2012-06-28

Reliability of the Mean Length of Utterance Measure in Samples of Children's Language

Katherine Marie Bigelow
Brigham Young University - Provo

Reliability of the Mean Length of Utterance Measure in Samples of Children's Language

Katherine M. Bigelow

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Ron W. Channell, Chair
Christopher Dromey
Shawn L. Nissen

Department of Communication Disorders

Brigham Young University

August 2012

ABSTRACT

Reliability of the Mean Length of Utterance Measure in Samples of Children's Language

Katherine Bigelow
Department of Communication Disorders, BYU
Master of Science

Mean length of utterance (MLU) is widely used in child language sample analysis as a way to quantify language development. The current study examines the split-half reliability of MLU and two alternative measures: MLU2 and median length of utterance (MdLU). The effects of utterance segmentation into phonological units (P-units) or communication units (C-units) on reliability were also studied. Sixty conversational child language samples were used which included ten children with language impairment. All measures were found to have high levels of split-half reliability, with MLU and MLU2 having higher levels of reliability than MdLU. There was no significant difference between MLU and MLU2. The differences in reliability when segmented into P-units or C-units were inconsistent. Overall, MLU and MLU2 are adequately reliable measures for clinical use.

Keywords: child language, MLU, MLU2

ACKNOWLEDGMENTS

I am grateful for the many people who have helped make this thesis possible. I thank Dr. Channell for his endless patience and guidance throughout this project. I am forever grateful to my parents for their continuous support and prayers. I thank my brother, Joseph, for his love and encouragement and for making me laugh despite the stress. I thank all of my family and friends for their continued love and support. I give thanks to Jeniel for helping me to remember what is most important, to stay motivated, and to keep a balanced life while in graduate school. And last but not least, I am so grateful to my amazing classmates who have been such a strength to me. I thank them for for sharing their thoughts and advice and, most of all, for their friendship.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF APPENDICES

DESCRIPTION OF STRUCTURE AND CONTENT

The body of this thesis is written as a manuscript suitable for submission to a peer-reviewed journal in speech-language pathology. An annotated bibliography is presented in Appendix A.

Introduction

Since its introduction, the quantification of language development using the average number of morphemes in a child's utterances has become widely used by both clinicians and researchers. This measure, known as the Mean Length of Utterance (MLU), is both conceptually and computationally simple. Although MLU is widely used and has been studied at length, questions remain about the use and calculation of MLU, including how to segment utterances, which utterances to include or exclude, and the effect of these issues on reliability.

MLU can be used in language sample analysis to help identify language impairment in populations where standardized testing is difficult (e.g. preschool children; Eisenberg, Fersko, & Lundgren, 2001). Age and MLU are significantly correlated, with an average increase of 1.2 morphemes per year (Miller & Chapman, 1981). This suggests that MLU may be used to quantify a child's linguistic level and help to identify those children whose language may need further evaluation.

Factors Impacting the Calculation of MLU

Three factors related to the calculation and use of MLU merit further examination. These factors include the method used to segment the utterances produced by the child, the choice as to which utterances to include as most representative of the child's abilities, and the method used to calculate the length of utterance. Roger Brown (1973) introduced the version of MLU which is commonly used nowadays in describing developmental language and its impairment. This MLU is calculated by collecting a language sample of 100 utterances and dividing the total number of morphemes by this number of utterances. In a review of Brown's (1973) book, Crystal (1974) criticized the consistency of Brown's MLU definition, as Brown gave no clear definition of an analyzable utterance nor a method of how to segment utterances in an objective and replicable way.

These three factors (utterance segmentation, utterance inclusion, and a statistical alternative) are also related to the reliability of measurement. Since it is widely known that a measure cannot be more valid than it is reliable (Salvia & Ysseldyk, 2007), an examination of the effects of these three factors on the reliability of MLU would be important for both clinicians and researchers.

Utterance segmentation. As the MLU is the average number of morphemes in a child's utterance, this measure is highly dependent upon the length of the utterances, which is determined by how the person transcribing the sample decides to segment the utterances.

Two different methods for segmenting utterances are used by child language researchers. Often language samples are segmented into phonological or prosody units (P-units). P-units are utterances that contain a complete thought (Miller & Chapman, 2004). The separation of a P-unit is based on intonation and pauses to determine where the utterance ends. Usually a falling intonation indicates the end of a statement and raised intonation indicates a question. A P-unit can theoretically contain multiple independent clauses along with any dependent clauses (Miller & Chapman, 2004), but in practice, a P-unit is limited to two independent clauses.

In contrast, communication units (C-units) are utterances that include only one independent clause and any dependent clauses attached to it (Miller & Chapman, 2004). The C-unit is an extension of the notion of T-unit (Hunt, 1965) to oral language; a T-unit consists of one main clause and its subordinate clauses. Segmenting C-units requires consideration of the grammatical structure, not the prosodic features of the utterance (Chapman, 1981; Loban, 1976).

Children often speak in strings of utterances joined by a conjunction, such as *and,* without pausing until after the conjunction, suggesting that their thought is not yet complete. Segmenting these strings of conjoined utterances into P-units may lead to an overestimation of MLU, as P-units allow multiple independent clauses to be conjoined. Conversely, C-units are not affected by the speaking habits of conjoining several clauses without pausing (Loban, 1976). A language sample divided into C-units may be more indicative of a child's language abilities and thus more reliable, but no research has been published regarding the reliability of these two different methods of segmentation.

Utterance inclusion. Johnston (2001) noted the ways in which pragmatic factors such as the use of questions by the conversational partner and the level of familiarity between the child and the partner could significantly affect the level of the syntactic complexity of a child's language productions. Building upon work by Klee (1992) and his associates, Johnston proposed an alternate method of calculating the MLU, which she called MLU2, which would be more robust against these pragmatic influences and, thus, possibly avoid skewed MLU results. In Johnston's MLU2, utterances biased by discourse are removed: (a) all elliptical question responses where implied material is omitted, (b) exact repetitions of self or adult partner, and (c) single word responses to yes/no questions. Johnston found an average increase of 18% in the MLU when these exclusions were made from the language sample and concluded that MLU2 may be a more representative measure of a child's syntactical abilities.

Perhaps the controlled removal of these pragmatic factors using MLU2 might also make the MLU2 a more reliable measure; however, no published studies have as yet investigated this issue.

Computation of utterance length. A third way to decrease variability and thus inconsistency in the quantification of the length of a child's utterances would be to compute the median rather than the mean length of these utterances. The median of a set of values is less affected by extreme scores than is the mean (Boslaugh & Watters, 2008). This calculation, herein called the MdLU, might help to avoid underestimating or overestimating a child's syntactic abilities in light of the pragmatic factors within the sample or the segmentation issues involved in transcribing the sample which have been described above. Eisenberg et al. (2001) mentioned the possible statistical advantages of MdLU. The MdLU has been used in two studies of the effects of utterance length on disfluency (Logan & Conture, 1995; Melnick & Conture, 2000) but no studies have been published regarding MdLU itself.

Reliability of MLU Measure

A pilot study. Cole, Mills, and Dale (1989) had examined the split-half reliability of MLU on samples of preschool children and found it to be $r = .94$, indicating a high level of internal consistency. Guided by this finding, Kemeny (2007) carried out pilot testing on the split-half reliability of both MLU and MLU2 using different types of utterance segmentation for 30 language samples from typically developing children. Kemeny's results suggested that the split-half reliability of MLU2 was equal to or greater than that of MLU. The difference between C-units and P-units was minimal; however, C-units had generally higher levels of reliability than P-units.

The current study. Kemeny's (2007) study was the first to examine the effects of utterance segmentation and utterance inclusion on the split-half reliability of MLU; however, expansion and generalization of that study to a larger number of child language samples as well as to samples from children with language impairment would be advantageous. The reliability of

MdLU and its relation to utterance segmentation and inclusion deserves scrutiny as well. Thus, the current study extends research comparing the split-half reliability of MLU, MLU2, and MdLU of child language samples segmented into P-units and C-units.

Method

Participants

The conversational language samples used in this study had been collected for various purposes as part of other studies.

Reno samples. Thirty language samples were collected by Fujiki, Brinton, and Sonneberg (1990) in the Reno, Nevada area. Ten of the children had language impairment (LI), 10 were matched to the children with LI for chronological age, and 10 were matched for language. The LI group ranged in age from 7;6 to 11;1. These children scored one or more standard deviations below the mean on two standardized expressive and receptive language tests. The group of children matched for language ranged from 5;6 to 8;4 and matched the language score within 6 months of the children with LI. The children matched for chronological age matched the children with LI within 4 months of age. Each language sample was elicited with toys and introducing familiar topics and lasted 30 minutes, containing anywhere between 178 and 611 utterances. The samples were mostly conversational in nature with some narrative and discourse elements.

Provo samples. Thirty language samples were collected by three graduate students for various research reasons from 30 children who lived in Provo, Utah and ranged in age from 2;6 to 7;11. The children were typically developing, spoke English as their primary language, and passed a pure-tone, bilateral hearing screening at 15 dB HL. Each sample was collected in the child's family's apartment, with generally only the child and the clinician present, and used various games, toys, and activities to facilitate conversation. These samples were used in a study by Channell and Johnson (1999); each sample contained approximately 200 child utterances.

Procedure

SALT format. Each language sample was converted to SALT format (Miller & Chapman, 2004) in order to use SALT software to calculate MLU. Each utterance was assigned a speaker code. Morphemes within words were divided with a slash and exact repetitions and mazes were placed in parenthesis. Two copies were made of each sample; one segmented into P-units, and the other into C-units.

P-units. One copy of the language sample was divided into P-units. A P-unit represents a complete thought which is highly based on prosody to determine the end of the utterance. A P-unit can include up to two independent clauses and their dependent clauses.

C-units. A second copy of the language sample was segmented into C-units. Only one independent clause and its dependent clauses make up a C-unit, which is not based on prosody but on syntax.

Each of the P-unit and C-unit files were split in half (even and odd utterances) resulting in four files to be used in analysis: P-unit even file, P-unit odd file, C-unit even file, C-unit odd file.

Reliability of segmentation. Ten percent of the samples were randomly selected and re-segmented for comparison to the original segmentation into P-units and C-units and the reliability was 97%.

Data Analysis

The MLU was then calculated with SALT software (Miller & Chapman, 2004) for the four files of each language sample. MLU2 (Johnston, 2001) was calculated using SALT by excluding elliptical responses to questions, single word yes/no responses, and exact self-repetitions or repetitions of the adult utterance. A utility program was used to calculate the MdLU. Split-half reliability was examined by correlating the values for the odd-numbered utterances with values from the even-numbered utterances in each sample.

Results

Paired *t*-tests were used to compare the measures of each sample at an alpha level of $p < .01$. Descriptive statistics are shown in Table 1, and paired *t*-test values are shown in Table 2.

Table 1

Descriptive Statistics for C-units and P-units

	N Units		MLU		MLU2		MdLU	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Overall								
C-units	336.98	84.98	5.24	1.19	6.22	1.25	4.71	1.07
P-units	289.58	76.25	5.97	1.87	6.71	1.66	5.15	1.56
Provo								
C-units	299.60	47.52	4.60	1.02	5.47	1.03	4.12	0.93
P-units	281.73	34.61	4.92	1.39	5.88	1.54	4.32	1.24
Reno								
C-units	374.37	97.68	5.88	1.00	6.96	0.99	5.30	0.88
P-units	297.43	102.49	7.02	1.69	7.53	1.33	5.98	1.40

Table 2

C-unit and P-unit Comparison: Paired t-Test Values

	N Units		MLU		MLU2		MdLU	
	t	*df*	*t*	*df*	*t*	*df*	*t*	*df*
Overall	6.98**	59	-6.51**	59	-5.64**	59	-4.83**	59
Provo	4.54**	29	-4.01**	29	-3.83*	29	-2.26*	29
Reno	7.07**	29	-6.26**	29	-5.16**	29	-4.60**	29

** $p < .01$
* $p < .05$

The samples contained more C-units ($M = 336.98$) than P-units ($M = 289.58$); this difference was statistically significant, $t(59) = 6.98$, $p < .01$. P-units were longer than c-units, as show by the MLU values. The MLU values for P-units ($M = 5.97$) were higher than the MLU values for C-units ($M = 5.24$); this difference was statistically significant $t(59) = -6.51$, $p < .01$.

The MLU2 values for P-units ($M = 6.71$) were higher than the MLU2 values for C-units ($M = 6.22$); this difference was statistically significant, $t(59) = -5.64$, $p < .01$. The MdLU values of P-units ($M = 5.15$) were also higher than the C-unit values ($M = 4.71$); this difference was also statistically significant, $t(59) = -4.83$, $p < .01$.

Split-Half Reliability

The observed levels of split-half reliability are shown in Table 3. These reliability levels were statistically significant for the MLU, MLU2, and MdLU measures at $p < .01$.

Table 3

Split-Half Reliability

	MLU	MLU2	MdLU
Overall			
C-units	.96**	.97**	.86**
P-units	.97**	.89**	.91**
Provo			
C-units	.94**	.95**	.85**
P-units	.94**	.90**	.85**
Reno			
C-units	.95**	.94**	.79**
P-units	.96**	.80**	.89**

** *p* < .01

Correlations with Age

Correlations with age are shown in Table 4. The Provo samples were correlated with age for both P-units and C-units, and these correlations were statistically significant, *p* < .01. The Reno samples, however, were not significantly correlated with age.

Table 4

Correlations with Age

	MLU	MLU2	MdLU
Provo			
C-units	.57**	.60**	.54**
P-units	.57**	.57**	.55**
Reno			
C-units	.21	.28	.22
P-units	.10	.28	.06

** *p* < .01

As the scores in the Provo samples were significantly correlated with age, the observed split-half reliability levels might be due in part to their shared association with age. Accordingly, partial correlations were used to remove the effects of age. These partial correlations are shown in Table 5. The sizes of the split-half reliability levels remained high and were statistically significant, $p < .01$, suggesting that split-half reliability was independent of the shared association of these reliability levels with age.

Table 5

Split-Half Reliability: Removing Age from Provo Correlations

	MLU	MLU2	MdLU
C-units	.91**	.93**	.80**
P-units	.92**	.87**	.80**

** $p < .01$

Perhaps the difference in correlation patterns between the Provo and Reno groups was due to the presence in the Reno samples of samples from children with language impairment. Accordingly, the scores of the typically developing children were analyzed separately from the children with language impairment. These data correlation values are presented in Table 6, where it can be seen that the patterns of the typically developing children and the children with language impairment were quite similar, and that correlation values from both subgroups continued to differ, according to the segmentation method and the length measure calculated.

Table 6

Split-Half Reliability of Reno Samples: Typical vs. Language Impairment

	MLU	MLU2	MdLU
Typical			
C-units	.97**	.93**	.75**
P-units	.93**	.73**	.83**
Language Impairment			
C-units	.80**	.85**	.67**
P-units	.93**	.71**	.73**

** *p* < .01

Discussion

High levels of reliability suggest that a measure has stability and consistency. The generally accepted interpretation for reliability levels is .90 for high reliability and .80 for adequate reliability (Gavin & Giles, 1996). With only one exception, all measures in the current study were at least adequate, and many had high reliability. Only one measure fell below the standard, which was .79 for MdLU in C-units from the Reno samples.

In general, MdLU had lower reliability overall than MLU and MLU2. The differences between MLU and MLU2 as well as the differences between C-units and P-units were minimal and inconsistent. For MLU, the difference in reliability for P-units and C-units was minimal. For MLU2, C-units had higher reliability, especially for the Reno samples. For MdLU, P-units had higher reliability, which was most apparent in the Reno samples.

Age was correlated with MLU, MLU2, and MdLU for the Provo samples, but there were no significant correlations with age for the Reno samples. This may be due to the differences in the nature of these two sets of samples. The Provo samples are developmental in nature; samples were taken from typically developing children from each age group, whereas the Reno samples were collected based on language impairment and language age matches as well as chronological age matches. However, when age correlations were removed from the Provo samples, MLU and MLU2 were highly reliable, suggesting that the split-half reliability of these measures is independent of age.

The results of this study differ somewhat from the findings of Kemeny's (2007) pilot study. Kemeny found the split-half reliability of MLU2 to be equal to or greater than that of MLU. The current study found that while both measures had high levels of reliability, MLU2 was not notably more reliable than the traditional MLU. Kemeny found that C-units had somewhat higher reliability than P-units, however, the current study found minimal and inconsistent differences. In the current study, MdLU was shown to be a clinically reliable measure, however, its reliability was not as high as MLU or MLU2; this finding agrees with the results of Kemeny's study.

The Provo language samples were used in both Kemeny's (2007) study and the current study. However, the results observed differed slightly, perhaps due to differences in the segmentation of P-units and C-units. Reliability of utterance segmentation was shown to be high in the current study; however, no reliability was analyzed for utterance segmentation in Kemeny's study. Nevertheless, the findings for the Provo samples were minimally different between the two studies. In contrast, patterns of split-half reliability associated with different segmentation or computational methods were quite different for the Reno samples, suggesting that conclusions drawn previously about the reliability of MLU vs. MLU2 and C-units vs. P-units may not apply to all types of language samples.

The findings of the present study also agree with the findings of by Gavin and Giles (1996). Based on repeated measures of the MLU, Gavin and Giles had found the test-retest reliability to be high. In the current study, the split-half reliability, which removes possible extraneous variables from the test-retest reliability, was also high for MLU. The present findings of the current study thus complement the findings of the Gavin and Giles study.

The findings of the present study also add to the insights offered by Johnston (2001), who found that MLU2 values were greater than MLU values. Johnston's (2001) study did not address the relative reliability of MLU and MLU2, but the descriptive statistics in the current study showed that MLU2 values are higher than MLU values. From Johnston's study, it might be inferred that using MLU2 would increase reliability by removing utterances affected by discourse variables, thus making the sample more representative of the child's actual language ability. However, for the Reno samples in the current study, this was not the case. For the P-unit Reno samples, the split-half reliability for MLU2 (.80) was much lower than for MLU (.96). It was noted, though, that segmentation of the Reno samples into C-units increased the reliability of MLU2 to almost equal to that of MLU.

Another study with relevance to the current study is that of Eisenberg et al. (2001), who suggested that MdLU might be a better indicator of linguistic ability but provided no supporting data. The present study showed that while MdLU is a reliable measure, both MLU and MLU2 offer higher levels of split-half reliability.

A possible explanation as to why the Reno samples did not follow the patterns found by Johnston (2001) and Kemeny (2007) might be because of the differences in the types of samples used and way the samples were collected. The Reno samples were much longer than the Provo samples. After the initial conversation between the clinician and child in the Reno samples, the language tasks became more narrative in nature and thus fewer utterances were removed for the MLU2 measure, which would minimize the differences between MLU2 and MLU. In the Provo samples, the clinician was usually meeting the child for the first time and the samples were shorter and more conversational in nature. These conversational Provo samples probably had more one-word answers and repetitions which would be deleted by MLU2, thus creating a bigger difference between MLU and MLU2.

Similarly, the pattern found in Kemeny's study was that the MLU2 was more reliable than MLU, and C-units were more reliable than P-units. These results were similar to the findings for the Provo samples in the current study, however, the Reno samples showed a different pattern. The question was raised that this difference could be due to the samples with language impairment included in the Reno group. However, when the samples from the Reno children with language impairment were analyzed separately from the samples from typically developing Reno children, the results for the children with language impairment followed a similar pattern as the typically developing children (see Table 6), suggesting that language impairment did not greatly affect the overall reliability of MLU and MLU2. The typically developing Reno children did not follow the same pattern as the typically developing Provo children. This suggests that the differences might be related to the type of sample collection, as mentioned earlier. Because the Reno samples were longer and more narrative in nature, they were probably less affected by MLU2 and thus did not show the same pattern that the Provo samples did.

As indicated by the inconsistent results of the current study, further research is needed on this topic to investigate the impact of different kinds of samples on the split-half reliability of MLU and MLU2. Future studies could analyze the reliability of MLU and MLU2 with different types of samples such as expository or narrative samples instead of conversational samples. It would also be helpful to determine if segmentation into C-units or P-units yields higher reliability for certain types of samples. A replication of the current study could also be done with new samples to determine if the results show a similar pattern. Another area needing research is how the reliability of MLU and MLU2 is affected by language impairment.

Nevertheless, findings from the current study showed that all three measures, MLU, MLU2, and MdLU, offered acceptable to high levels of split-half reliability. MdLU's reliability was not as high as that of MLU and MLU2. Reliability levels varied between different methods of utterance segmentation and also varied between different sample types, but –except in one instance– all were acceptably high. In general, clinicians can have confidence in the internal consistency of these measures when analyzing conversational language samples.

References

Boslaugh, S., & Watters, P. A. (2008). *Statistics in a nutshell: A desktop quick reference.* Sebastopol, CA: O'Reilly.

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

Channell, R. W., & Johnson, B. W. (1999). Automated grammatical tagging of child language samples. *Journal of Speech, Language, and Hearing Research, 42*, 727-734.

Chapman, R. S. (1981). Computing mean length of utterance in morphemes. In J. F. Miller (Ed.), *Assessing language production in children: Experimental procedures* (pp. 22-25). Baltimore, MD: University Park Press.

Cole, K. N., Mills, P. E., & Dale, P. S. (1989). Examination of test-retest and split-half reliability for measures derived from language samples of young handicapped children. *Language, Speech, and Hearing Services in Schools, 20,* 259-268.

Crystal, D. (1974). Review of the book *A first language*: The early stages. *Journal of Child Language, 1*, 289-334.

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. doi: 10.1044/1058-0360(2001/028)

Fujiki, M., Brinton, B., & Sonnenberg, E. A. (1990). Repair of overlapping speech in the conversations of specifically language-impaired and normally developing children. *Applied Psycholinguistics, 11*, 201-215.

Gavin, W. J., & Giles, L. (1996). Sample size effects on temporal reliability of language sample measures of preschool children. *Journal of Speech, Language, and Hearing Research, 39*(6), 1258-1262.

Hunt, K. W. (1965). *Grammatical structures written at three grade levels.* Champaign, IL: National Council of Teachers of English.

Johnston, J. R. (2001). An alternate MLU calculation: Magnitude and variability of effects. *Journal of Speech, Language, and Hearing Research, 44*, 156-164. doi: 10.1044/1092-4388(2001/014)

Kemeny, A. (2007). *Split-half reliability of MLU and MLU2 in two methods of utterance segmentation.* Unpublished Thesis, Brigham Young University, Provo, UT.

Klee, T. (1992). Developmental and diagnostic characteristics of quantitative measures of children's language production. *Topics in Language Disorders, 12,* 28-41.

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

Logan, K., & Conture, E. (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). The systematic and nonsystematic speech errors and stuttering of children who stutter. *Journal of Fluency Disorders, 25*, 21-45.

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*, 154-161.

Miller, J. F., & Chapman, R. S. (2004). Systematic analysis of language transcripts (SALT, v8.0) [Computer software and manual]. Madison, WI: Language Analysis Laboratory, Waisman Center, University of Wisconsin-Madison.

Salvia, J., & Ysseldyke, J. E. (2007). *Assessment in special and inclusive education* (10th ed.). Boston: Houghton Mifflin.

Appendix A: Annotated Bibliography

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

Brown claims that MLU is an excellent measure of grammatical development because the child's new grammatical knowledge increases utterance length. MLU is calculated by segmenting the language sample into utterances and dividing the total number of morphemes by the number of utterances. Brown specified that a language sample of 100 utterances should be used to compute the MLU. Repetitions of phrases or words should only be counted once, unless the word is repeated for emphasis. Filler words, as in *umm* or *oh*, should not be counted but *no, yeah*, and *hi* are counted as one morpheme each. Compound words that are made up of two or more free morphemes are counted as single words (e.g. *birthday, choo-choo*, and *night-night* count as one morpheme each). Diminutives (e.g. *mommy, doggie*), irregular past tense verbs (e.g. *got, went*), and catenatives (e.g. *gonna, wanna, hafta*) are counted as one morpheme. All inflections, such as possessive-s, plural-s, third person singular-s, regular past-d, and progressive-ing, are counted as separate morphemes (e.g. *her* counts as one morpheme, whereas *hers* counts as two morphemes).

A child's MLU corresponds with age. Brown describes five stages of morphological and syntactical development. A child with an MLU from 1.0-2.0 is in Stage I and is usually between 12-26 months old. Stage II includes children of about 27-30 months with an MLU from 2.0-2.5. Stage III children have an MLU of 2.5-3.0 and are usually between 31-34 months old. Stage IV includes children with and MLU of 3.0-3.75 (about 35-40 months old). Stage V constitutes children with an MLU of 3.75-4.5 (about 41-46 months old). Once a child reaches Stage V and has an MLU of 4.5 (at around 3 years 10 months old), the length of a child's utterance is more dependent upon the situation, not what the child knows. Therefore, when a child reaches Stage V, MLU is no longer a valid measure of syntactical development. Brown lists eight specific rules for calculating MLU. He admits that no claim can yet be made that these rules are complete or correct, but that the rules are a way to make the data comparable.

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, Language, and Hearing Research, 25*(1), 124-128.

Many factors affect reliability of MLU including familiarity of the examiner, familiarity with the stimulus topic, and instructions given with the stimulus. MLU-M is the mean length of utterance in morphemes. The purpose of the study was to assess the temporal reliability of MLU-M for children beyond Brown's Stage V of language development by obtaining MLU-M data for the same children on three successive days. The study resulted in a lack of stability of MLU-M values from day to day for children beyond Brown's Stage V. Several factors may cause this instability. For example, as children's language develops, they use embedding and deletion which makes their utterances more complex but not necessarily longer.

Chapman, R. S. (1981). Computing mean length of utterance in morphemes. In J. F. Miller (Ed.), *Assessing language production in children: Experimental procedures* (pp. 22-25). Baltimore, MD: University Park Press.

Chapman’s method for computing MLU is based on Brown’s (1973) method, only 50 utterances are used instead of 100 and the first page of the transcript is included. A child’s utterances are segmented using rising and falling intonation to indicate where one utterance ends and a new one begins. Unintelligible utterances are not included in the morpheme count, but exact repetitions of utterances are included. A morpheme is defined as the smallest meaningful unit of a language: a noun such as *book* is one morpheme, or a plural-*s* is one morpheme. Chapman lists the same rules as Brown (1973) for counting morphemes in order to calculate MLU.

The total number of counted morphemes is divided by the number of utterances. Chapman indicates that the language sample used must be representative. There are many things that can influence the MLU. Familiarity (of the listener, the setting, or the topic) greatly influences the length of a child’s utterances. Other factors to consider are imitation of the speaking partner, self-repetitions, a lot of questions and answers, routine speech (like counting or nursery rhymes), and utterances with clauses conjoined by the conjunction *and*. If any of these happen at a high rate, the language sample may not be a good representation of the child’s actual language abilities.

Chapman explains that T-units or communication units (Loban, 1976) may be the most representative form of segmenting utterances in a language sample. T-units separate clauses conjoined by *and* as well as ignore false starts.

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(3), 349-358. doi: 10.1044/1058-0360

Language measures are used to assess language development and skills. There are two kinds of language measures: standardized psychometric tests and measures of spontaneous speech from natural language samples. Assessing language in children with autism presents challenges. For example, it may be difficult to elicit a naturalistic language sample due to the pragmatic and conversational language deficits associated with autism. No research has been done to compare the use of standardized test and spontaneous speech measures in children with autism. The purpose of this study was to look at the relationship between standardized and spontaneous speech measures of language in children with autism. Results indicated that both kinds of assessment measured the same linguistic abilities in these children. The children with autism performed lower on both standardized and spontaneous speech measures. The result of the study indicated that MLU is a good measure of the grammatical ability for children with autism.

Crystal, D. (1974). Review of the book *A first Language*: The early stages. *Journal of Child Language, 1*, 289-334.

This article reviews Brown's book about syntactical language development. The stages of language development developed by Brown are based on MLU. Stage I begins with an MLU of 1.0. Stage II includes an MLU from 2.0 to 2.5. Brown states that, up until Stage V, MLU is an excellent indicator of grammatical development because most new language knowledge increases the length of utterances in morphemes. By Stage V, MLU is no longer a good indicator of syntactical development because the child can produce more complex sentences in ways that do not necessarily increase the MLU. MLU is a useful comparative tool. Crystal's main complaint with MLU is the inconsistency in using it and having to make subjective analytical decisions. Crystal pointed out the flaws of Brown's rules for calculating MLU. Brown never gave a definition of an utterance nor how to segment the utterances of a language sample in a consistent manner.

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. doi 10.1080/02699200410001716165

The MLU is commonly used by clinicians for diagnosing language impairment and measuring progress, but what and how MLU reflects linguistic knowledge is not clear. MLU was originally used to measure grammatical development, but studies since then have shown that utterance length and grammatical complexity are not perfectly correlated. A strong correlation has been found between MLU and the number of different words; therefore, a larger vocabulary leads to longer utterances. This study looked at the relation between MLU and morphosyntactic and syntactic abilities. Language samples from 44 typically developing children were analyzed and four measures used: MLU, number of different words, total number of words, and a tense accuracy composite, which is a measure of morphsyntax. Results showed that the number of different words had the greatest effect on the variance of MLU with morphosyntax also affecting MLU. MLU appears to be greatly affected by children’s expressive vocabulary. The study concluded that MLU measures expressive vocabulary skills more than morphosyntactic ability. Also, MLU is somewhat affected by non-linguistic factors like talkativeness, assertiveness, pragmatics, etc. Multiple forms of assessment should be used in addition to MLU. Interpretation should be made with the context in mind.

Dunn, M., Flax, J., Sliwinski, M., & Aram, D. (1996). The use of spontaneous language measures as criteria for identifying children with specific language impairment: An attempt to reconcile clinical and research incongruence. *Journal of Speech, Language, and Hearing Research*, *39*(3), 643-654.

There is no generally accepted method for identifying children with specific language impairment (SLI). This study discussed the differences in standardized test performance, MLU, and errors in spontaneous language between groups of preschool children diagnosed as SLI but who were not identified through standardized psychometric discrepancy criteria. The validity of quantitative measures of MLU, syntax, and pragmatics as criteria for differentiating between normally developing children and children with SLI was also examined. Results suggested that a combination of MLU, percent of structural errors, and chronological age were useful variables in predicting SLI.

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. doi: 10.1044/1058-0360(2001/028)

Because of the limitations of standardized testing in differentiating children with normal language and children with language impairment, language sample analysis (LSA) is being used more with preschool children. Mean length of utterance (MLU) is the most common LSA procedure used by clinicians. Longer utterances are not always more grammatically complex than shorter utterances. MLU therefore, is not a measure of syntactic development, but a measure of average utterance length. MLU will identify some, but not all preschool children with language impairment. Even though a child might have an MLU in the normal range, he or she may still have a language impairment. A low MLU is supporting evidence of a diagnosis of language impairment but should not be the only form of assessment in confirming this diagnosis.

Gavin, W. J., & Giles, L. (1996). Sample size effects on temporal reliability of language sample measures of preschool children. *Journal of Speech, Language, and Hearing Research, 39*(6), 1258-1262.

The purpose of the study was to evaluate the temporal reliability of four language sample measures: total number of words, number of different words, mean length of utterance in morphemes (MLU-m), and mean syntactic length. The temporal reliability was found to be dependent upon the language sample size. The test-retest reliability was higher for longer samples that had around 175 intelligible utterances.

Hunt, K. W. (1965). *Grammatical structures written at three grade levels.* Champaign, IL: National Council of Teachers of English.

Hunt's research is based on written language. Before proper analysis of a language sample, extraneous utterances and words (e.g. mazes) must be removed. Since children use *and* a lot, sentence (or utterance) length varies greatly. Determining the length sentences for analysis is very subjective, making assessment measure results variable. Hunt presents the "minimal terminable unit," or T-unit for short, as a more representative way to segment utterances into the shortest grammatical sentences. A T-unit consists of one main clause and its subordinate clauses. Segmenting a language sample into t-units may be a better indicator of language development.

Johnston, J. R. (2001). An alternate MLU calculation: Magnitude and variability of effects. *Journal of Speech, Language, and Hearing Research, 44*(1), 156-164. doi: 10.1044/1092-4388(2001/014)

The study looked at how MLU was affected by minimizing discourse-based variability and by removing some certain question responses and repetitions. Answering questions and predictability of conversations with familiar people are correlated with a reduced MLU. An alternate MLU was created (MLU2) to overcome these limitations of the traditional MLU. To calculate MLU2, discourse biased utterances were removed including: exact self repetitions, exact repetitions of the adult partner, single-word yes/no responses, and *Wh* questions that elicited only a minimal answer were all removed from the sample. Doing so increased MLU by 3%-49% with an average increase of 18%. This clearly shows that discourse variables can have a major influence on MLU.

Kemeny, A. (2007). *Split-half reliability of MLU and MLU2 in two methods of utterance segmentation.* Unpublished Thesis, Brigham Young University, Provo, UT.

This pilot study examined the split-half reliability MLU, MLU2, and MdLU. This study used 30 conversational language samples which had been previously used in studies by Channell and Johnson (1999) and were of typically developing children living in Provo, Utah. The samples were already segmented into P-units and an additional copy of each sample was segmented into C-units. The SALT software (Miller & Chapman, 2004) was used to calculate MLU and MLU2.

Results showed that MLU2 had a higher split-half reliability than the traditional MLU and MdLU had the lowest reliability overall. Both MLU2 and MLU have adequate reliability levels and can be used clinically. C-units had a slightly higher reliability than P-units. Future research should include different types of samples and examine the effects of language impairment on the MLU measures.

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.

The relationship between MLU and age was examined in 48 preschool children, half of which had SLI and half had typical language. Samples consisted of a conversation between the mother and child during free play. MLU and age were significantly correlated for both the SLI and the typically developing groups. MLU was significantly lower in the SLI group than typically developing children of the same age. Results also showed that the rate of MLU change was similar in each group, which means

that although MLU for the SLI group is lower, it increased almost as much for the typically developing children over the two year age range.

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.

MLU and lexical diversity were studied in a group of Cantonese-speaking children. Results confirmed that MLU is correlated linearly with age. This correlation was not as high as the correlation for a group of children speaking American English, possibly due to grammatical differences between the languages. MLU and lexical diversity were not correlated in partial correlations removing the effects of age. This study also compared children with SLI to other age-matched and language-matched children using MLU and lexical diversity. MLU and lexical diversity were significantly higher in typically developing children than children with SLI if the same age. Results indicate that a combination of age, MLU, and lexical diversity could be used clinically to identify SLIU in Cantonese-speaking children. The children with SLI matched for language had similar MLU and lexical diversity values. MLU and lexical diversity values can be used clinically to differentiate between children with and without language impairment.

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

The communication unit (C-unit) is the smallest group of words that still maintains the same meaning. Hunt (1965) used that same unit in his research on written language called it a T-unit. A C-unit is made up of an independent clause and any of its modifying clauses. Mazes are words or phrases that are not necessary to convey the meaning of the communication unit and should be marked with brackets prior to analysis.

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*(2), 325-337. doi: 10.1044/1092-4388(2006/026)

MLU differences have been shown to correlate with syntax use. Adolescents with Down syndrome (DS) were found to have a lower MLU in interviews than typically developing (TD) children. The adolescents with DS did not, however, have lower MLU than TD children in narrative samples elicited though wordless picture books, implying that picture supports increase MLU for the adolescents with DS. The study concluded that picture support did increase the MLUs for the DS group.

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.

A significantly correlated relationship between age and mean length of utterance was found in this study of 123 middle- to upper-class children ages 17 to 59 months. Language sample consisted of the children conversing with their mothers in free play. MLU was found to increase with age at an average of 1.2 morphemes per year. Results of the study can help with finding children whose language development requires further

evaluation, quickly finding children at a particular linguistic stage, and predicting the age most likely to be associated with a certain MLU. Age-appropriateness can be determined for a child's MLU. One caution is that MLU is only a general indicator of language structural development and should not be used as a sole basis for clinical decisions. Other cautions are that MLU is highly sensitive to contextual variations in how the language sample was elicited. Also, this study only included middle- to upper-class children, and therefore data can only be applied to similar populations.

Miller, J. F., & Chapman, R. S. (2004). Systematic analysis of language transcripts (SALT, v8.0) [Computer software and manual]. Madison, WI: Language Analysis Laboratory, Waisman Center, University of Wisconsin-Madison.

The manual explains how to format language transcriptions prior to analysis by the SALT software. Each utterance is given a speaker code. Mazes are put in parenthesis. Slashes are used to separate morphemes. Utterances can be segmented into phonological units (P-units) or communication unit (C-units).

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(4), 793-808. doi: 10.1044/1092-4388(2006/056)

This study examined the concurrent validity and temporal stability of MLU in children with specific language impairment (SLI) and typically developing children. The language samples were conversational in nature and included 39 children with SLI and

40 typically developing children. Correspondence between MLU and developmental sentence scoring (DSS), index of productive syntax (IPSyn), and MLU in words was used to determine the concurrent validity. There was high correlation between MLU, DSS, and IPSyn. There was no significant difference between the SLI group and the typically developing group. Results showed that MLU is a reliable and valid measure of language development for longitudinal growth patterns from 3 to 8 years of age.

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. doi: 10.1044/1092-4388(2009/08-0183)

MLU is a robust measure of language acquisition in children and is used clinically in diagnosing children with language impairment. An MLU of one standard deviation below the mean for the child’s age group is considered a language impairment. This study provided an age progression of MLU for children with and without SLI. The conversational language samples were collected at 6-month intervals. The children with SLI showed lower MLU levels. Results indicate that MLU is a reliable and valid measure of language acquisition and language impairment. The data in the study can be used clinically to estimate a child’s performance compared to normative data.