

SPLIT-HALF RELIABILITY OF MLU AND MLU2 IN TWO METHODS OF
UTTERANCE SEGMENTATION

by

Alyse Kemeny

A thesis submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Master of Science

Department of Communication Disorders

Brigham Young University

December 2007

Copyright © 2007 Alyse Kemeny

All rights reserved

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a thesis submitted by

Alyse Kemeny

This thesis has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

Date

Ron W. Channell, Chair

Date

Martin Fujiki

Date

Shawn Nissen

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the thesis of Alyse Kemeny in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

Date

Ron W. Channell
Chair, Graduate Committee

Accepted for the Department

Ron W. Channell
Graduate Coordinator

Accepted for the College

K. Richard Young
Dean, David O. McKay School of Education

ABSTRACT

SPLIT-HALF RELIABILITY OF MLU, MLU2, AND MEDIAN LENGTH OF UTTERANCE IN TWO METHODS OF UTTERANCE SEGMENTATION

Alyse Kemeny

Department of Communication Disorders

Master of Science

Concerns regarding Mean Length of Utterance (MLU) has led to adaptations of this method of analysis. A recent study by Johnston (2001) introduced an alternative to MLU called MLU2. The current study investigates the split-half reliability of MLU and MLU2 as well as another alternative, Median Length of Utterance (Med-LU). Split-half reliability was found for these methods when segmented into Phonological and Communication Units. Split-half reliability of MLU2 was generally higher than that of MLU, and both were higher than Med-LU. The study suggests that MLU2 may also be a valuable tool for clinicians in analyzing child language.

ACKNOWLEDGMENTS

I am so grateful for all of the many people who have helped me to finish this project. Thank you Dr. Channell for all of your hard work, patience and guidance throughout this process. Thanks to all of my classmates for their advice and friendship over the last two years. I am so grateful for the love and support of my family, especially my wonderful parents and parents-in-law. I could not have made it through without the love and support of my sweetheart Carlos whose many prayers and words of encouragement have made this possible. I'm grateful for his sacrifice of many months apart while I completed my education. I am most especially grateful to my Father in Heaven for the many tender mercies I have felt all along the way.

TABLE OF CONTENTS

	Page
Introduction.....	1
Review of Literature	5
Utterance Segmentation.....	5
History of MLU.....	8
Alternatives to MLU.....	14
Median Length of Utterance.....	16
Reliability of MLU	16
Summary.....	17
Method	18
Participants	18
Procedure.....	18
SALT format.....	18
P-units	18
C-units.....	19
Data Analysis.....	19
Results.....	21
Table 1: Descriptive Statistics for P-Units and C-Units.....	21
C-Units vs. P-Units.....	21

Split-half Reliability	22
Age	22
Discussion.....	24
References.....	27

Introduction

Clinicians who use language samples to assess child language are confronted with several unresolved issues. Among these issues are the questions of how to segment individual utterances within the language sample and which utterances to include when analyzing the sample. The present study examines these issues primarily in relation to the Mean Length of Utterance (MLU) measure and an alternative MLU calculation called MLU2 (Brown, 1973; Johnston, 2001).

Since the Brown (1973) study, MLU has been one of the most widely used methods of the quantitative assessment of the development of child syntax (Eisenberg, Fersko, & Lundgren, 2001; Hickey, 1991; Klee & Fitzgerald, 1985). Although popular among researchers and clinicians, MLU is not without its limitations. Several researchers have become frustrated with the ambiguity of the methodology for calculating MLU and have concerns as well regarding the procedure's reliability.

As MLU is the average of the number of morphemes in a child's utterances, it is highly dependent upon decisions which affect the length of those utterances. Two different methods exist for how language samples are segmented into utterances. One method separates utterances into phonological units (P-units; Miller, 2004). A P-unit is described as an utterance that includes a complete thought. This division relies mostly upon prosodic features to determine when the utterance is complete. To prevent run-ons, however, a P-unit cannot contain more than two conjoined independent clauses with any dependent clauses. Unless noted otherwise, researchers and clinicians assume that utterances are segmented according to P-units (Chapman, 1981; Miller, 2004).

An alternate way to segment utterances is called the communication unit (C-unit; Loban, 1976). A C-unit differs from the P-unit in that C-units cannot include more than one independent clause and any related dependent clauses (Chapman, 1981; Loban, 1976). A C-unit does not consider the prosodic contour of the utterance, rather, it considers only the utterance's grammatical structure, as it separates each independent clause into separate C-units.

C-unit segmentation could highlight information regarding a child's syntactic abilities. Children frequently use conjunctions such as *and* to conjoin their utterances and often do not use pauses between utterances until after using a conjunction, indicating that their message is not yet complete. These language style patterns may lead to overestimation of a child's MLU when language samples are segmented into P-units, as P-units allow two or more independent clauses to be conjoined. C-units may provide a viable alternative that allows a clinician to look at the length of each individual independent clause with its modifiers. Thus, the length of C-units will not depend on these speaking habits of conjoining several sentences without pausing, but will reflect a child's true length of individual utterances, defined as an independent clause and its modifiers (Loban, 1976). Loban considered C-units more productive in informing the clinician about a child's language growth and abilities than the P-unit. However, research has not yet examined these claims, such as by comparing the reliability of samples segmented using these two methods. Differences in observed reliability might promote the use of one of these segmentation methods over the other.

Researchers dissatisfied with the ambiguity of the methodology of MLU have also proposed alternatives or adaptations to MLU calculation to provide generalization

and increased reliability to the procedure. Several studies have recommended the use of word instead of morpheme counts (Arlam-Rupp, van Niekirk-de Hahn, & van de Sandt-Koenderman, 1976; Hickey, 1991; Rom & Leonard, 1990). Chapman (1981) encouraged calculation of a second MLU that would exclude imitative utterances, single-word utterances, and answers to questions. Johnston (2001) followed up on this proposal and created an alternative MLU called MLU2. Johnston suggested that this alternative method might provide a better representation of a child's language abilities by excluding one-word answers, self-repetitions, and imitations of a conversational partner's utterances, all of which could affect the obtained size of the MLU and either underestimate or overestimate a child's actual language abilities. However, little is known regarding the MLU2, such as its reliability, and further study is necessary to validate the use of this method.

In Johnston's (2001) study, the Systematic Analysis of Language Transcripts (SALT; Miller & Chapman, 2004) software was used to compute MLU; however, researchers were required to exclude individual utterances by hand prior to computing MLU2. This manual culling of utterances would require clinicians to spend more time and require more effort to compute MLU2. More recent versions of the SALT software allow the clinician to exclude these types of utterances automatically, prior to calculating MLU. The clinician can then efficiently calculate both MLU and MLU2 using the SALT software without spending more time to exclude certain types of utterances by hand.

Another possible alternative to MLU that has not yet been explored is the calculation of the median length of utterance (Med-LU). Med-LU has been briefly mentioned in the literature for studies related to stuttering (Logan & Conture, 1995;

Melnick & Conture, 2000). However, Med-LU has not yet been used as a measure comparable to MLU. Med-LU could provide a better representation of a child's typical length of utterance, as the median is less affected by extreme scores than the mean (Mendenhall, Beaver, & Beaver, 2005). Med-LU has been proposed as a viable alternative to MLU, however, no data for this method have been reported (Eisenberg et al., 2001).

Understanding the reliability of a particular measure is necessary in order to justify its use. Several studies have discussed the reliability of MLU (Chabon, Kent-Udolf, & Egolf, 1982; DeThorne, Johnson, & Loeb, 2005; Klee & Fitzgerald, 1985; Rice, Redmond, & Hoffman, 2006; Rondal, Ghiotto, Bredart, & Bachelet, 1987); however, little is known regarding the reliability of MLU2 or Med-LU. Split-half reliability examines the intra-sample reliability of these measures. Other types of reliability, such as temporal reliability, may require more than one language sample, and the comparison of multiple language samples is inevitably influenced by extraneous factors, as no two samples are collected under exactly the same conditions. Analyzing the split-half reliability of MLU, MLU2, and Med-LU will allow a clear comparison of these measures without the influence of extraneous variables.

The purpose of the present study is to compare the split-half reliability of MLU, MLU2, and Med-LU when language samples have been segmented into either P-units or C-units.

Review of Literature

This review examines the two commonly used methods of utterance segmentation that will be examined in the present study. Next, the research on MLU is reviewed. Finally, a summary of suggested alterations to MLU resulting from researchers' concerns with the methodology and reliability of MLU is given.

Utterance Segmentation

In the book introducing the currently used MLU procedure, Brown (1973) did not define how he segmented the utterances used in his research. Crystal (1974) criticized Brown for this ambiguity in definition, indicating frustration with not being able to replicate Brown's study. Chapman (in Miller, 1981) clarified rules for calculating MLU and recommended using the terminal intonation contour as the criterion for segmenting utterances. Typically, utterances in language samples are transcribed following this phonological criterion, which Loban (1976) termed the P-unit. The determination of a P-unit depends on the presence of a pause and the use of inflection in the speaker's voice that indicates the completion of a thought.

Other techniques for segmenting utterances also exist. Hunt (1965) defined the T-unit for analyzing written language as a main clause with any subordinate clauses (Klecan-Aker & Hedrick, 1985). This same idea was adapted for spoken language, and the name was changed to the C-unit by Loban in 1976. The C-unit followed Hunt's definition of the T-unit, and was defined as an independent clause with all of its modifiers in spoken language. The C-unit cannot be further segmented without resulting in a fragmented sentence. Scott (1988) provided a similar definition of the C-unit, and further explains that clauses beginning with a coordinating conjunction count as a separate

utterance, except when the subject of the sentence is not given in the coordinating clause. Scott provided the following example to illustrate when coordinating conjunctions should not be separated: *the boy went to the store and bought some Coke* (p. 48). This same example would be considered two separate utterances had the speaker stated: *the boy went to the store and he bought some Coke*.

The definitions of C-units and P-units used in the present study come from the SALT manual (Miller & Chapman, 2004). The SALT manual states that P-units are primarily segmented by documenting the completion of a thought, using either rising or falling intonation, and also the presence of a pause. However, to avoid run-on utterances, only two independent clauses may be conjoined in one P-unit. Conversely, segmentation of C-units strictly follows the grammatical-rule based definition proposed by Loban (1976) and Hunt (1965) as “an independent clause and its modifiers” (Miller, 2004, p. 48). C-units do not depend on a pause or the intonation of the speaker, but simply on grammatical rules.

Although language sample transcription is a common practice in both research and clinical settings, little attention is paid to ensure that everyone transcribes language samples the same way, and a variety of ways to transcribe samples exist (Reed, MacMillan, & McLeod, 2001). Reed et al. examined the effects of how different methods of segmenting utterances affected the syntactic analyses of language samples, including (a) MLU in words, (b) MLU in morphemes, (c) number of dependent clauses, (d) number of independent clauses, (e) number of dependent clauses per utterance, (f) number of independent clauses per utterance, and (g) number of utterances in the sample. Reed et al. contrasted the differences among four definitions of utterance: (a) Developmental

Sentence Scoring (DSS; Lee, 1974), (b) T-unit (Hunt, 1965), (c) C-unit (Loban, 1976), and (d) Tone unit, as used in the Language Assessment, Remediation and Screening Procedure (LARSP; Crystal, Fletcher, & Garman, 1976). Of these methods, segmentation according to DSS is closest to the definition of a P-unit. Reed et al. found statistically significant differences for samples segmented according to the DSS method, as DSS allows utterances to include coordinated clauses, which results in longer utterances. The other three types of segmentation require the separation of coordinated clauses. T-unit and C-unit definitions differed significantly in relation to the number of utterances in a sample. Tone units influenced the number of dependent and independent clauses in the language samples, by identifying more independent clauses and less dependent clauses.

Though utterance segmentation has not received attention in many articles concerning MLU calculation, several researchers have noted difficulties with segmentation. For example, Klee and Fitzgerald (1985) stated that they did not attempt to use a formal definition of utterance in their study on the reliability and validity of MLU. The two authors compared their own views on utterance segmentation on 25% of their data and resolved differences through discussion. Klee and Fitzgerald explained that they segmented utterances according to “major clausal syntactic units, intonation contours, pauses and speaker turns,” but this does not provide enough information to replicate the study (p. 255). Another example is Miller and Chapman (1981) who state only that utterances in the samples used for their study were segmented mostly by terminal intonation contour, but they did not provide any further explanations of this rule. Eisenberg (2001) expressed concern that the variability in utterance segmentation would greatly impact the validity of MLU, as differences in segmentation would lead to varying

numbers of utterances within the same sample, and the calculation of MLU depends upon the number of utterances in a sample. Thus, individuals trained in different methods of utterance segmentation would have differing results for MLU calculation.

History of MLU

In 1973 Brown promoted a new method for quantifying language development in morphemes called the MLU. Brown described MLU as “an excellent simple index of grammatical development because almost every new kind of knowledge increases length” (p.53). MLU counts morphemes to find the average length of utterances in a language sample. Brown felt that this would provide a better match than age for child speech. Brown divided MLU values into five arbitrary stages that represent the order of development of “grammatical complexity of constructions” (p. 59). Since 1973, MLU has been widely used and examined in both research and clinical settings.

In a review of Brown’s book, Crystal (1974) criticized Brown for not having a clearly defined methodology for computing MLU, specifically, the lack of a clear definition of an utterance. Crystal stated that he could not segment the utterances due to Brown’s ambiguity in how he segmented the utterances used in his original data. Crystal disapproved of MLU due to its inconsistencies and need for ad hoc decisions. Crystal later developed his own technique of analyzing language samples based on adult speech called LARSP (Crystal et al., 1976).

A number of studies have looked at the application of MLU to be used in languages other than English. Arlman-Rupp et al. (1976) studied MLU in Dutch children. Dromi and Berman (1982) modified MLU to calculate morphemes per utterance in Hebrew. Hickey’s (1991) study looked at applying MLU to the acquisition of Irish.

Thordardottir and Weismer (1998) adapted MLU for assessing language development in Icelandic. Klee, Stokes, Wong, Fletcher, and Gavin (2004) studied MLU in Cantonese-speaking children with and without Specific Language Impairment (SLI).

Rondal and DeFays (1978) examined the question of an adequate sample size for MLU. Language samples from 40 children, ages ranging from 1;8 to 2;8, (years; months) were used to compute MLU values for progressively longer blocks of utterances. Rondal and DeFays compared reliability of language samples increasing in length in blocks of 25 utterances from 25 to 200 utterances. Reliability scores exceeded .80 for samples of 50 utterances or more. Results indicated that sample sizes beyond 50 utterances only slightly improved MLU reliability.

Schachter, Shore, Hodapp, Chalfin, and Bundy (1978) studied difference in MLU between male and female children. Schachter et al. found that toddler girls have significantly longer MLUs than boys of similar age, class, and race. Schachter et al. concluded that girls developed language earlier than boys.

A significant correlation between age and MLU was found by Miller and Chapman (1981) in 123 children of ages 1;5 to 4;11. Miller and Chapman found an increase of variance of predicted MLU as age increased. MLU's linear relationship with age was significant up to four years of age. However, Miller and Chapman cautioned the reader not to make clinical decisions based solely upon MLU, and that MLU should be used only as a "general indicator" of syntactic development (p. 157). Further cautions regarding MLU's sensitivity to contextual variables were also noted, including: "the nature of the interaction, the person with whom the child is interacting, materials present, and the intent of the language addressed to the child" (p. 158). Miller and Chapman's

study has been cited as a reference for data on MLU and for clarification in the computation of MLU (Eisenberg et al., 2001). Several other studies have looked at this same question of how well age and MLU are related (Blake, Quartaro, & Onorati, 1993; Chan, McAllister, & Wilson, 1998; Conant, 1987; Klee & Fitzgerald, 1985; Klee, Schaffer, May, Membrino, & Mougey, 1989).

Chabon, Kent-Udolf, and Egolf (1982) examined the temporal reliability, inter-examiner reliability, and intra-examiner reliability of MLU in children beyond Brown's Stage V. Thirty typically developing children were divided into three equal groups, separated by age: 3;6 (year; month) to 4;6, 5;6 to 6;6, and 8;6 to 9;6. The children participated in conversational interviews and picture description tasks three days in a row. The study concluded that MLU was not temporally reliable for children beyond Stage V though inter-examiner and intra-examiner reliability was high.

Klee and Fitzgerald (1985) analyzed the value of MLU beyond Brown's Stage II (Brown, 1973) on 18 typically developing children between 2 and 4 years in age. Klee and Fitzgerald found no significant correlation with age and concluded that MLU did not successfully distinguish differences in grammatical development of the children. As part of their study, Klee and Fitzgerald looked at the intra-sample variability of MLU by estimating the standard error of the mean to determine the fluctuation of MLU scores within a language sample. A child's linguistic stage, as defined by Brown, could vary two to three stages, depending upon which 100-utterance block in a language sample was used to compute the MLU.

Conant (1987) re-examined Klee and Fitzgerald's (1985) data and discovered higher correlations between age and MLU for 3-year-old children. Conant concluded that

Klee and Fitzgerald's data did in fact correspond with findings of Miller and Chapman (1981) wherein age and MLU were not related for 2-year-old children, but were correlated for 3-year-olds.

Bountress, Bountress, and Tonelson (1988) conducted a study on the effects of race on MLU values. MLU values of 42 African American and Caucasian children were compared when examined by either an African American or Caucasian clinician. No significant differences were found.

Klee et al. (1989) examined the relationship between age and MLU in 48 children between the ages of 2;0 and 4;2. Some 24 of the children were typically developing, and the other 24 had been diagnosed with SLI. Klee et al. found results similar to Miller & Chapman (1981) that age and MLU were highly correlated for both groups. Other findings indicated that the predicted MLU values for the SLI group were lower than their typically developing peers, but the rate that MLU changed in each group was alike.

Scarborough, Rescorla, Tager-Flusberg, and Sudhalter (1991) used MLU to predict syntactic complexity of language samples from typically developing preschoolers, ages 2;0 to 4;0, and from children and adolescents with delayed language, Fragile X syndrome, Down syndrome, and autism. MLU was compared to the Index of Productive Syntax (IPSyn) and strong correlations were found for MLU values between 1.0 and 4.5 for normal and the disordered populations; however, the correlation was weaker after 3.0 morpheme level. Scarborough et al. concurred with Rondal et al. (1987) that MLU is not as reliable for an estimate of syntactic complexity above 3.0 morphemes. Furthermore, in the disordered population, MLU appeared to overestimate the complexity of subjects' syntax more frequently than that of typically developing subjects. Likewise, Rescorla,

Dahlsgaard, and Roberts (2000) found a high correlation between MLU and IPSyn in children identified as late-talking toddlers between 3;0 and 4;0.

Klee (1992) reviewed studies of MLU and examined other quantitative measures of language sampling as well. Klee examined six different measurements, including: (a) the total number of utterances, (b) the total number of complete and intelligible utterances, (c) mean syntactic length (MSL), (d) the total number of words, (e) the number of different words, and (f) the type-token ratio. Klee investigated how well these procedures correlated with age in typically developing children and children with SLI, and how well the measure distinguished between these two groups of children. Klee found that the total number of utterances, the total number of complete and intelligible utterances, and the type-token ratio did not change with age or differentiate between groups. However, alterations of the type-token ratio that were based on a constant number of words instead of utterances were correlated with age in both groups, but could not distinguish between the two groups. The MSL, the total number of words, and the number of different words were correlated with age and could statistically differentiate between typically developing children and children with SLI.

Several studies have been conducted to examine MLU in disordered populations. MLU has been studied in children with developmental delays (Yoder, Spruytenburg, Edwards, & Davies, 1995). Further studies include those focusing on the MLU values of children with autism (Condouris, Meyer, & Tager-Flusberg, 2003), SLI (Dunn, Flax, & Sliwinski, 1996; Hewitt, Hammer, Yont, & Tomblin, 2005; Johnston, 2001; Klee, et al. 1989; Miller & Deevy, 2003), and Down Syndrome (Harris, 1983; Rondal, Ghiotto, Bredart, & Bachelet, 1988; Miles, Chapman, & Sindberg, 2006; Rondal, 1978).

Rollins, Snow, and Willet (1996) studied individual variations in language development of 36 children to understand how semantic and morphologic understanding corresponded with MLU. Rollins et al. studied both between-child and within-child variations of MLU scores from their longitudinal study and found that MLU was not highly correlated with semantic and morphological growth. Rollins et al. cautioned readers in using MLU as a language match.

Bornstein (2002) studied the stability of MLU in children's spontaneous language in different situations. Natural language samples were obtained as 2-year-old children played by themselves with mother close by, played directly with their mothers, and interacted with their mothers in situations their mothers considered optimal for language production. Results showed that children produced the longer MLUs in the situations when the observer was not present, and that girls had longer MLUs than boys. Bornstein cautioned that samples obtained during observation may underestimate a child's MLU.

Miller and Chapman's SALT software, originally developed in 1983, allows for an automated calculation of MLU. This automated calculation is as accurate as calculating MLU by hand. A more recent version (v8, 2004) provides for the option of also automatically calculating MLU2 without manually excluding utterances. SALT also provides a guide for how to transcribe language samples, and the definitions of utterance segmentation provided in this manual were used in the present study.

DeThorne, Johnson, and Loeb (2005) analyzed how much the expressive vocabulary and the morphosyntax level contributed to the variance in MLU. Expressive vocabulary, measured by the number of different words, and the tense accuracy composite accounted for a significant percentage of variance in MLU. DeThorne et al.

gave three possible explanations for this influence: (a) the same cognitive device controls the development of a child's vocabulary and morphosyntax, (b) a child's lexical skills directly affect MLU, and (c) the possibility of the influence of nonlinguistic variables such as pragmatics and personality.

Parker and Brorson (2005) found that MLU calculated in morphemes and MLU calculated in words were almost perfectly correlated in the 40 typically developing children of ages 3;0 to 3;10. Parker and Brorson suggested that MLU in words and MLU in morphemes are equally effective as measures of gross language development in children. Hickey (1991) found similar results, and recommended using MLU in words as it is easier, faster to obtain, and does not require additional ad hoc decisions.

Rice, Redmond, and Hoffman (2006) studied conversational samples of children with SLI and typically developing children. Their study examined the concurrent validity and temporal stability of MLU equivalency among children with SLI, MLU-matched peers, and age-equivalent controls. Concurrent validity was examined by comparing the MLU, DSS, IPSyn, and MLU in words. Rice et al. found a high correlation among the MLU, DSS, and IPSyn analyses.

Alternatives to MLU

Due to frustrations with the reliability and validity of MLU, many researchers have turned to alternate measures. Miller and Chapman (1981) cautioned the reader that MLU is "sensitive to contextual variables such as the nature of the interaction" (p. 158). This sensitivity led Klee and Fitzgerald (1985) to propose MSL, an alternative to MLU which focused on mean syntactic length but would exclude contextually sensitive utterances. MSL excludes single-morpheme utterances. MSL is found by calculating the

mean length of utterance in morphemes of 100 consecutive, intelligible utterances with two or more morphemes, beginning with utterance #76 in the language sample. The purpose in excluding single-morpheme utterances was to “eliminate a possible pragmatic influence upon mean utterance length imposed by single-morpheme responses” (p. 255). Klee and Fitzgerald predicted that MSL would provide a more accurate representation of a child’s syntactic length of utterances.

MLU and Klee and Fitzgerald’s (1985) alternate MSL were evaluated by Rondal, Ghiotto, Bredart, and Bachelet (1987) to determine a relationship with age, intra-sample reliability, and the grammatical validity of these measures. Rondal et al.’s findings contradicted those of Klee and Fitzgerald in that MLU related well to age, was reliable and was able to predict grammatical development in the range of their studied population, which included 21 typically developing children of ages 1;8 to 2;8. MSL values differed only slightly from MLU values. The study concluded that MLU and MSL were reliable and valid measures of syntactic complexity up to 3.0 and 4.0 respectively.

Recently, Johnston (2001) published an article on an alternative method of calculating MLU called MLU2. Johnston followed the recommendations of Miller and Chapman (1981) and Klee and Fitzgerald (1985) in excluding single-word yes/no responses, imitative utterances, and elliptical question responses prior to calculating MLU. Johnston proposed that this alternate method would provide a more representative sample of a child’s language capabilities by excluding contextually sensitive utterances that may underestimate or inflate the appearance of the child’s language. Johnston’s alternative method increased MLU values between 3% and 49% for all of the language samples. Johnston explains that this corresponds to two children with an MLU of 3.5

where one could receive an MLU2 of 3.8, and the other an MLU2 of 4.5. This finding shows that the nature of the interaction affects MLU calculation, and can account for some of the variability of MLU. Controlling this variable could produce increased reliability for the measure; however, further study is required to confirm this assumption.

Med-LU

Another possible measure of syntactic abilities is the Med-LU. Studies related to childhood stuttering have used Med-LU (Logan & Conture, 1995; Melnick & Conture, 2000). Eisenberg (2001) suggested that the Med-LU may be a more appropriate and representative measure than MLU because the median of the utterance lengths in a sample would be less affected by outliers than would the mean. Med-LU might thus provide a more stable representation of a child's level of syntactic complexity; however, no data have been collected regarding the reliability or the validity of Med-LU as a measurement of utterance length.

Reliability of MLU

Gavin and Giles (1996) studied the temporal reliability of four quantitative measures of syntactic language: (a) the total number of words, (b) the MLU, (c) the number of different words (NDW), and (d) the MSL. Gavin and Giles used SALT to analyze language samples of different length based on the duration of 12 or 20 minutes, or the number of complete and intelligible utterances, ranging from 25 to 175 in increments of 25 utterances. Gavin and Giles found that MLU, MSL, and NDW were adequately reliable, but the total number of different words was not. The 20-min samples had an increased temporal reliability, with MLU scores having the highest reliability coefficients. Furthermore, temporal reliability increased as the number of total utterances in the language samples increased.

Cole, Mills, and Dale (1989) examined test-retest and split-half reliability of MLU in 10 children, ages 4;4 to 6;8 with language delay. Cole et al. obtained two language samples of over 100 utterances 2 weeks apart. Cole et al. found high correlation between the two samples, although the second sample was generally slightly higher. Similarly, Cole et al. found no significant differences between odd and even numbered utterances in the samples. The authors compared first and second halves of the language samples and found that the first half was generally longer than the second, however, no statistical differences between halves were found. Cole et al. further compared differences between 50 and 100 utterance samples and found that 50 utterance samples included 70-80% of the lexical information in the 100 utterance samples. Two 50 utterance samples were recommended for analysis, as they appeared to provide a more representative picture of a child's language than a single 100 utterance sample.

Summary

Language samples can be segmented into either P-units or C-units, a decision which would affect MLU. Other utterance inclusion or exclusion rules affect MLU as well, and have prompted development of MLU2. The use of Med-LU has been suggested but not tested. The purpose of the present study is to compare the split-half reliability of MLU, MLU2, and Med-LU when language samples have been segmented into either P-units or C-units. If differences in reliability are seen between P-units and C-units, or between MLU, MLU2, and Med-LU, clinicians might be able to make more informed choices as to which segmentation and utterance quantification tools to use.

Method

Participants

Language samples had been previously collected by three graduate students for various research reasons from 30 children (3 in each 6-month interval) who ranged in age between 2;6 and 7;11. These language samples had been used in studies by Channell and Johnson (1999). The children were typically developing, lived in Provo, Utah, spoke English as their primary language, and passed a pure-tone, bilateral hearing screening at 15 dB HL. The samples consist of each child participating in naturalistic play and conversational interactions in their own home with one of three graduate students in Speech Language Pathology.

Procedure

SALT format. The language samples were formatted to comply with SALT (Miller & Chapman, 2004) transcription conventions and guidelines. Each utterance was given a speaker code, and inflectional morphemes within words were divided by using a slash. Mazes and exact repetitions were placed in parentheses. These modifications were made in order to correctly count MLU using the SALT software. One copy of each sample was divided into P-units and one copy was divided into C-units.

P-units. Original transcriptions of the 30 language samples were written as P-units. According to the SALT manual (Miller & Chapman, 2004), a P-unit represents documentation of a complete thought. Thought completion is generally characterized by a rise or fall in intonation and the presence of a pause. When conjoined and complex sentences do not contain pauses or changing intonation, then thought completion is determined by independent and dependent clauses. In these instances, P-units are separated after two conjoined independent clauses. Dependent clauses are conjoined to

their independent clauses, and are not segmented. The following is an example of one P-unit from one of the language samples:

we have to go ten minute/s, and then we get there on time.

This utterance consists of two independent clauses, conjoined by the word *and*. The two independent clauses were not separated in order to allow the speaker to complete an entire thought.

C-units. Separate copies of the original samples were then divided into C-units by separating each independent clause. C-units differ from P-units because they are not segmented by intonation or pauses, but consist of one independent clause and its modifiers. The same example given for P-units would differ in the C-unit format by separating the two independent clauses:

we have to go ten minute/s.

and then we get there on time.

Copies of the P-unit and C-unit files were then split into even and odd utterances using a utility program. Thus, each of the original language samples yielded six files to be analyzed, including: a P-unit file, P-unit odd file, P-unit even file, C-unit file, C-unit odd file, and a C-unit even file.

Data Analysis

SALT (2004) was used to calculate the MLUs of these six versions of each of the original language samples. The analysis was arranged so that all of the utterances would be included in the analysis of MLU. SALT was also used to calculate MLU2. This was done by excluding the following types of utterances: single-word yes/no responses, imitative utterances, and elliptical question responses. SALT has the option of excluding

all of these types of utterances from analysis under the *Setup* and *Analysis Set* program options. In this way, the rules established by Johnston (2001) could be followed by using options available from SALT (2004). A utility program was used to calculate the Med-LU for all six files for each original language sample.

Results

All comparisons between samples and measures were made using paired t tests at an alpha level of $p < .01$. Descriptive statistics are shown in Table 1.

Table 1

Descriptive Statistics for C-units and P-units

	N Totals		MLU		Median		N Filtered		MLU2	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
C-units	310	47.53	4.57	0.96	4.1	0.91	241	46.60	5.45	0.97
P-units	292	34.57	4.87	1.33	4.3	1.21	226	29.42	5.80	1.41

C-units vs. P-units

Samples contained more C-units ($M = 310$) than P-units ($M = 292$); this difference was statistically significant, $t(29) = 4.60$, $p < .01$. The MLU values for P-units ($M = 4.87$) were higher than the MLU values for C-units ($M = 4.57$); this difference was also statistically significant $t(29) = 3.88$, $p < .01$. Similarly, the Med-LUs of P-units ($M = 4.3$) were higher than those of C-units ($M = 4.1$), and this difference was also statistically significant, $t(29) = 2.26$, $p < .05$.

When samples were filtered for the MLU2 analysis, there were still more C-units ($M = 241$) than P-units ($M = 226$); this difference was statistically significant, $t(29) = 4.11$, $p < .01$. The MLU2 values for P-units ($M = 5.80$) were higher than MLU2 values for C-units ($M = 5.45$); this difference was also statistically significant $t(29) = 3.75$, $p < .01$.

Within files divided into C-units, MLU2 values ($M = 5.45$) were higher than MLU values ($M = 4.57$); this difference was statistically significant $t(29) = 25.16$, $p < .01$. This was also the case for files divided into P-units, where MLU2 values ($M = 5.80$) were higher than MLU values ($M = 4.87$); this difference was also statistically significant $t(29) = 19.07$, $p < .01$.

Split-half Reliability

For both C-units and P-units the split-half reliability of MLU2 was slightly higher than that of MLU, and both were higher than the Med-LU. For C-units, MLU2 was $r = .952$, MLU was $r = .939$, and Med-LU was $r = .847$. For P-units, MLU2 was $r = .946$, MLU was $r = .940$, and Med-LU was $r = .860$. Each of these correlations was statistically significant, $p < .01$.

Age

For both C-units and P-units, age was correlated with MLU, MLU2, and Med-LU measures. For C-units, age and MLU were $r = .569$, age and MLU2 were $r = .585$, and age and Med-LU were $r = .544$. For P-units, age and MLU were $r = .568$, age and MLU2 were $r = .569$, and age and Med-LU were $r = .553$. Each of these correlations was statistically significant, $p < .01$.

As age was also correlated ($r = .417$) with the number of C-units, partial correlations were used to remove the shared effects of sample length from the relationships of age and the MLU, MLU2, and Med-LU measures. This resulted in a decrease in correlation values; age and MLU were $r = .461$, age and MLU2 were $r = .407$, and age and Med-LU were $r = .467$. Each of these partial correlations was statistically significant at $p < .05$.

Age was also slightly correlated ($r = .282$) with the number of P-units. Using partial correlations to remove the shared effects of sample length from the relationships of age and the measures resulted in a slight decrease in correlation values; age and MLU were $r = .535$, age and MLU2 were $r = .426$, and age and Med-LU were $r = .534$. Each of these partial correlations was statistically significant at $p < .05$.

Discussion

The results of this study show that split-half reliability of MLU2 was equal if not greater than the split-half reliability of MLU. Med-LU was not as reliable as either MLU or MLU2. Findings also illustrated that age was correlated with MLU, MLU2 and Med-LU, and slightly correlated with the number of C-units and P-units. Also, the results indicated that the levels of reliability for C-units were generally higher than those of P-units, but the difference was minimal.

High levels of reliability indicate that a test score is stable and has meaning. More specifically, split-half reliability is important to establish so that the clinician may trust that this measure has internal consistency. Generally, experts have set .90 as the standard for achieving adequate reliability (Gavin & Giles, 1996). Higher reliability levels are preferable; however, clinicians must determine what level of reliability is required for a test to have value.

The results of this study concur with the findings of Johnston (2001) who found that MLU2 values were greater than MLU values. Johnston predicted that use of MLU2 could improve reliability by excluding those utterances that are contextually sensitive, or more likely affected by situational variables than a child's natural linguistic abilities. The current study examined the split-half reliability of MLU2 and found the reliability for MLU2 to be higher than the split-half reliability of traditional MLU calculation. Similar to Johnston's findings, the present study found that MLU2 was not more highly correlated with age than MLU. Johnston had concluded that this is due to age being an imperfect predictor of linguistic development.

Eisenberg et al. (2001) raised the question of possibly using Med-LU. Data from this study show that Med-LU is a reliable measure, however, MLU and MLU2 both showed greater split-half reliability.

Although MLU2 was generally more reliable than MLU, clinicians should be guarded in applying these results. The current study only evaluated conversational samples of children during play. Further investigation is required for language samples done in different contexts. Johnston (2001) pointed out that MLU and MLU2 have value in different contexts, and recommended using MLU2 when a large portion of the language sample involves responses to questions.

The language samples used in the current study were collected by student clinicians in the child's home. Further research should look at the split-half reliability of children's language in different settings, different contexts, and with different clinicians. Furthermore, the present study only evaluated the language of typically developing children. Future research might investigate the reliability of these measures in children with language impairment.

Results from this study indicated overall that segmentation of utterances into C-units was only slightly more reliable than P-unit segmentation. C-units may provide the clinician with important information regarding a child's average length of independent clauses with modifiers, unaffected by a child's use of several conjoined utterances without pausing. However, more research is necessary to determine the benefits of using C-units instead of P-units. Clinicians may decide to use C-unit segmentation if a particular child repeatedly uses conjunctions between utterances. Clinicians must decide whether use of these conjunctions is to form complete utterances, or only a string of

several different utterances. P-unit segmentation is valuable in capturing a more complete thought, instead of focusing solely on syntax. Decisions regarding use of C-units instead of P-units should not be based solely on split-half reliability, as differences between the two methods appear to be minimal.

Based on the findings of this study, clinicians could use either MLU or MLU2 in child language analysis as both have adequate levels of split-half reliability. Med-LU was also found to be reliable, but not as reliable as MLU and MLU2. MLU2 appears to be a reliable alternative to MLU that may provide further insight into a child's actual utterance length. As a result, MLU2 may serve as a useful method for clinicians when analyzing conversational samples of children.

References

- Arlman-Rupp, A., de Haan, D., & van de Sandt-Koenderman, M. (1976). Brown's early stages: Some evidence from Dutch. *Journal of Child Language*, 3, 267-274.
- Blake, J., Quartaro, G., & Onorati, S. (1993). Evaluating quantitative measures of grammatical complexity in spontaneous speech samples. *Journal of Child Language*, 20, 139-152.
- Bornstein, M. H., Painter, K. M., & Park, J. (2002). Naturalistic language sampling in typically developing children. *Journal of Child Language*, 29, 687-699.
- Bountress, M. G., Bountress, N. G., & Tonelson, S. W. (1988). The influence of racial experimenter effects upon mean length of utterance. *Clinical Linguistics and Phonetics*, 2, 47-53.
- Brown, R. (1973). *A first language: The early stages*. Cambridge, MA: Harvard University Press.
- Chabon, S. S., Kent-Udolf, L., & Egolf, D. B. (1982). The temporal reliability of Brown's mean length of utterance (MLU-M) measure with post-Stage V children. *Journal of Speech and Hearing Research*, 25, 124-128.
- Chan, A., McAllister, L., & Wilson, L. (1998). An investigation of the MLU-age relationship and predictors of MLU in 2- and 3-year-old Australian children. *Asia Pacific Journal of Speech, Language and Hearing*, 3, 97-108.
- Chapman, R. S. (1981). Computing mean length of utterance in morphemes. In J. F. Miller, *Assessing language production in children: Experimental procedures* (pp. 23-36). Baltimore: University Park Press.

- Conant, S. (1987). The relationship between age and MLU in young children: A second look at Klee and Fitzgerald's data. *Journal of Child Language*, *14*, 169-173.
- Condouris, K., Meyer, E., & Tager-Flusberg, H. (2003). The relationship between standardized measures of language and measures of spontaneous speech in children with autism. *American Journal of Speech-Language Pathology*, *12*, 349-358.
- Crystal, D. (1974). [Review of the book *A First Language: The Early Stages*]. *Journal of Child Language*, *1*, 289-334.
- Crystal, D., Fletcher, P., & Garman, M. (1976). *The grammatical analysis of language disability*. London: Edward Arnold.
- DeThorne, L. S., Johnson, B. W., & Loeb, J. W. (2005). A closer look at MLU: What does it really measure? *Clinical Linguistics & Phonetics*, *19*, 635-648.
- Dromi, E., & Berman, R. A. (1982). A morphemic measure of early language development: Data from modern Hebrew. *Journal of Child Language*, *9*, 403-424.
- Dunn, M., Flax, J., & Sliwinski, M. (1996). The use of spontaneous language measures as criteria for identifying children with SLI: An attempt to reconcile clinical and research incongruence. *Journal of Speech and Hearing Research*, *39*, 3, 643-654.
- 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*, 323-342.
- Gavin, W. J., & Giles, L. (1996). Sample size effects on temporal reliability of language sample measures of preschool children. *Journal of Speech and Hearing Research*, *39*, 1258-1262.

- Harris, J. (1983). What does mean length of utterance mean? Evidence from a comparative study of normal and Down's Syndrome children. *British Journal of Disorders of Communication, 18*, 153-169.
- 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*, 197-213.
- Hickey, T. (1991). Mean length of utterance and the acquisition of Irish. *Journal of Child Language, 18*, 553-569.
- Hunt, K. W. (1965). *Grammatical structures written at three grade levels*. (Research Report No. 3). Champaign, IL: National Council of Teachers of English.
- Klecan-Aker, J. S., & Hedrick, D. L. (1985). A Study of the syntactic language skills of normal school-age children. *Language, Speech, and Hearing Services in Schools, 16*, 171-186.
- Klee, T. (1992). Developmental and diagnostic characteristics of quantitative measures of children's language production. *Topics in Language Disorders, 12*, 28-41.
- Klee, T., & Fitzgerald, M. D. (1985). The relation between grammatical development and mean length of utterance in morphemes. *Journal of Child Language, 12*, 251-269.
- Klee, T., Schaffer, M., May, S., Membrino, I., & Mougey, K. (1989). A comparison of the age-MLU relation in normal and specifically language-impaired preschool children. *Journal of Speech and Hearing Disorders, 54*, 226-233.
- Klee, T., Stokes, S. F., Wong, A. M.-Y., Fletcher, P., & Gavin, W. J. (2004). Utterance length and lexical diversity in Cantonese-speaking children with and without

- specific language impairment. *Journal of Speech, Language, and Hearing Research, 47*, 1396-1410.
- Loban, W. (1976). *Language development: Kindergarten through grade twelve*. Urbana, IL: National Council of Teachers of English.
- Logan, K. G., & Conture, E. G. (1995). Length, grammatical complexity, and rate differences in stuttered and fluent conversational utterances of children who stutter. *Journal of Fluency Disorders, 20*, 35-61.
- Melnick, K. S., & Conture, E. G., (2000). Relationship of length and grammatical complexity to the systematic and nonsystematic speech errors and stuttering of children who stutter. *Journal of Fluency Disorders, 25*, 21-45.
- Miles, S., Chapman, R., & Sindberg, H. (2006). Sampling context affects MLU in the language of adolescents with down syndrome. *Journal of Speech, Language, and Hearing Research, 49*, 325-337.
- Miller, C. A., & Deevy, P. (2003). A method for examining productivity of grammatical morphology in children with and without specific language impairment. *Journal of Speech, Language, and Hearing Research, 46*, 1154-1165.
- Miller, J. F., & Chapman, R. S. (1981). The relation between age and mean length of utterance in morphemes. *Journal of Speech and Hearing Research, 24*, 154-161.
- Miller, J. F., & Chapman, R. S. (2004). *Systematic analysis of language transcripts (SALT, v8.0)* [Computer software]. Madison, WI: Language Analysis Laboratory, Waisman Center, University of Wisconsin-Madison.

- 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*, 365-376.
- Reed, V. A., MacMillan, V., & McLeod, S. (2001). Elucidating the effects of different definitions of utterance on selected syntactic measures of older children's language samples. *Asia Pacific Journal of Speech, Language and Hearing, 6*, 39-45.
- Rice, M. L., Redmond, S. M., & Hoffman, L. (2006). Mean length of utterance in children with specific language impairment and in younger control children shows concurrent validity and stable and parallel growth trajectories. *Journal of Speech, Language, and Hearing Research, 49*, 793-808.
- Rollins, P. R., Snow, C. E., & Willett, J. B. (1996). Predictors of MLU: Semantic and morphological developments. *First Language, 16*, 243-259.
- Rondal, J. A. (1978). Patterns of correlations for various language measures in mother-child interactions for normal and Down's Syndrome children. *Language and Speech, 21*, 242-252.
- Rondal, J. A., & DeFays, D. (1978). Reliability of mean length of utterance as a function of sample size in early language development. *The Journal of Genetic Psychology, 133*, 305-306.
- Rondal, J. A., Ghiotto, M., Bredart, S., & Bachelet, J. (1987). Age-relation, reliability and grammatical validity of measures of utterance length. *Journal of Child Language, 14*, 433-446.

- Rondal, J. A., Ghiotto, M., Bredart, S., & Bachelet, J. (1988). Mean length of utterance of children with Down Syndrome. *American Journal on Mental Retardation*, *93* (1), 64-66.
- Scarborough, H. S., Rescorla, L., Tager-Flusberg, H., Fowler, A. E., & Sudhalter, V. (1991). The relation of utterance length to grammatical complexity in normal and language-disordered groups. *Applied Psycholinguistics*, *12*, 23-45.
- Schachter, F. F., Shore, E., Hodapp, R., Chalfin, S., & Bundy, C. (1978). Do girls talk earlier?: Mean length of utterance in toddlers. *Developmental Psychology*, *14*, 388-392.
- Thordardottir, E. T., & Weismer, S. E. (1998). Mean length of utterance and other language sample measures in early Icelandic. *First Language*, *18*, 1-32.
- Yoder, P. J., Spruytenburg, H., Edwards, A., & Davies, B. (1995). Effect of verbal routine contexts and expansions on gains in the mean length of utterance in children with developmental delays. *Language, Speech, and Hearing Services in Schools*, *26*, 21-32.