2009-06-02

A Comparison of Seven Automated Measures of Syntactic Complexity

Laura Elizabeth Wilde
Brigham Young University - Provo

A COMPARISON OF SEVEN AUTOMATED MEASURES OF SYNTACTIC COMPLEXITY

by

Laura E. Wilde

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

August 2009

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a thesis submitted by

Laura E. Wilde

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 L. Nissen

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the thesis of Laura E. Wilde 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

Date	Ron W. Channell Graduate Coordinator

Accepted for the College

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

ABSTRACT

A COMPARISON OF SEVEN AUTOMATED MEASURES OF SYNTACTIC COMPLEXITY

Laura E. Wilde

Department of Communication Disorders

Master of Science

This study compared seven syntactic measures which can be automatically generated by the Computerized Profiling (CP) software: Mean Length of Utterance in morphemes or words (MLUm or MLUw), Mean Syntactic Length (MSL), the Index of Productive Syntax (IPSyn), the Picture Elicited Scoring Procedure (PESP) for the Language Analysis Remediation and Screening Profile (LARSP), the Syntactic Complexity Score (MSC) scoring of LARSP, and Developmental Sentence Scoring (DSS). Language samples came from 192 children, 106 typically developing children, ages 5;6 to 11;2 and 86 children with language impairment, ages 5;6 to 11;1. Patterns of correlation were consistent for children with or without language impairment. All measures were computed with CP software, and all coding decisions that were made by the software were accepted.

The three measures of length (MLUm, MLUw, and MSL) were highly intercorrelated. MSC correlated with the measures of length and with DSS. DSS correlated with the length measures, though not as highly as MSC. DSS also correlated with IPSyn. IPSyn correlated moderately with PESP, correlated less with MSC, and correlated the least with the measures of length. PESP correlated moderately with each measure. PESP, DSS, and IPSyn correlated more highly for the children with language impairment. These measures correlated highly sometimes and sometimes they did not correlate much. This suggests that they are measuring different aspects of syntactic ability.

ACKNOWLEDGMENTS

I would like to express appreciation to Dr. Ron Channell for his patience, his expertise, and his willingness to assist me with all problems I encountered. I would also like to thank my husband Brian Wilde. He was supportive, patient, and encouraging during the completion of this study.

Table of Contents

Page

List of Tables

Table Page

Introduction

Syntactic development is characterized by increases in the length and the complexity of utterances. Samples of a child's language are collected, transcribed, and analyzed for evidence of these increases in length and for the use of syntactic forms which emerge gradually during childhood. Often these findings are summarized quantitatively to allow a clinician to track a child's syntactic development over time. Quantitative analyses of syntactic complexity can be useful for making clinical decisions and identifying areas of concern in a child's language sample.

Though these quantitative measures are of value, the collection, transcription, and analysis of children's language samples consume a significant amount of time (Long 2001) or may involve the use of skills which a clinician may not have (Long, 1996). Automation of these measures of syntactic development has been suggested as one possible solution to these time and training demands.

Currently, available automated measures of syntactic length or complexity include variations of the mean length of utterance (MLU) as described by Brown (1973), the Index of Productive Syntax (IPSyn; Scarborough, 1990), Developmental Sentence Scoring (DSS; Lee, 1974), and two quantifications of the Language Assessment Remediation and Screening Profile (LARSP; Crystal, Fletcher, & Garman, 1989), which are Ward and Fisher's (1990) picture-elicited screening procedure (PESP) and Blake and Quartaro's (1990) measure of syntactic complexity (MSC).

Several studies have also compared manually calculated versions to the automated versions of these measures. Long and Channell (2001) compared the MLU, DSS, IPSyn, and LARSP analyses calculated by the Computerized Profiling (CP; Long, Fey, & Channell, 2000) software to manually calculations of these measures. Agreement

between automated and manual scoring ranged from very high (for MLU) to very low (for the subclause level of LARSP). Channell (2003) compared the manual and CP calculation of DSS using samples from school-aged children with and without language impairment and found agreement to be moderate. Sagae, Lavie, and MacWhinney (2005) found that the level of point-to-point reliability between automated and manual versions of IPSyn calculated by their software was very close to the levels of inter-rater reliability among human scorers.

No study has yet compared these automated measures of syntactic complexity; however, several studies have compared the manual versions of these measures. For example, Cheung and Kemper (1992) compared eleven different measures of syntactic complexity and how well these measures could index age-group differences in 30 adult speakers. These measures were: MLU in words, MCU, DSS, IPSyn, DLevel, Directional Complexity (DComplexity), two variants of Yngve depth, and two variants of Frazier's node count. All measures were completed by manual analysis. Cheung and Kemper found that all of the measures except for IPSyn accurately accounted for age-group differences in adult speakers. In another study, Kemper, Rice, and Chen (1995) compared different measures of syntactic complexity in children ages 5-10. These measures included MLU, mean clauses per utterance (MCU), DSS, Developmental Level (DLevel), IPSyn, and propositional density. Measures were completed manually by experienced clinicians and researchers. In this study, MCU, DSS, and DLevel were highly correlated to each other and showed significant development in syntactic abilities up to the age of 7. After age 7, development appeared to level off. Kemper et al. suggested that the correlation of these three measures supports the idea that they all measure the same

underlying developmental function. MLU, IPSyn, and propositional density did not show syntactic development for the children ages 5-10 that were studied.

Thus, studies have compared the manually calculated versions of these syntactic measures and have provided evidence for the usefulness of these measures in analyzing language samples. Although automated versions of these measures exist, no studies have been performed to evaluate the concurrent validity of these measures. The automated measures can be computed by one or more software packages; differences between automated and manual versions may well affect the concurrent validity of these measures. However, the results produced by the automated versions of these measures have not been directly compared to one another. The present study seeks to make comparisons among the seven automated measures of syntactic complexity which are available: MLU in words (MLUw), MLU in morphemes (MLUm), MSL, IPSyn, DSS, and the PESP and MSC scorings of automated LARSP analysis. Comparing these automated measures of syntactic complexity will provide the first evidence of concurrent validity for these measures.

Review of Literature

This review will cover five measures of syntactic complexity, including two variations of MLU, as well as software which can calculate these measures.

MLU

Development of and support for MLU. MLU has been used as a measure of syntactic length for many years. In Brown's (1973) book *A First Language: The Early Stages,* MLU was described as an index of grammatical development. Though length of utterance had been studied earlier, Brown standardized the practice of computing MLU in

terms of morphemes rather than words. Brown found that counting individual morphemes was a better measure of syntactic complexity than counting words.

MLU measures a child's average number of morphemes per utterance. Brown (1973) developed stages of development that he associates with certain MLU scores. These stages of development identify certain syntactic constructions that a child should be capable of producing. MLU is still widely used by clinicians today (DeThorne, Petrill, Hayiou-Thomas, & Plomin, 2005). It can be calculated quickly and easily and can be interpreted easily by most clinicians as well. MLU correlates significantly with age and can be used to quickly identify children who are in need of further linguistic testing (Miller & Chapman, 1981). MLU is also associated with several different areas of linguistic competence in typically developing children.

Rice, Redmond, and Hoffman (2006) recently studied MLU as an indicator of language development in children who were typically developing as well as in children with language disorders. This study found that MLU was a good indicator of grammatical language development over time and that there was no difference in the correlation of MLU with IPSyn in the group of children with language disorders. Thus, for young children, MLU had useful diagnostic value. MLU is still being used as a measure of syntactic complexity by many clinicians (Kemp & Klee, 1997).

MLU can be calculated by three different software packages: Child Language Analysis (CLAN; MacWhinney, 1991), Systematic Analysis of Language Transcripts (SALT; Miller & Chapman, 2000), and CP (Long et al., 2000).

Limitations of MLU. There are several shortcomings of MLU, and it should not be the sole measure upon which clinicians base their analysis (DeThorne et al., 2005; Miller

& Chapman, 1981). MLU can be greatly affected by linguistic context, language sample elicitation method, and other nonlinguistic factors. Klee and Fitzgerald (1985) found that MLU's correlation with grammatical competence decreased with age; beyond Brown's Stage II, MLU was found to have limited value. Scarborough, Rescorla, Tager-Flusberg, Fowler, and Sudhalter (1991) found that MLU had a strong correlation with age when the MLU was below 3.0 and that this correlation decreased significantly for MLUs between 3.0 and 4.5. Also, the correlation of MLU with IPSyn in groups of children with language disorders was found to be different than for the groups of children who were typically developing, in that MLU overestimated IPSyn measures in the group of children with disorders.

Rescorla, Dahlsgaard, and Roberts (2000) studied MLU and IPSyn outcomes for children who were identified as late-talkers and age matched children who were typically developing at ages 3;0 and 4;0. In the late-talking group, MLU correlated significantly with IPSyn at both ages, but in the typically developing group, MLU correlated with IPSyn only at age 3;0, thus confirming that for children who are typically developing, MLU is a weaker measurement as a child gets older. One last important note about MLU is that MLU becomes more reliable as the time of the language sample increases (Gavin & Giles, 1996). Often clinicians use MLU to measure small language samples (less than 50 utterances), but this is not a valid use of this index.

Variants of MLU. Two variants of MLU have also been developed that seek to limit the type of child utterances that are included in the calculation of MLU. First, Klee and Fitzgerald (1985) developed Mean Syntactic Length (MSL) which is calculated by excluding answers to yes/no questions. Klee argued that using MSL rather than MLU

helped to remove the possible influence of pragmatics on the MLU measure. Klee and Fitzgerald found that in a group of children who were typically developing, MSL had a higher correlation with age than MLU. Second, Johnston (2001) developed MLU-2, an even more exclusive variant of MLU. MLU-2 is calculated by analyzing a child's language sample after removing elliptical question responses, single word yes/no responses, and imitative responses. Johnston found MLU-2 to be a more valid measure of a child's language level than MLU.

IPSyn

Development of and support for IPSyn. Scarborough (1990) developed IPSyn as a measure of the emergence of various linguistic constructions in child language samples. IPSyn was developed in response to research that looked at large groups of children but had insufficient means of quantifying grammatical competence. Scarborough developed this measure as a quick way to quantify large amounts of data for research purposes. IPSyn codes 100-utterance speech samples for 56 different syntactic and morphological forms. Scores are computed by scoring for overall syntactic proficiency as well as scoring in four different areas of grammatical development: noun phrase, verb phrase, questions/negations, and sentence structure (Scarborough, 1990). IPSyn looks at the correct usage of grammatical constructs rather than inappropriate usage. Also, IPSyn is designed to comment on the emergence of grammatical constructs and not the mastery of them. IPSyn endeavors to evaluate individual differences in syntactic development of children.

Previous to the development of IPSyn, indexes of syntactic development that were commonly used were MLU, DSS, LARSP, and Miller's Assigning Structural Stage

(MASS; 1981). Scarborough (1990) noted that each of these measures was useful for certain clinical and research purposes, but each had drawbacks as well. LARSP is an exhaustive analysis of linguistic complexity, but it cannot give a quantitative value. LARSP and MASS are both very time consuming analyses. MLU does not provide specific information about various grammatical constructs. MASS requires a sample with 50 utterances containing both a subject and a verb. Scarborough developed IPSyn in response to these limitations.

Since its development, IPSyn has been used in research as a measure of linguistic complexity in several research articles. Scarborough found in her longitudinal study that for 15 children from 2;0 to 4;0, MLU and IPSyn correlated at 0.94. Correlation was much higher at younger ages and decreased as age increase. MLU changed little from 3;0 to 4;0, but IPSyn showed a great amount of grammatical development within this period. This may indicate that IPSyn is a more sensitive measure than MLU past 3;0. A later study done by Scarborough et al. (1991) confirmed this relationship between MLU and IPSyn. MLU and IPSyn correlated at 0.92, and the correlation was much higher for the children under 3;0. IPSyn scores were found to increase significantly in older age groups (36-42 months and 42-48 months), providing further evidence that IPSyn may be a more differentiating measure in older age groups than MLU.

Rescorla et al. (2000) supported IPSyn as a better measure for children over 3;0. This study looked at toddlers who were late-talkers and toddlers who were typically developing. Language samples were analyzed using both MLU and IPSyn. Of the 34 toddlers who were late-talkers, at age 3;0, 59% scored below the tenth percentile for MLU and 66% scored below the tenth percentile for IPSyn. At age 4;0, only 29% of the

toddlers who were late-talkers scored below the tenth percentile for MLU, but 71% scored below the tenth percentile for IPSyn. This data may support the idea that IPSyn is a more sensitive measure for children over 3;0 and above 3.0 MLU.

In a comparison of IPSyn and DSS, Holdgrafer (1995) scored language samples of 29 different children ages 3;7 to 5;0, ten of whom were considered to have language delays. Scores between IPSyn and DSS were strongly correlated, but only the IPSyn scores were able to differentiate between the children who were typically developing and the children with disorders. This study suggested that IPSyn may be a more sensitive measure when evaluating children with disorders. IPSyn can be calculated by the CP software, but it cannot be computed by CLAN or SALT software packages.

Limitations of IPSyn. IPSyn is an efficient quantitative measure of syntactic complexity and is fairly well supported by research. However, there are some shortcomings of this measure. Though IPSyn may be a useful diagnostic tool, it does not provide detailed information about specific syntactic and morphological rules (Scarborough, 1990). IPSyn can give a general level of a child's syntax, but it cannot give detailed information about the areas in which a child is behind. Also, because IPSyn was originally developed for a research project, the population on which it was tested is not large or diverse enough to develop standardized scores. Scarborough recommended that clinicians determine their own standards based on data from local children. This may be too time-consuming for most clinicians, and therefore IPSyn may not be a practical measure for working clinicians. Scarborough noted in her 1990 article that IPSyn "may prove to be helpful as a screening tool" (p. 13).

Sagae et al. (2005) wrote software that was used to compute several different measures of syntactic complexity. They called this software Grammatical Relations (GR). Sagae et al. used GR as well as CP to compute IPSyn automatically. They compared the automated scores to scores computed manually by trained child language analysts. Their study found that GR scores had higher reliability with the manual versions than the CP computed scores. Also, the point-to-point reliability between automated and manual versions was very close to the inter-rater reliability among human scorers.

DSS

Development of and support for DSS. Lee (1974) developed DSS in an attempt to predict what grammatical constructs a child would use in typical conversation. DSS measures spoken syntax. Previous to the development of DSS, syntactic language tests such as the Northwestern Syntax Screening Test or the Illinois Test of Pyscholinguistic Abilities were used to quantify a child's syntactic abilities. However, these tests did not account for individual differences in children's acquisition of language. Lee argued that a child's syntax develops as they learn to use more syntactic and morphological rules or as they start to use more syntactic elements in a single sentence. By looking at the different syntactic and morphological rules, a grammatical load can be determined for each sentence a child produces. As a child's syntactic abilities increase, the grammatical load of the sentences they produce will increase as well.

DSS is computed from a language sample of at least 50 utterances. Each utterance must be a complete sentence with a subject and a verb, though the utterances do not have to be grammatically correct. DSS looks at the following grammatical constructs: indefinite pronouns or noun modifiers, personal pronouns, main verbs, secondary verbs, negatives, conjunctions, interrogative reversals, and wh-questions. These categories (if

present in an utterance) are given scores from 1-8 based on Lee's determined level of difficulty for the grammatical production.

DSS is beneficial for diagnostic purposes as well as for showing development because it is a quantitative measure (Hughes, Fey, & Long, 1992). Studies by Liles and Watt (1984) and Johnston and Kamhi (1984) both found that the DSS scores for children with language impairment were significantly lower than DSS scores for children who were typically developing. These studies also suggested that children with language impairment may use constructs that are similar in complexity to their age-matched counterparts, but children with language impairment may make more grammatical mistakes than their age-matched peers. In a study by Hughes et al. with 31 children, only one child who was language impaired based on observation and nonstandardized language-sample analyses was judged to be typically developing by the DSS measure.

Because language samples are organized into different areas of syntax, a DSS score can help clinicians identify specific areas in which a child may be below the norm. By marking different grammatical areas that a child is producing, clinicians can see what constructions a child is consistently producing incorrectly (Hughes et al., 1992). These areas can help a clinician determine which areas should be targeted during therapy.

DSS can also be used to track progress of a child throughout therapy and thus determine the effectiveness of the therapy. Fey, Cleave, Long, and Hughes (1993) identified 21 children as scoring below the tenth percentile using DSS scoring. Eleven of these children received treatment and ten were grouped into a delayed-treatment group (receiving no treatment until after the specified period for retesting). DSS scores that were then obtained post-treatment for the group that received treatment showed

significant gains in DSS scores. The delayed-treatment group showed no significant gains in DSS scores. This study suggested that DSS scores can be used to provide evidence for treatment efficacy.

Another benefit of DSS as a measure of syntactic complexity is the availability of computer-assisted instruction (CAI). Hughes, Fey, Kertoy, and Nelson (1994) studied a program of computer-assisted instruction for DSS scoring. In this study, graduate students were split into two groups, one that received CAI and one that studied DSS on an individual basis. Students were tested before they received instruction or studied individually, and then they were tested again afterwards. Results found that CAI and individual study were comparable in terms of test score improvements of the graduate students. CAI is an effective tool to help clinicians learn how to use DSS and properly score utterances.

Lively (1984) also published an article that outlined common scoring errors made by clinicians using DSS and how these errors should be remedied. This article serves as extra help and clarification about some issues that Lee did not explicitly define. DSS can be calculated by the CLAN and CP software, but DSS cannot be computed by SALT software.

Limitations of DSS. DSS has received criticism for several reasons. First, DSS is a complex scoring system and it takes a lot of time (Fristoe, 1979). This may make it impractical as a clinical tool for language analysis. Also, some of the scoring values for certain structures have been criticized, specifically scoring of the word *like* when used as a preposition or as a conjunction. This word is scored the same even though grammatically it functions differently. Most of Lee's (1974) other rules follow

grammatically-based rules, so this discrepancy has fostered skepticism (Hughes et al., 1992).

Johnson & Tomblin (1975) studied the reliability of DSS when using a 50-utterance language sample and determined that reliability was 0.75 and that reliability for most of the individual grammatical categories was below this (with the exception of personal pronouns and main verbs). Johnson and Tomblin found that another measure of syntactic complexity, mean length of response (MLR), had a reliability of 0.85 for a 50 utterance sample. For DSS to have that same 0.85 reliability, a 95 utterance sample was required. However, Johnson and Tomblin noted that though MLR requires a smaller sample size, this should not be construed as evidence that MLR is the better measure.

Another limitation of DSS is that two children with very different grammatical abilities may receive the same score (Hughes et al., 1992). Also, one child may be tested and then retested a year later and receive the same score, even though their specific syntactic abilities may be very different on these two occasions. DSS score alone cannot give a complete evaluation of a child's linguistic abilities.

PESP

Development of and support for PESP. Ward and Fisher (1990) developed a fourth quantitative measure of syntactic development called the Picture-Elicited Screening Procedure (PESP). This measure was first developed as an attempt to update and improve an older test called the Renfrew Action Picture Test. The pictures in this test were designed to elicit complex syntactic structures from a child without clinicians asking questions.

PESP is based on LARSP categories, but it attempts to produce a quantitative measure and to allow a measure to be obtained quicker than with LARSP. The procedure

for obtaining a PESP score, was outlined by Ward and Fisher (1990). The first step of the directions indicates that the clinician should "LARSP each utterance". This means that each utterance should be parsed, analyzing each clausal element and all embedded structures as well. The directions for obtaining PESP are as follows:

1. LARSP each utterance.
2. Mark the structure on the LARSP profile sheet.
3. Count the number of marked structures at each Stage. Disregard how often a structure has been logged—once is enough.
4. Multiply the number of marked structures at each Stage by the Stage number.
5. Total the scores for each Stage.
6. The total of all the Stages is the PESP score.

Ward and Fisher (1990) noted that they expected PESP to be used normally as an informal measure, using the cards to elicit language and each clinician developing their own intuitive norms. However, Ward and Fisher did approve the use of the PESP scoring system in cases where quantitative measures must be used for screening purposes or to indicate syntactic development over time. PESP can be calculated by CP software, but the CLAN and SALT software cannot compute PESP.

Limitations of PESP. The main limitation of PESP is that it is not supported by research. Ward and Fisher's (1990) article is the only available article that studied PESP as a measure of syntactic complexity. Ward and Fisher also explicitly stated that PESP is an informal measure and that further testing is necessary for quantifying data of a child's syntactic abilities in a clinical setting. Also, because Ward and Fisher encouraged each

clinician to develop their own intuitive norms, data cannot be standardized or compared for research purposes.

MSC

Development of and support for MSC. Blake and Quartaro (1990) developed another measure of syntactic complexity which, like PESP, attempts to quantify LARSP results. This measure is based on the number of grammatical categories combined in an utterance; these categories are subject, verb, object, and complement (Blake, Quartaro, & Onorati, 1993). This measure, like MSL and MLU-2, looks only at multiword utterances because single word utterances do not make use of syntactic rules. Each LARSP clause unit, including subject, verb, object, etc., is counted and given one point. The subject and object categories can be made of a noun phrase, a noun, or a pronoun. The verb category includes the main verb and any auxiliary verbs, particles, and infinitives that have the same subject as the main verb. The complement category can be made of a prepositional phrase, a predicate adjective, a predicate noun or pronoun, or an adverb (p. 143). The subject, verb, and object categories are all counted as one unit, and each separate compliment is counted. Clause units are totaled and used in calculating complexity.

Blake et al. (1993) found that MLU, LARSP mean clausal stage, and their own measure of syntactic complexity were highly correlated, though LARSP mean phrasal stage was less correlated. This study provided evidence that MSC may be a valid measure of syntactic complexity in children's language. MSC can be calculated by CP software, but not by the CLAN or SALT software packages.

Limitations of MSC. Like PESP, the main limitation of MSC is that it is not supported by research. Blake et al.'s (1993) article is the only available article that studied MSC as a measure of syntactic complexity in children's language samples. More

research is necessary to confirm the validity of MSC as a measure of syntactic complexity.

Software Packages for Automated Analysis

Three different software packages calculate the automated syntactic analysis of language samples. These packages are CLAN, SALT, and CP. Each software package differs in the focus of the analysis, the scope of analyses performed, the amount of learning required to operate the software programs, as well as several other notable ways that will be outlined.

SALT. SALT is a program created by Miller (2000) that runs only on Windows, and unlike CLAN and Computerized profiling, it is not free. The current price is $99. There is also a student version available for $30 (www.saltsoftware.com). The manual for the software has a tutorial with instructions for formatting language samples. Video tutorials are available online, and users can also contact support by phone (an 800 number) or by email.

SALT software computes several measures that are categorized into several categories including transcript length, syntax/morphology, semantics, discourse, intelligibility, mazes and abandoned utterances, verbal facility and rate, and omissions and error codes. Transcript length includes such measures as total utterances and number of complete words. Syntax/morphology includes measures of MLU in words (MLUw) and MLU in morphemes (MLUm). MLU-2 can also be approximated by choosing which types of utterances to include in the calculation.

Semantics measures include type-token ratio and number of different word roots. In the discourse category, measures include percentage of responses to questions, mean turn length, and number of utterances with overlaps. The intelligibility category measures

include the percent of intelligible utterances. In the mazes and abandoned utterances category, measures can be made of utterances with mazes, number of mazes, percent of maze words compared to total number of words, and abandoned utterances. More in-depth measures of mazes and abandoned utterances can be made as well. Verbal facility and rate measures include words per minute, between utterances pauses, within utterance pauses, and times for pauses if they are input by the clinician. The category of omissions and error codes measures omitted words, omitted bound morphemes, word-level error codes and utterance-level error codes.

A reference database is available online. Manual formatting of language samples involves slash-coding of inflectional morphemes. Files can be imported in CLAN or Computerized profiling format, but the slash-coding is not included on import; this must be done manually.

Miller (1991) studied three different measures of syntactic length calculated using SALT: MLU, number of different words (NDW), and total number of words (TNW). Miller found that all three measures correlated highly with age. Miller also found that all three measures contributed something unique to the prediction of age, but that MLU did not "contribute to the prediction of age beyond the variance accounted for by TNW and NDW" (p. 218). Miller concluded that "a composite measure of NDW and TNW [was] the best predictor of age in both the conversation and narrative sampling conditions" (p. 218-219).

CLAN. The CLAN package is a suite of programs designed to assist in language sample coding and analysis (www.childes.psy.cmu.edu). This software was created by MacWhinney (1991) and is supported by a federal grant and can be downloaded by users

without cost. The documentation for CLAN is extensive, and tutorial information is available for new users.

Language samples can be formatted, searched, and organized with the CLAN package. It is available in Windows, Mac, and Linux versions. Language samples are formatted according to guidelines given in the tutorial information. Clinicians can receive support for questions about the software or help with problems by posting on a bulletin board called *chi-bolts* which is dedicated to CLAN support.

Once CLAN has been downloaded and the sample has been correctly formatted, the clinician can run the desired analyses by typing or selecting commands which name the file of the sample, specify the analyses to be done along with any options selected, and specify the file to receive the output. Files can be imported in SALT format, allowing clinicians to format a sample once but have access to both CLAN and SALT analyses.

CLAN produces many different measures; some are measures of syntactic complexity or length, and others are measures of discourse, morphology, and other areas of language analysis. These measures include: MLU, the mean length of the five longest utterances, the mean length of turn, and type-token ratio. CLAN can also do many different searches for words, word combinations, unique word combinations, and specific instances of words that precede or follow a target word. CLAN does a DSS analysis, although research has shown that this automated DSS has low accuracy and must be checked and corrected by clinicians. CLAN also performs a morphosyntactic analysis, though the manual cautions users that this function is for serious users who are willing to commit significant time to learning about this function and how it can be used on language sample analysis.

CP. CP (Long et al., 2000) is another software package that is free and can be downloaded by clinicians. This program is supported by its first author, Long, and extensive on-line help is available. Tutorial files also help guide the user through the program's basic use. CP is menu-driven and DOS based, thus it will run under Windows or Windows emulation software. User support is handled through email to the first author of the software. Language samples can be imported in CLAN or SALT format.

CP can measure total number of utterances, number of different sentence types, and an index of utterances produced. Semantic measures that can be performed are Profile In Semantics-Lexical (PRISM-L), Analysis of Propositions (APRON), and Early Vocabularies. CP can also perform LARSP, IPSyn, DSS, and Black English Sentence Scoring (BESS). Phonological analyses that CP can perform are Profile of Phonology (PROPH) and Profile of Prosody (PROP). CP will also perform Conversational Acts Profile (CAP) and Narrative Analysis Procedure (NAP) analyses.

Long and Channell (2001) compared automated analyses of syntactic complexity measures to analyses done by hand. Scores for MLU, IPSyn, DSS, and LARSP were obtained using Computerized Profiling software (CP; Long et al., 2000), and these scores were compared to scores that were obtained by hand. All four automated measures were found to correlate sufficiently with non-automated measures.

Channell (2003) performed a comparison of automated DSS and manually coded DSS analysis. In this study, automated DSS and manual DSS had a 78% overall agreement for children who were typically developing. Though manual DSS scores were significantly higher than automated scores, the two had a correlation of $r = .97$. Different categories had agreement levels from 0% to 98%. Levels of agreement for children with

language impairment were approximately 2% lower than for children who were typically developing. Channell concluded that clinicians should continue to check and correct automated DSS scores and that more work on improving automated DSS scoring was needed.

Summary

These software packages can calculate a number of measures, some of which are quantitative measures of syntactic length or complexity. Quantitative measures can be valuable for tracking a child's progress over time. However, it is not known how these measures correlate with each other when used on the same sets of language samples, and how much of this correlation is simply due to a shared association with age. The present study makes such a comparison by using CP to perform automated analysis measures and thus provides evidence for the concurrent validity of these measures.

Method

Samples

Many child language samples were collected by different researchers for a variety of purposes. These samples are the Reno Samples, the Jordan Samples, the Weismer Samples, and the Wymount Samples.

Reno samples. Fujiki, Brinton, and Sonnenberg (1990) collected samples from 30 children: 10 children with specific language impairment, 10 peers who were matched for language age, and 10 peers who were matched for chronological age. These samples were collected as part of a longer sample for a study on conversational repairs in children with SLI. All children in the SLI group had been receiving speech-language therapy since first grade and were receiving treatment at the time of the study as well as being seen by a

learning disabilities specialist, primarily for communication disorders. The average age of the children with SLI was 9;1 years.

Children identified has having SLI demonstrated deficits in both language comprehension and language production but scored within normal limits on a nonverbal intelligence test. This was determined by scores that were at least one standard deviation below age level or below the 15th percentile on two of several expressive and receptive language tests. These children had no history of hearing loss, mental retardation, or sensory deficits. The children in the CA group had an average age of 9;0 years. The children in the LA group had an average age of 6;9 years. In children from both the CA and LA groups, there was no history of hearing loss, speech and language problems, mental retardation, behavioral disturbance, neurological impairment, or academic difficulties that required remedial services.

This study examined these language samples in three groups: samples from children with SLI (RCLI), samples from peers who were matched for LA (RCLA), and samples from peers who were matched for CA (RCCA).

Jordan samples. Collingridge (1998) collected samples of 12 females and 9 males (21 children total) between 6 and 10 years old. All children had language impairment as determined by speech language pathologists. Children were all judged to have intelligible articulation and adequate language skills to participate in a conversation by their speech language pathologists. Ten of the children were attending self-contained classrooms for children with learning disabilities or communication disorders, and eleven of the children were receiving pull-out speech and language therapy. All children were attending an elementary school in the Jordan school district in Utah and were receiving speech

language therapy for communication disorders. Most children came from low to middle income families.

Weismer samples. The Weismer samples were gathered by Weismer (http://childes.psy.cmu.edu/manuals/) as part of a study on children with language impairments and how a limited processing capacity might account for linguistic deficits. The language samples include 112 children who were participating in a 5 year long study. There were 56 participants identified as late talking (WPILT) and 56 control participants (WCP) that were matched for chronological age, nonverbal cognition, and socioeconomic status. Children identified as late talking were identified with the MacArthur-Bates CDI, which measures words produced at 24 months. Children qualified as late talkers if they scored at or below the 10th percentile. Children were evaluated yearly at ages 2;6, 3;6, 4;6, and 5;6 as part of the longitudinal study. This study will use the samples collected at age 4;6 during a playtime conversation (Ec54), the samples collected at age 4;6 during an interview (Int54), and the samples collected at age 5;6 during an interview (66).

Wymount samples. The Wymount samples were gathered by Barber (1989), Chamberlain (1989), and Taylor (1989) as part of three separate thesis studies. The children ranged from 2;6 to 7;11 in age, and none were considered to have language or speech impairments. All children lived in a student housing complex at Brigham Young University in Provo, Utah. Three children from each six month age interval were randomly selected from a pool of volunteers. Each child passed a hearing screening. A language sample of at least 200 child utterances was collected from each child participant, and generally only the child and the examiner were present during the sample

collection. The first ten minutes of each sample were considered to be a warm-up period and were not transcribed.

Procedure

Format of language samples. All language samples were formatted according to SALT software specifications except for the noting of certain morphemes using the slash character. CP accepts language samples in SALT format with slash coding of morphemes, but the slash coding is removed once the language samples are entered into the CP program, so this formatting step was disregarded.

CP software. CP software was used to perform automated measures of syntactic length and complexity. CP is a software package that is free and can be downloaded by clinicians. CP is menu-driven and DOS based, thus it runs under Windows or Windows emulation software. User support is handled through email to the first author of the software. Language samples can be imported in CLAN or SALT format. CP can perform a variety of syntactic and morphological measures. Measures of interest in the present study included MLU, DSS, IPSyn, PESP, and MSC.

Reliability

As a measure of reliability, a different clinician separately coded language samples from 10% of the children. Inter-rater reliability for the samples was 100%.

Data Analysis

Pearson's correlation was used to analyze correlation among all measures of syntactic length and complexity. Also, for samples in which the measures correlated with age, partial correlations were used to remove the effects of age on the correlations between measures.

The Reno samples represent three groups (children with language impairment, age-matched control subjects, and similar language test score matched control subjects). Scores on the measures of syntactic complexity were compared across groups using one-way ANOVAs. The Weismer samples have two groups (children who were identified as late-talkers and typical children) and scores on grammatical complexity measures of these groups were compared using *t*-tests.

Results

Reno Samples

RCLI. The Pearson's correlations between age and each measure for this group are presented in Table 1. The three measures of length include MLUw, MLUm, and MSL. These three measures were highly correlated with each other. MSC, a measure of length and complexity, was also highly correlated to the other measures of length. DSS correlated highly with all measures and correlated most highly with MSC. PESP and IPSyn correlated highly, but overall, IPSyn had the lowest correlations with the other measures.

RCLA. Pearson's correlations among age and all measures for this group are presented in Table 2. For the RCLA, the three length measures correlated with age and with each other. The MSC correlated with all three length measures as well as with DSS and IPSyn. Because the correlations among these measures might be due to their correlation with age, partial correlations were used to remove the effects of age; these correlations are shown in Table 3. The same general pattern of correlation among the measures still exists, though correlation magnitude decreased about 10%.

Table 1

Pearson's Correlations for the RCLI

	MLUw	MLUm	MSL	MSC	PESP	DSS	IPSyn
Age	-.06	.04	-.03	-.09	.20	-.12	.07
MLUw		.98**	.90**	.94**	.85**	.96**	.86**
MLUm			.93**	.93**	.91**	.96**	.88**
MSL				.83**	.73*	.83**	.67*
MSC					.82**	.97**	.81**
PESP						.89**	.90**
DSS							.88**

Note: * $p < .05$, ** $p < .01$

Table 2

Pearson's Correlations for the RCLA

	MLUw	MLUm	MSL	MSC	PESP	DSS	IPSyn
Age	.65*	.67*	.67*	.62	.40	.60	.43
MLUw		.99**	.95**	.88**	.50	.63	.53
MLUm			.97**	.86**	.43	.62	.50
MSL				.87**	.28	.62	.63
MSC					.45	.82**	.73*
PESP						.29	-.07
DSS							.53

Note: * $p < .05$, ** $p < .01$

Table 3

Partial Correlations for the RCLA

	MLUm	MSL	MSC	PESP	DSS	IPSyn
MLUw	.98**	.91**	.80**	.35	.39	.36
MLUm		.94**	.76*	.24	.37	.32
MSL			.78*	.02	.37	.51
MSC				.28	.71*	.65
PESP					.07	-.30
DSS						.37

Note: * $p < .05$, ** $p < .01$

RCCA. Pearson's correlations among age and all measures for this group are presented in Table 4. In this group, age did not correlate with any other measures. The three length measures, MSC, and DSS were all significantly intercorrelated. PESP and IPSyn did not correlate significantly for this group.

Table 4

Pearson's Correlations for the RCCA

	MLUw	MLUm	MSL	MSC	PESP	DSS	IPSyn
Age	-.01	.04	.21	.08	-.20	.31	-.26
MLUw		.99**	.89**	.88**	-.05	.84**	.40
MLUm			.90**	.86**	-.08	.86**	.35
MSL				.95**	-.11	.96**	.22
MSC					.11	.89**	.36
PESP						-.16	.58
DSS							.19

Note: * $p < .05$, ** $p < .01$

Comparison of Reno Groups. Table 5 presents ANOVA data regarding differences among the Reno groups for each of the syntax measures. The groups differed on the MLUw, MLUm, MSL, MSC, and DSS measures. Post-hoc analyses using the Student-Newman-Keuls procedure revealed that for all measures except DSS, the RCLI and RCLA clustered, or formed a homogeneous subset; the RCCA did not cluster. For DSS, the RCCA and RCLA clustered, and the RCLA and RCLI clustered.

Table 5

Analyses of Variance for the Reno Groups

	F	η^2
MLUw	14.39**	.52
MLUm	15.84**	.54
MSL	18.03**	.57
MSC	10.36**	.43
PESP	1.62	.11
DSS	6.59**	.33
IPSyn	0.86	.06

Note: * $p < .05$, ** $p < .01$, $df = 2, 27$ for all comparisons

Wymount Samples

Pearson's correlations among age and all measures for this group are presented in Table 6. In this group, all measures correlated with age and with each other. Partial correlations were performed to remove the shared correlation with age. These correlations are presented in Table 7. The three length measures correlated highly with each other and to MSC. DSS was more moderately correlated with the length measures and MSC. PESP correlated slightly less with the length measures, MSC, and DSS. IPSyn

was moderately but significantly correlated with MSC, DSS, and PESP, but it did not correlate significantly with the length measures.

Table 6

Pearson's Correlations for the Wymount Group

	MLUw	MLUm	MSL	MSC	PESP	DSS	IPSyn
Age	.55**	.56**	.58**	.63**	.52**	.64**	.47**
MLUw		.99**	.98**	.92**	.57**	.79**	.46*
MLUm			.98**	.92**	.58**	.80**	.47**
MSL				.94**	.59**	.80**	.50**
MSC					.58**	.88**	.61**
PESP						.43*	.59**
DSS							.68**

Note: * $p < .05$, ** $p < .01$

Table 7

Partial Correlations for the Wymount Group

	MLUm	MSL	MSC	PESP	DSS	IPSyn
MLUw	.99**	.98**	.89**	.40*	.69**	.27
MLUm		.98**	.88**	.41*	.69**	.28
MSL			.91**	.42*	.69**	.32
MSC				.38*	.79**	.47*
PESP					.14	.46*
DSS						.56**

Note: * $p < .05$, ** $p < .01$

Jordan Samples

Pearson's correlations among age and all measures for this group are presented in Table 8. Age correlated with PESP. The length measures correlated with each other, with

MSC, and with DSS. DSS correlated moderately with all other measures; MSC correlated with all measures except PESP. IPSyn correlated with PESP, as well as with MSC and DSS. Partial correlations removing age are presented in Table 9. With age removed, correlations among length measures were essentially unchanged. Correlations involving other measures decreased slightly.

Table 8

Pearson's Correlations for the Jordan Group

	MLUw	MLUm	MSL	MSC	PESP	DSS	IPSyn
Age	.22	.26	.28	.17	.56**	.34	.32
MLUw		.99**	.99**	.78**	.24	.47*	.27
MLUm			.99**	.77**	.29	.51*	.31
MSL				.79**	.30	.50*	.33
MSC					.27	.77**	.66**
PESP						.49*	.64**
DSS							.87**

Note: * $p < .05$, ** $p < .01$

Table 9

Partial Correlations for the Jordan Group

	MLUm	MSL	MSC	PESP	DSS	IPSyn
MLUw	.99**	.99**	.77**	.15	.43	.22
MLUm		.99**	.76**	.18	.46*	.25
MSL			.79**	.17	.45	.27
MSC				.21	.76**	.64**
PESP					.39	.59*
DSS						.86**

Note: * $p < .05$, ** $p < .01$

Weismer Samples

WCP. The Pearson's correlations between measures for the WCP are presented in Table 10. The three measures of length from the WCP were significantly and highly correlated with each other. MSC correlated significantly with the measures of length and correlated moderately with DSS. PESP and IPSyn correlated significantly with each other. Across most of the measures, correlations seemed to drop for the Int54 samples.

Most measures were consistent across the three samples of each child (fluctuation in *r* less than .2). There were several exceptions to this pattern. The correlation between PESP and MSL ranged from $r = .260$ to $r = .738$. The correlation between IPSyn and MSL ranged from $r = .269$ to $r = .777$. The correlation between PESP and MSC ranged from $r = .293$ and $r = .665$. The correlation between IPSyn and MSC ranged from $r = .404$ to $r = .766$. All other correlations between measures stayed consistent between the three samples of each child (less than .2 difference in correlations).

Table 10

Pearson's Correlations for the WCP Samples

		MLUw	MLUm	MSL	MSC	PESP	DSS	IPSyn
Age	Ec54	-.06	-.03	-.04	-.08	-.08	-.08	.08
	Int54	.00	.04	-.06	.00	.18	.12	.16
	66	.07	.04	.02	.12	.06	.05	.20
MLUw	Ec54		.99**	.97**	.85**	.76**	.73**	.74**
	Int54		.99**	.87**	.82**	.60**	.76**	.58**
	66		.99**	.98**	.93**	.58**	.87**	.69**
MLUm	Ec54			.99**	.87**	.76**	.72**	.78**
	Int54			.84**	.79**	.66**	.76**	.63**
	66			.99**	.93**	.60**	.86**	.71**
MSL	Ec54				.88**	.74**	.73**	.78**
	Int54				.89**	.26	.65**	.27
	66				.95**	.54**	.87**	.66**
MSC	Ec54					.67**	.84**	.77**
	Int54					.29*	.64**	.40**
	66					.50**	.88**	.68**
PESP	Ec54						.59**	.73**
	Int54						.51**	.77**
	66						.46**	.67**
DSS	Ec54							.69**
	Int54							.54**
	66							.65**

Note: * $p < .05$, ** $p < .01$

The partial correlations for the WCP samples are presented in Table 11. The patterns among the partial correlations in the WCP samples were similar to the patterns among the Pearson's correlations. The same four areas where the correlations were not consistent across the three samples in the Pearson's correlations were the same areas where the correlations were not consistent across the three samples in the partial correlations. The correlation between PESP and MSL ranged from $pr = .127$ to $pr = .716$. The correlation between IPSyn and MSL ranged from $pr = .166$ to $pr = .711$. The

correlation between PESP and MSC ranged from *pr* = .207 and *pr* = .555. The correlation between IPSyn and MSC ranged from *pr* = .352 to *pr* = .690. In addition to these four correlations, the correlation between MLUm and IPSyn was not consistent across the three samples in the partial correlations (*r* = .468 to *r* = .708).

Table 11

Partial Correlations for the WCP Samples

		MLUm	MSL	MSC	PESP	DSS	IPSyn
MLUw	Ec54	.99**	.96**	.81**	.68**	.65**	.65**
	Int54	.99**	.90**	.84**	.38**	.72**	.41**
	66	.99**	.99**	.94**	.58**	.86**	.67**
MLUm	Ec54		.99**	.82**	.70**	.63**	.71**
	Int54		.87**	.82**	.47**	.72**	.47**
	66		.99**	.94**	.60**	.86**	.68**
MSL	Ec54			.84**	.72**	.65**	.71**
	Int54			.89**	.13	.62**	.17
	66			.95**	.60**	.87**	.66**
MSC	Ec54				.54**	.80**	.69**
	Int54				.21	.61**	.35*
	66				.56**	.88**	.69**
PESP	Ec54					.42*	.64**
	Int54					.42**	.61**
	66					.49**	.55**
DSS	Ec54						.60**
	Int54						.44**
	66						.65**

Note: * $p < .05$, ** $p < .01$

WPILT. The Pearson's correlations between measures for the WPILT samples are presented in Table 12. Most measures were consistent across the three samples of each child, with correlation values fluctuating less than .2. There were several exceptions to this pattern. The correlation between MLUw and DSS ranged from *r* = .696 to *r* = .932. The correlation between MSC and IPSyn ranged from *r* = .597 to *r* = .802. Finally, the

correlation between PESP and DSS ranged from $r = .350$ to $r = .669$. All other correlations between measures stayed consistent between the three samples of each child (less than .2 difference in correlations). Overall, the range of correlations between the three samples fluctuated less in the WPILT samples than in the WCP samples.

Table 12

Pearson's Correlations for the WPILT Samples

		MLUw	MLUm	MSL	MSC	PESP	DSS	IPSyn
Age	Ec54	.01	.04	.11	-.01	-.15	.03	-.02
	Int54	-.18	-.17	-.14	-.12	-.32	-.19	-.25
	66	.20	.19	.16	.13	.04	.04	.13
MLUw	Ec54		.99**	.86**	.74**	.59**	.79**	.70**
	Int54		.99**	.98**	.92**	.64**	.70**	.57**
	66		.99**	.98**	.93**	.56**	.93**	.75**
MLUm	Ec54			.88**	.75**	.59**	.79**	.69**
	Int54			.99**	.94**	.63**	.71**	.56**
	66			.99**	.93**	.56**	.93**	.74**
MSL	Ec54				.80**	.58**	.78**	.56**
	Int54				.96**	.62**	.73**	.58**
	66				.92**	.51**	.92**	.72**
MSC	Ec54					.49**	.85**	.60**
	Int54					.59**	.75**	.60**
	66					.64**	.91**	.80**
PESP	Ec54						.35*	.59**
	Int54						.67**	.76**
	66						.56**	.70**
DSS	Ec54							.59**
	Int54							.73**
	66							.72**

Note: * $p < .05$, ** $p < .01$

The partial correlations between measures for this group are presented in Table 13. The patterns among the partial correlations were similar to the patterns among the Pearson's correlations. However, there were more correlations that were not consistent

across the three samples in the partial correlations than the Pearson's correlations. The correlation between PESP and MLUw ranged from *pr* = .294 to *pr* = .619. The correlation between DSS and MLUw ranged from *pr* = .634 to *pr* = .909. The correlation between MLUm and DSS ranged from *pr* = .653 to *pr* = .907. The correlation between PESP and MSL ranged from *pr* = .287 to *pr* = .628. The correlation between DSS and PESP range from *pr* = .273 to *pr* = .611. The correlation between IPSyn and PESP ranged from *pr* = .337 to *pr* = .658.

Table 13

Partial Correlations for the WPILT Samples

		MLUm	MSL	MSC	PESP	DSS	IPSyn
MLUw	Ec54	.99**	.86**	.74**	.60**	.79**	.70**
	Int54	.99**	.98**	.92**	.64**	.70**	.57**
	66	.99**	.98**	.93**	.56**	.93**	.75**
MLUm	Ec54		.88**	.75**	.59**	.79**	.69**
	Int54		.99**	.94**	.63**	.71**	.56**
	66		.99**	.93**	.56**	.93**	.74**
MSL	Ec54			.80**	.58**	.78**	.56**
	Int54			.96**	.62**	.73**	.58**
	66			.92**	.51**	.92**	.72**
MSC	Ec54				.49**	.85**	.60**
	Int54				.59**	.75**	.60**
	66				.64**	.91**	.80**
PESP	Ec54					.35*	.59**
	Int54					.67**	.76**
	66					.56**	.70**
DSS	Ec54						.59**
	Int54						.73**
	66						.72**

Note: * $p < .05$, ** $p < .01$

Comparison of Weismer Groups. *T*-tests were used for the Weismer samples to compare the scores on the measures of syntactic complexity between the two groups (WCP and WPILT). Due to the large number of comparisons in the Weismer samples, the critical alpha level was set at .005 instead of .05. In the Ec54 group, PESP was the only measure that was statistically significant between the two groups at $p = .005$. In the Int54 group, both MSL and MSC were statistically significant between the two groups at $p = .005$. In the 66 group, none of the measures were significant at $p = .005$.

Discussion

This study examined five different measures of syntactic complexity (with two additional variations of MLU) and how these measures correlated when used to examine the same sets of child language samples. The five different syntactic measurements can be categorized into three different constructs of measuring language: those that assess length, those that use a system that gives different grammatical elements different weights according to complexity, and those that look at an inventory of grammatical elements.

General Patterns

Across all the language samples, the three measures of length (MLUw, MLUm, and MSL) were highly correlated. MSC and DSS, the weighted measures which measure both length and complexity, tended to correlate, though interestingly, MSC tended to correlate more highly with the length measures than with DSS, especially in the typical samples. IPSyn and PESP, both inventory measures that look at specific syntactic elements, correlated in some of the samples but not in all samples. Also, this correlation tended to diminish in the partial correlations.

IPSyn, in general, did not correlate highly with the other measures once age had been controlled; the measure it correlated with most highly was DSS. This is surprising since IPSyn is an inventory measure and DSS is a weighted measure. Table 14 shows the frequency of correlations between each measure for all samples studied. The number of times each measure correlated significantly with another measure is listed, with a maximum of 11 and a minimum of 0.

Table 14

Frequency of Correlations for All Samples

	MLUm	MSL	MSC	PESP	DSS	IPSyn
MLUw	11	11	11	8	10	8
MLUm		11	11	8	10	8
MSL			11	7	10	7
MSC				8	11	10
PESP					9	9
DSS						9

As can be seen in Table 14, the three length measures correlated with each other in every sample. Overall, PESP and IPSyn were the two measures that correlated the least frequently with other measures. DSS correlated almost as frequently as the length measures.

Patterns Between Groups of Children with LI and Typical Children

For the Reno samples, every measure but PESP and IPSyn showed a significant difference between the group of children with language impairment and the group of typical age-matched peers. For the Weismer samples, PESP was the only measure that showed a significant difference between the two groups in the Ec54 group. In the Int54

group, only MSL and MSC showed a significant difference between the two groups, and in the 66 group, none of the measures showed a significant difference between the two groups.

Measures of Syntactic Complexity Studied

Length measures. Three different measures of length were compared in the present study. To answer the question of why three different measures of length were studied, a brief overview of the different measures is required. The three length measures studied were MLUw, MLUm, and MSL. MLUw is very easy to compute, but it was rejected when Brown (1973) presented the MLUm measure. MLUw does not account for inflectional morphemes. However, this measure is still used by many clinicians who wish to make quick and immediate judgments about a child's language. MLUm correlates significantly with age in younger children (Miller & Chapman, 1981). It has also been found that MLUm is associated with several areas of linguistic competence in typically developing children. MSL is similar to MLUm with one difference. Klee and Fitzgerald (1985) developed MSL which is calculated by excluding answers to yes/no questions.

The benefit of using automated versions of the length measures is that it would take approximately the same amount of time to compute all three measures using CP as it would to compute one of the measures manually. Using CP, the clinician can look at all three scores (MLUw, MLUm, and MSL) and compare the scores. This can help give a more complete picture of the language sample and can help the clinician determine what role context played in the score.

Though there are differences between the three length measures, all three measures were highly correlated in all of the different groups. There are a few reasons for this high correlation. MLUm is often used in place of MLUw because it accounts for

inflectional morphemes and can show more information about syntactic development. However, the difference between MLUm and MLUw becomes more consistent as a child gets older. The samples studied in the present study were samples from children 54 months and older. Because the children in the sample were older, the difference between MLUm and MLUw was likely consistent, which accounts for the high correlation between these two measures. MSL, in each group of samples, was less correlated with MLUw and MLUm than the two MLU measures. This could be because MSL excludes one-word responses to yes/no questions, so MSL counted less of the utterances than both of the MLUs. However, despite this minor difference, MSL was still very highly correlated to the two versions of MLU.

Weighted measures. MSC is a measure that attempts to quantify LARSP and looks at the syntactic inventory of a child's language sample as well as length (Blake et al., 1993). The manual version of MSC might be used by clinicians who wish to look more in depth at a child's syntactic development and have time to take a more exhaustive inventory. Manually, MSC takes more time than the length measures, but MSC looks at syntactic elements as well as length and may give a more complete picture of a child's linguistic abilities. The automated version, in contrast, takes no more time than any of the other measures to compute. The benefit of the automated version is that the score is obtained quickly. If a clinician wishes to use MSC as a measure either to show progress in therapy or to compare a child's language to peers, a numbered score would be of value to the clinician.

DSS, like MSC, looks at syntactic elements and gives each element a weighted value. DSS looks at specific grammatical constructs and gives an utterance more points

for elements that are more complex. The manual version of DSS might be used by a clinician who is confident in their grammatical abilities and who wishes to take an exhaustive inventory of the child's language sample. The benefit of using the automated version of DSS is, once again, saved time for the clinician. The manual version of DSS can take several hours and requires a great amount of grammatical knowledge on the part of the clinician. The automated version of DSS, in contrast, can be computed in several minutes and it can be computed by clinicians who are less confident in their grammatical abilities. One drawback of the automated version, however, is that the coding for embedded structures is not very accurate, and some of the other coded levels may need to be checked by the clinician (Channell, 2003) if performance in specific areas is of interest.

DSS, unlike MSC, makes use of data-derived weights for different syntactic elements. MSC simply counts up the elements in a child's utterance, so a sentence with more elements in it will get a higher score. MSC, therefore, is more susceptible to length; longer utterances will receive higher scores. This might explain why MSC was so highly correlated with the length measures. DSS was not as highly correlated with the length measures as MSC, but DSS was more highly correlated with MSC than any of the other measures. This is not surprising, because both MSC and DSS rely on weighted values.

Inventory measures. PESP, like MSC, is a quantification of LARSP. However, unlike MSC, sentences do not get higher scores for having more elements in them. The manual version of PESP might be used in a clinical setting because it is more quickly derived than some of the other inventory measures; however, with the automated versions, all measures can be computed within minutes. Because PESP is not supported

by research, it may not be the best choice for clinicians. However, it can give a clinician a quick list of all the elements of interest and reference each utterance that contains the element. PESP was more highly correlated with the other measures than IPSyn, but the correlations were not as high as correlations between the previously mentioned measures.

IPSyn was designed for preschool language samples and, like PESP, uses a list of morphemes that are counted for scoring. However, many of the morphemes that IPSyn looks at are morphemes that should develop in the preschool years. IPSyn was not highly correlated with the other measures. Of all the measures studied, IPSyn had the lowest correlations with other measures. This may be because IPSyn is not an appropriate measure for looking at language samples of children who are older than five years old. Also, IPSyn requires that a 100 utterance block is used. Some of the Weismer language samples were not 100 utterances in length and could not be used. IPSyn measures morphemes that should develop in the preschool years, so one might expect that a typical child over the age of 5 would "top out" on their score of IPSyn. It is interesting to note that this was not the case with the samples in the present study. Because IPSyn looks at only 100 utterances and no more, perhaps it misses some of the utterances that might indicate development in a child's language sample.

The results of the present study are consistent with the results obtained by Kemper et al. (1995). In their study, the measures of syntactic length were highly correlated, and DSS was also correlated with the length measures. Also in the Kemper et al. study, IPSyn failed to show child development for children ages 5 to 10, which is consistent with the findings of the present study. In the present study, PESP and IPSyn

were the only two measures that did not distinguish a difference between the group of children with language impairment and the typical group for the Reno samples.

Limitations and Suggestions for Future Research

Pre-collected samples were used in this study. These samples were not ethnically, socioeconomically, or regionally balanced. This may limit the extent to which these results can be generalized. Also, sample size may have affected how accurate the findings are. Some of the Weismer samples were fairly short (less than 100 utterances); the correlations in this group of samples were lower and less consistent than the correlations in the Reno samples, which were much longer samples.

Future research could repeat the present study with new samples that were balanced in terms of ethnicity, socioeconomic status, and region. Also, a future study could collect samples that are all 150 child utterances or longer to ensure more accurate numbers for each measure. Another interesting area to research would be how these measures compare in populations of bilingual children or children from ethnically diverse backgrounds.

Another limitation of the present study is the availability of automated measures. This study compared MLU, MSC, DSS, PESP, and IPSyn. These were the only measures for which there was an available automated version with CP software. However, these are not the only measures that are used clinically to analyze child language samples. Future research could determine which measures are the most commonly used measures in clinical settings, and then use those measures in the comparisons. Some measures may not be available in automated format; however, there are several different software packages that perform automated measures of syntactic complexity. A future study could use different software packages to compare the different measures.

The present study compared automated measures to one another. However, Channell (2003) suggested that the automated version of DSS may not be as accurate as the manual version. Sagae et al. (2005) compared the automated version of IPSyn to the manual version and also found that the automated version was not as accurate. It is possible that the relationships between the different measures of syntactic complexity may have been skewed because of these inaccuracies. A future study could compare the manual versions of these measures and compare how the manual versions correlate to each other as well as to the automated versions of the measures.

Implications

Measures of syntactic complexity are useful for clinicians working with children. Formal tests often leave out important information about a child's linguistic capabilities whereas informal language samples can give a more complete description of the child's capabilities (DeThorne et al., 2005; Scarborough, 1990). Software packages that can compute the syntactic measures automatically can help speed up the process of computing these measures. For clinicians who have limited time to spend on assessment, automatic versions can make the process quicker and therefore make clinicians more likely to use informal measures. The present study found consistent patterns of correlation among these automated measures of syntactic complexity and thus provides the first evidence of concurrent validity for these measures.

The patterns of correlation seen in the present study generally held true across several different sample sets. Though these samples were not specifically balanced geographically or ethnographically, the samples do represent a diverse population of children. Samples were collected by different examiners in different sampling situations

from both typically developing children and children with language impairment, and had been collected in Utah, Nevada, and Wisconsin.

The patterns correlations in the present study were consistent across samples from diverse children, but some measures were consistently strongly correlated and others were consistently less correlated. This finding suggests that these different kinds of measures might be measuring different aspects of syntactic complexity, rather than measuring a single aspect but doing it poorly. Because these automated measures are simple and rapid to compute, clinicians who wish to use these automated measures might consider using one measure from each of the three different types: measures of length, measures of inventory, and weighted measures. Such a strategy might provide the best quantitative assessment of a client's syntactic complexity.

References

Barber, M. (1989). Children's repetition of sentences previously produced spontaneously. Unpublished Master's Thesis, Brigham Young University.

Blake, J., & Quartaro, G. (1990). *Manual for recording, transcribing, and analyzing preschool children's speech samples*. York University Department of Psychology Report No. 189.

Blake, J., Quartaro, G., & Onorati, S. (1993). Evaluating quantitative measures of grammatical complexity in spontaneous speech samples. *Journal of Child Language, 20*, 139-152.

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

Channell, R. W. (2003). Automated developmental sentence scoring using computerized profiling software. *American Journal of Speech Language Pathology, 12*(3), 369-375.

Cheung, H., & Kemper, S. (1992). Competing complexity metrics and adults' production of complex sentences. *Applied Psycholinguistics, 13*, 53-76.

Collingridge, J. D. (1998). *Comparison of DSS scores from online and subsequent language sample transcriptions.* Unpublished master's thesis, Brigham Young University, Provo, UT.

Crystal, D., Fletcher, P., & Garman, M. (1989). *The grammatical analysis of language disability (2nd ed.)*. London, England: Cole and Whurr.

DeThorne, L. S., Petrill, S. A., Hayiou-Thomas, M. E., & Plomin, R. (2005). A closer look at MLU: What does it really measure? *Clinical Linguistics and Phonetics, 19*, 635-648.

Fey, M. E., Cleave, P. L., Long, S. H., & Hughes, D. L. (1993). Two approaches to the facilitation of grammar in children with language impairment: an experimental evaluation. *Journal of Speech and Hearing Research, 36*(1), 141-157.

Fristoe, M. (1979). Developmental sentence analysis. In F. L. Darley (Ed.), *Evaluation of appraisal techniques in speech and language pathology*. Reading, MA: Addison-Wesley.

Fujiki, M., & Brinton, B. (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 and Hearing Research, 39*, 1258-1262.

Holdgrafer, G. (1995). Comparison of two methods for scoring syntactic complexity. *Perceptual and Motor Skills, 81*, 498.

Hughes, D. L., Fey, M. E., Kertoy, M. K., & Nelson, N. W. (1994). Computer-assisted instruction for learning developmental sentence scoring: An experimental comparison. *American Journal of Speech Language Pathology, 3*(3), 89-95.

Hughes, D. L., Fey, M. E., & Long, S. H. (1992). Developmental sentence scoring: Still useful after all these years. *Topics in Language Disorders, 12*(2), 1-12.

Johnson, M. R., & Tomblin, J. B. (1975). The reliability of developmental sentence scoring as a function of sample size. *Journal of Speech and Hearing Research, 18*, 372-380.

Johnston, J., & Kamhi, A. (1984). The same can be less: syntactic and semantic aspects of the utterances of language-impaired children. *Merrill Palmer Quarterly, 30*, 65-86.

Johnston, J. R. (2001). An alternate MLU calculation: magnitude and variability of effects. *Journal of Speech, Language, and Hearing Research, 44*, 156-164.

Kemp, K., & Klee, T. (1997). Clinical language sampling practices: Results of a survey of speech-language pathologists in the United States. *Child Language Teaching and Therapy, 13*, 161-176.

Kemper, S., Rice, K., & Chen, Y. (1995). Complexity metrics and growth curves for measuring grammatical development from five to ten. *First Language, 15*, 151-166.

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.

Lee, L. L. (1974). *Developmental sentence analysis*. Evanston, IL: Northwestern University Press.

Liles, B., & Watt, J. (1984). On the meaning of 'language delay'. *Folia Phoniatrica, 36*, 40-48.

Lively, M. (1984). Developmental sentence scoring: Common scoring errors. *Language, Speech, and Hearing Services in Schools, 15*, 154-168.

Long, S. H. (1996). Why Johnny (or Joanne) can't parse. *American Journal of Speech-Language Pathology, 5*(2), 35-42.

Long, S. H. (2001). About time: A comparison of computerized and manual procedures for grammatical and phonological analysis. *Clinical Linguistics and Phonetics, 15*(5), 399-426.

Long, S. H., & Channell, R. W. (2001). Accuracy of four language analysis procedures performed automatically. *American Journal of Speech-Language Pathology, 10*, 180-188.

Long, S. H., Fey, M. E., & Channell, R. W. (2000). Computerized Profiling (CP) (Version 9.2.7, MS-DOS) [Computer software]. Cleveland, OH: Department of Communication Sciences, Case Western Reserve University.

MacWhinney, B. (1991). *The CHILDES project: Tools for analyzing talk*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Miller, J. F. (1981). *Assessing language production in children*. Baltimore, MD: University Park Press.

Miller, J. F. (1991). *Research on child language disorders: A decade of progress*. Austin, TX: PRO-ED.

Miller, J. F., & Chapman, R. (1981). The relationship of age and MLU in morphemes. *Journal of Speech and Hearing Research, 24*(154-161).

Miller, J. F., & Chapman, R. S. (2000). Systematic Analysis of Language Transcripts (*SALT*, Version 6.1, Windows) [Computer software]. Madison, WI: Language Analysis Laboratory, Waisman Center on Mental Retardation and Human Development.

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

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.

Sagae, K., Lavie, A., & MacWhinney, B. (2005). *Automatic Measurement of Syntactic Development in Child Language.* Paper presented at the 43rd Meeting of the Assiciation for Computational Linguistics, Ann Arbor, Michigan.

Scarborough, H. S. (1990). Index of Productive Syntax. *Applied Psycholinguistics, 11*, 1-22.

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).

Ward, E., & Fisher, J. (1990). A picture-elicited screening procedure. *Child Language Teaching and Therapy, 6*(2), 147-159.

Appendix

Steps for Computing Measures with CP

1. Open CP. Type Ret three times to bring up the Main Menu.

2. Type 1, 1, 3; then type 2 for SALT format. Select the desired file. Type C for Continue, then type 1 for "C" as target, and type 1 again for the child utterances to be analyzed. Type Ret to accept all default classifications of the *'s* element. This step may need to be repeated several times until all defaults are accepted. If an utterance is longer than 20 words, a number must be typed to split the utterance. Type a number to choose a point at which to split the utterance. Then type Ret to accept the corpus file name. Type Esc to return to the Main Menu.

3. Type 5 for LARSP, type 1 to create the LARSP file, select the desired file, then type Y (for Yes) to code all repetitions as stereotypes. Type Ret, then type Y to "Analyze all single-word utterances as Stage 1". When it finishes all the utterances, type 3 to tabulate the LARSP file. Type Ret three times to skip the top of the profile, then type Ret to start the tabulation. Type P for LARSP Profile, type 1, then review profile to get the Number of Utterances, MLU in words, MLU in morphemes, and MSL. The Blake & Quartaro MSC measure can be found at the bottom of the page. Then type Esc 2 times to return to the Main Menu.

4. Type 5 for LARSP again, then type 6 this time to choose PESP Score. Select the desired file, then type V for View/Print. The score is on the next to last line. Type Esc 2 times to return to the Main Menu.

5. Type 7 for DSS, type 1 for Create DSS, then select the desired file. Type C for Continue, then type V for View Profile, then type N for Norms. Type in a dummy age ("66"), then type Ret to get the DSS score. Type Esc 2 times to return to the Main Menu.

6. Type 6 for IPSyn, type 1 to Create IPSyn, then select the desired file. Type Ret to select the default 25 limit. Type Ret to begin on utterance 1. Then a pop-up window asks "Run Index Utterances to Identify Repetitions?" Type Y for Yes, then type C for code repetitions. Type Y for Continue, then type Esc to return to the Main Menu. Type 6 for IPSyn again. Type 1 to Create File, select the desired file, then type Ret to accept the limit of 25. Type Ret to begin on utterance 1. A pop-up window will ask "Run Utterance continue to find cutoff?" Type Y, then type Esc. Type Ret on Limit 25, then type Ret for Begin on 1, and type Ret for End on Calculated End Utterance. Type Ret to truncate, then type E for Edit/Print Profile. The IPSyn score is three-fourths of the way down the page.

Note: Ret = Return, Esc = Escape