

Hall-Mills, S., & Apel, K. (2015). Linguistic feature development across grades and genre in elementary writing. *Language, Speech, and Hearing Services in Schools*, *46*(3), 242–255. https://doi.org/10.1044/2015_LSHSS-14-0043

Heilmann, J., & Malone, T. O. (2014). The rules of the game: Properties of a database of expository language samples. *Language, Speech, and Hearing Services in Schools*, *45*(4), 277–290. https://doi.org/10.1044/2014_LSHSS-13-0050

Heilmann, J., Nockerts, A., & Miller, J. F. (2010). Language sampling: Does the length of the transcript matter? *Language, Speech, and Hearing Services in Schools*, *41*(4), 393–404. https://doi.org/10.1044/0161-1461(2009/09-0023)

Heilmann, J., Tucci, A., Plante, E., & Miller, J. F. (2020). Assessing functional language in school-aged children using language sample analysis. *Perspectives of the ASHA Special Interest Groups*, *5*(3), 622–636. https://doi.org/10.1044/2020_PERSP-19-00079

Hewitt, L. E., Hammer, C. S., Yont, K. M., & Tomblin, J. B. (2005). Language sampling for kindergarten children with and without SLI: Mean length of utterance, IPSYN, and NDW.

Journal of Communication Disorders, *38*(3), 197–213. https://doi.org/10.1016/j.jcomdis.2004.10.002

Hoff, E. (2014). *Language development* (5th ed.). Wadsworth Cengage Learning.

Jefferson, G. (1984). Transcript notation. In J. Atkinson & J. Heritage (Eds.), *Structures of social interaction: Studies in conversation analysis* (pp. 134-162). Cambridge: Cambridge University Press.

Johansson, V. (2009). Lexical diversity and lexical density in speech and writing: A developmental perspective. *Lund Working Papers in Linguistics*, *53*, 61–79.

Kallay, J. E., & Redford, M. A. (2020). Clause-initial AND usage in a cross-sectional and longitudinal corpus of school-age children's narratives. *Journal of Child Language*, 1–22. https://doi.org/10.1017/S0305000920000197

Klee, T., & Fitzgerald, M. D. (1985). The relation between grammatical development and mean length of utterance in morphemes. *Journal of Child Language*, *12*(2), 251–269. https://doi.org/10.1017/S0305000900006437

Loban, W. (1976). *Language development: Kindergarten through grade twelve*. National Council of Teachers of English.

Lu, X. (2009). Automatic measurement of syntactic complexity in child language acquisition. *International Journal of Corpus Linguistics*, *14*(1), 3–28. https://doi.org/10.1075/ijcl.14.1.02lu

Lüdecke, D. (2018). *Sjstats: Statistical Functions for Regression Models.* Zenodo. https://doi.org/10.5281/ZENODO.1284472

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282.

Miller, J.F., Andriacchi, K., & Nockerts, A., (2019). *Assessing Language Production Using SALT Software: A Clinician's Guide to Language Sample Analysis* (3rd ed.). SALT Software, LLC.

Miller, J. F., & Chapman, R. S. (1981). The relation between age and mean length of utterance in morphemes. *Journal of Speech, Language, and Hearing Research*, *24*(2), 154–161. https://doi.org/10.1044/jshr.2402.154

Miller, J. F., Freiberg, C., Holland, M.-B., & Reeves, M. A. (1992). Implementing computerized language sample analysis in the public school. *Topics in Language Disorders*, *12*(2), 69–82. https://doi.org/10.1097/00011363-199202000-00008

Miller, J.F., & Iglesias, A. (2019). *Systematic Analysis of Language Transcripts (SALT), Student Version 20-1*. [Computer software]. Madison, WI: SALT Software, LLC.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. W. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology*, *63*(7), 737–745. https://doi.org/10.1016/j.jclinepi.2010.02.006

Moyle, M. J., Karasinski, C., Ellis-Weismer, S., & Gorman, B. K. (2011). Grammatical morphology in school-age children with and without language impairment: A discriminant function analysis. *Language, Speech, and Hearing Services in Schools*, *42*(4), 550–560.

Nespor, M., & Vogel, I. (2012). *Prosodic phonology*. De Gruyter Mouton.

Nippold, M. A. (2009). School-age children talk about chess: Does knowledge drive syntactic complexity? *Journal of Speech, Language, and Hearing Research*, *52*(4), 856–871. https://doi.org/10.1044/1092-4388(2009/08-0094)

Nippold, M. A. (2014). *Language sampling with adolescents: Implications for intervention* (2nd ed.). Plural Publishing Inc.

Nippold, M. A., Frantz-Kaspar, M. W., Cramond, P. M., Kirk, C., Hayward-Mayhew, C., & MacKinnon, M. (2015). Critical thinking about fables: Examining language production and comprehension in adolescents. *Journal of Speech, Language, and Hearing Research*, *58*(2), 325–335. https://doi.org/10.1044/2015_JSLHR-L-14-0129

Nippold, M. A., Hesketh, L. J., Duthie, J. K., & Mansfield, T. C. (2005). Conversational versus expository discourse: A study of syntactic development in children, adolescents, and adults. *Journal of Speech, Language, and Hearing Research*, *48*(5), 1048–1064. https://doi.org/10.1044/1092-4388(2005/073)

Nippold, M. A., Mansfield, T. C., & Billow, J. L. (2007). Peer conflict explanations in children, adolescents, and adults: Examining the development of complex syntax. *American Journal of Speech-Language Pathology*, *16*(2), 179–188. https://doi.org/10.1044/1058-0360(2007/022)

Parker, M. D., & Brorson, K. (2005). A comparative study between mean length of utterance in morphemes (MLUm) and mean length of utterance in words (MLUw). *First Language*, *25*(3), 365–376. https://doi.org/10.1177/0142723705059114

Paul, R., Hernandez, R., Taylor, L., & Johnson, K. (1996). Narrative development in late talkers: Early school age. *Journal of Speech, Language, and Hearing Research*, *39*(6), 1295–1303. https://doi.org/10.1044/jshr.3906.1295

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Redford, M. A. (2013). A comparative analysis of pausing in child and adult storytelling. *Applied Psycholinguistics*, *34*(3), 569–589. https://doi.org/10.1017/S0142716411000877

Rescorla, L., Dahlsgaard, K., & Roberts, J. (2000). Late-talking toddlers: MLU and IPSyn outcomes at 3;0 and 4;0. *Journal of Child Language*, 27(3), 643-664. https://doi.org/10.1017/S0305000900004232

Rice, M. L., Smolik, F., Perpich, D., Thompson, T., Rytting, N., & Blossom, M. (2010). Mean length of utterance levels in 6-month intervals for children 3 to 9 years with and without language impairments. *Journal of Speech, Language, and Hearing Research*, *53*(2), 333–349. https://doi.org/10.1044/1092-4388(2009/08-0183)

Rondal, J. A., Ghiotto, M., Bredart, S., & Bachelet, J.-F. (1987). Age-relation, reliability and grammatical validity of measures of utterance length. *Journal of Child Language*, *14*(3), 433–446. https://doi.org/10.1017/S0305000900010229

Rosenberg, S., & Abbeduto, L. (1987). Indicators of linguistic competence in the peer group conversational behavior of mildly retarded adults. *Applied Psycholinguistics*, *8*(1), 19–32. https://doi.org/10.1017/S0142716400000047

Rstudio Team. (2020). *Rstudio: Integrated Development for R*. Rstudio, PBC. http://www.rstudio.com

Shipley, K. G., & McAfee, J. G. (2009). *Assessment in speech-language pathology: A resource manual* (4th ed). Delmar Cengage Learning.

Shriberg, L. D., Kwiatkowski, J., & Hoffmann, K. (1984). A procedure for phonetic transcription by consensus. *Journal of Speech, Language, and Hearing Research*, *27*(3), 456–465. https://doi.org/10.1044/jshr.2703.456

Scott, C. M. (1988). Producing complex sentences. *Topics in Language Disorders*, *8*(2), 44–62. https://doi.org/10.1097/00011363-198803000-00006

Ukrainetz, T. A., & Blomquist, C. (2002). The criterion validity of four vocabulary tests compared with a language sample. *Child Language Teaching and Therapy*, *18*(1), 59–78. https://doi.org/10.1191/0265659002ct227oa

Watkins, R. V., Kelly, D. J., Harbers, H. M., & Hollis, W. (1995). Measuring children's lexical diversity: Differentiating typical and impaired language learners. *Journal of Speech, Language, and Hearing Research*, *38*(6), 1349–1355. https://doi.org/10.1044/jshr.3806.1349

Wickham, H. (2016). *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag.

Wickham, H., & Henry, L. (2018). *Tidyr: Easily Tidy Data with "spread()" and "gather()" Functions* [R package version 0.8.1]. https://CRAN.R-project.org/package=tidyr

Wiig, E., Semel, E., & Secord, W. (2013). *Clinical Evaluation of Language Fundamentals-Fifth Edition* (5th ed.). Pearson Assessments.

Appendix A: Question Prompts for Each Story

Questions for *Dog Comes Home*

1. What do you think the girl is thinking here?
2. Why do you think she's putting the dog in her bag?
3. Why do you think the girl's getting so dirty?
4. Why is there a white dog in the bathtub now?
5. What do you think the mother's going to do now?
6. What would you say to the mom if you were the girl here?

Questions for *Bunny Goes to School*

1. Why do you think the bunny jumped out of the backpack?
2. Why do you think some students are afraid and some students are laughing?
3. What would you do if a bunny came to your school?
4. What was the boy's idea?
5. How did the mom know to come to the school?
6. Why do you think the mom came to the school?
7. What do you think will happen when the boy when he goes home?
8. What do you think the teacher's thinking now?

Appendix B: Sample Narratives

Narrative by a 5-year-old participant for *Dog Comes Home*

She found a little dog under there.
And then (she) it started licking her.
And then it got her clothes all dirty.
And then she put it in her bag.
And then she told the dog to be quiet.
And then she goed in bathtub.
And then they were washing off.

Narrative by an 8-year-old participant for *Dog Comes Home*

A girl found a dog under a porch.
She found it.
and she made friends with it.
but she thought she wouldn't be able to bring it home.
Because her mom didn't want her to maybe.
I don't really know probably cause someone in her family's allergic to it.
I don't really know.
(um tries to she tries to) she smuggles the dog in in her backpack.
(um) her mom tells her to have a bath cause she's so dirty.
The dog comes out of the backpack and hops in the bath (or bath).
(um) and the mom is angry.
And it looks like she's swearing.