2006-07-14

Accuracy of Automated Developmental Sentence Scoring Software

Carrie Ann Judson
Brigham Young University - Provo

ACCURACY OF AUTOMATED DEVELOPMENTAL SENTENCE SCORING SOFTWARE

by

Carrie Judson

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 2006

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a thesis submitted by

Carrie Judson

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 Carrie Judson in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

Date

Ron W. Channell
Chair, Graduate Committee

Accepted for the Department

Ron W. Channell
Graduate Coordinator

Accepted for the College

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

ABSTRACT

ACCURACY OF AUTOMATED DEVELOPMENTAL SENTENCE SCORING SOFTWARE

Carrie Judson

Department of Communication Disorders

Master of Science

Developmental Sentence Scoring (DSS; Lee 1974) is a well established, structured method for analyzing a child's expressive syntax within the context of a conversational speech sample. Automated DSS programs may increase efficiency of DSS analysis; however the program must be accurate in order to yield valid and reliable results. A recent study by Sagae, Lavie, and MacWhinney (2005) proposed a new method for analyzing the accuracy of automated language analysis programs. This method was used in addition to previously established methods to analyze the accuracy of a new automated DSS program, entitled DSSA (Channell, 2006).

Previously collected language samples from 118 children between the ages of 3 and 11 years in age were manually and automatedly coded for DSS. The overall accuracy of DSSA was about 86%, while the mean point difference was approximately .7. DSSA generally scored language samples of children achieving lower manual DSS scores or

children with language impairment with less accuracy than those of other children. While some precautions may need to be taken, accuracy levels are sufficiently high to allow the fully automated use of DSSA as an alternative to manual DSS scoring.

ACKNOWLEDGMENTS

I would like to acknowledge my professors, clinic supervisors, and internship supervisors at both Idaho State University and Brigham Young University for their instruction and genuine interest in my education. I would also like to thank Ron Channell for his encouragement, guidance, and willingness to assist me in my research efforts. Finally, I would like to thank my husband for his support in this endeavor.

Table of Contents

Page

List of Tables

Table Page

Introduction

Developmental Sentence Scoring (DSS; Lee, 1974) continues to be a widely used method for analyzing a child's syntax within a language sample. In the most recent published surveys of clinician use of language sample analysis, DSS was the technique most frequently mentioned (Hux, Morris-Friehe, & Sanger, 1993; Kemp & Klee, 1997). DSS is based on the typical sequence of acquisition of eight grammatical areas used in standard American English and thus provides a structural framework for analysis. Fifty consecutive, non-imitated, non-repeated utterances containing a subject and verb serve as the basis of DSS. The user awards up to eight points for grammatical forms found in the child's utterances, with later developing forms receiving more points. The points awarded each utterance are summed, and this sum is divided by the number of sentences analyzed to arrive at an overall DSS score.

Some of the strengths of DSS include the possibility of referencing the DSS score to normative data for comparison of a child to chronological age peers. DSS also provides information that is useful in making clinical decisions (Hughes, Fey, & Long, 1992). Although DSS has been found to discriminate between normally developing children and those with language impairments (Liles & Watt, 1984), it was not designed to be used alone in judging whether or not a child has a language disorder, because it does not analyze all aspects of a child's language (Lee, 1974). However, the results of DSS can assist clinicians in choosing treatment goals and evaluating efficacy of treatment. In addition, adaptations of DSS exist for analyzing versions of Black English (Nelson, 1976) and Spanish (Toronto, 1976).

DSS does have limitations as well, including insufficiency in the size and diversity of the normative group (Fristoe, 1979; Vaughn-Cook, 1983); consequently, normative results must be interpreted with caution. While adaptations have been made for use of DSS in analyzing Black English (Nelson, 1976), the validity of this adaptation has been questioned (Vaughn-Cooke, 1983). Another weakness includes the possibility that the specified minimal sample size of 50 sentences may be inadequate, yielding unreliable results (Johnson & Tomblin, 1975) and that typical development of grammatical forms established by Lee (1974) may be inaccurate (Klee, 1985). In addition, because many different language profiles could be represented by the same DSS score, simplification of syntax abilities into a DSS score without additional analysis may lead to inferior assessment (Hughes et al., 1992; Klee, 1985). A final weakness includes the complexity and time required in using DSS (Fristoe, 1979).

Automated DSS programs have been made in recent years including Hixson's DSS Computer Program (DSSCP; 1983), Child Language Analysis (CLAN; MacWhinney 1991, 2006), and Computerized Profiling (CP; Long, 1986; Long & Fey, 1993; Long, Fey, & Channell, 2000). CP allows for efficient analysis (Gregg & Andrews, 1995; Long, 2001; Long & Masterson, 1993) and has also been found to be fairly accurate (Channell, 2003; Long, 2001; Long & Channell 2001). However, improvements need to be made in existing automated DSS programs before they can be used without post-editing.

Channell (2006) recently released a new software program, called DSSA, which performs automated DSS analysis. The accuracy of this program has not yet been carefully scrutinized and compared to other automated DSS software programs. In

addition, a recent study by Sagae, Lavie, and MacWhinney (2005) proposed a method called *Point Difference* for comparing automated scores to manual scores. Until now, quantification of DSS accuracy has been based on per cell agreement, not the overall score. The purpose of this study was to evaluate the accuracy of DSSA in terms of point by point comparison to manual scoring as well as in terms of point difference in order to determine if DSSA's accuracy is comparable with manual DSS scoring to the extent that DSSA could replace manual scoring.

Review of Literature

Description of Language Sampling

The merits and limitations of standardized and non-standardized assessment procedures have long been debated. Language sample analysis is one informal method of choice among many speech-language pathologists working with children. Gallagher (1983) referred to language sampling as "the centerpiece of child language assessment" (p. 2). Emphasizing the importance of language sampling, Klee (1985) stated, "A language sample allows for the direct assessment of linguistic behaviour, behaviour which is not influenced by the constraints and restrictions imposed by standardized testing" (p. 183). However, language sampling and the analysis that follows can be complex and time consuming.

Prevalence of Language Sampling

Language sampling is commonly used by speech-language pathologists in California. Wilson, Blackmon, Hall, and Elcholtz (1991) surveyed speech-language pathologists working in California schools concerning assessment procedures used. Of the 253 individuals responding to a question about use of language sampling, only 5 individuals reported that they did not use language sampling. On average, language sampling practices were reportedly used in 75% of assessments. Under California law, children could qualify for speech or language intervention services without the use of language sampling, yet Wilson et al. found that language sampling continued to be a commonly used method among speech-language pathologists.

A survey by Hux et al. (1993) provides further insight as to speech-language pathologists' practices and attitudes concerning language sampling. A total of 500

surveys were sent to school-based speech-language pathologists from ten Midwestern states. Two-hundred thirty-nine individuals participated in the survey. While 82% of respondents reported that language sampling was not required by local or state agencies, it continued to be commonly used. Language sampling was most often used for supplementing standardized tests (80%), planning intervention (77%), and evaluating efficacy of treatment (62%) or program (54%). Hux et al. reported that a majority of those surveyed indicated that they felt that language sampling was reliable and useful in identifying individuals with language delays.

Although previous studies had been conducted to assess language sampling practices of speech-language pathologists, no surveys had used randomly selected participants throughout the United States before the Kemp and Klee (1997) survey. Kemp and Klee surveyed the language assessment practices of speech-language pathologists working with preschool children. Five hundred surveys were sent to randomly selected speech-language pathologists living in the continental United States; 253 individuals responded. Eighty-five percent of clinicians surveyed reported using language sample analysis for assessment purposes. Thirty-seven percent of those surveyed responded to a question regarding reasons for not using language sample analysis; of these individuals, lack of time was the most common reason given (86%).

Description of Developmental Sentence Scoring

Developmental Sentence Scoring (DSS), a component of Developmental Sentence Analysis (DSA), was developed by Lee (1974) to provide a basis for analyzing a child's syntax, and among other things, offers norm-referenced data for comparison of a child's performance to that of same age peers. Results were anticipated to help clinicians

determine if a child's syntax in conversational speech is within functional limits compared to same age peers, and is improving over time. DSS also assists clinicians in identifying weak areas in grammar use that could be targeted in treatment.

DSS was designed for children who speak standard American English (Lee, 1974). Normative data was developed by Lee based on language samples collected from 200 children ages 2;0 (years;months) to 6;11 from middle income families in the middle west and Maryland who spoke standard American English. Results of five males and five females were used to develop norms for each three month interval. Normative group members did not demonstrate signs of hearing loss, or behavior anomalies, and were judged by an interviewer to be intelligible. The children had normal developmental and social histories as well. Later, norms were developed by Stephens, Dallman, and Montgomery (1988) for children age 6;0 to 9;11.

Lee (1974) specified that utterances used for DSS analysis should consist of 50 complete sentences, defined as containing a subject and a verb in a subject – predicate relationship. Imperatives, including negative imperatives, also qualified as complete sentences since the subject is implied. Lee specified that the 50 utterances selected should be consecutive. In addition, the clinician should not score any repetitions of the same utterances and should exclude any phrases that were imitated by the child.

Grammatical forms of the utterances are classified into one of eight categories, which include (1) indefinite pronoun or noun modifier, (2) personal pronoun, (3) main verb, (4) secondary verb, (5) negative, (6) conjunction, (7) interrogative reversal in questions, and (8) wh-question (Lee, 1974). Between one and eight points are awarded according to order in which grammatical forms are typically acquired, with higher point

values being assigned to forms that develop later. An additional "sentence point" is awarded each utterance which is correct in every way, thus allowing a minimal assessment of aspects of language beyond the included eight categories.

Descriptions of DSS date back to 1971; a final version of DSS was made by Lee in 1974. Several changes were made over this three year span. The 1971 version (Lee & Canter, 1971) included the same eight grammatical categories as the finalized version; however some grammatical forms were initially assigned different point values. In order to allow for inter-category comparison, some changes were made in initial point value assignments (Lee, 1974). For example, *and* was initially given a point value of one because it is typically the first conjunction to develop in a child's speech; however, *and* typically develops after some pronouns and verb tenses. Therefore, this conjunction was later given a higher score, allowing for comparison across categories.

DSS results assist clinicians in making clinical decisions (Hughes et al., 1992). Because DSS does not address all aspects of language, other assessment methods should be used in conjunction with DSS in order to determine whether or not a child has a language impairment (Lee, 1974; Lively, 1984). DSS is useful in choosing treatment goals and in evaluating a child's progress during treatment (Lee, 1974; Lively, 1984). For example, children's incorrect attempts at grammatical forms are not awarded points but can be noted. Clinicians can evaluate grammatical structures that the child has mastered, those that are emerging, and those that are absent in order to choose treatment goals or evaluate progress. For these reasons and others, DSS has become one of the most commonly used procedures for analyzing children's syntactic development within conversational speech (Hux et al., 1993; Kemp & Klee, 1997; Lively, 1984).

Reliability of DSS and Related Procedures

A study by Johnson and Tomblin (1975) analyzed the reliability of utterance sample sizes larger and smaller than the 50 utterance sample size recommended by Lee (1974). Johnson and Tomblin selected 50 preschool children between the ages of 4;8 and 5;8 from the University of Iowa Institute for Child Development who had normal hearing. The children were presented with several types of stimuli, and language samples were obtained. Twenty-five sentences meeting previously established DSS requirements were broken into five segments. DSS total scores and category scores were calculated for each five-sentence segment. A total of 250 score values (i.e. five scores from each of the 50 subjects) were used to find a formula for predicting reliability and standard error of measurement as a function of sample size ranging from 5 to 250 sentences.

As one might predict, results indicated that as the sample size increased, the reliability of the DSS total score and grammatical forms categorical scores increased (Johnson and Tomblin, 1975). With a sample size of 50, as recommended by Lee (1974), the total reliability was estimated to equal .75 with a standard error of measurement equaling 4.43 (Johnson and Tomblin, 1975). The authors suggested using sample sizes as large as 175 utterances to improve reliability but acknowledged the difficulty of obtaining such a large sample.

Gavin and Giles (1996) found that temporal reliability of syntactic measures increased with increased length of language samples. Twenty preschool children participated in the study, which analyzed language samples in terms of four measures, including number of different words, total number of words, mean syntactic length, and mean length of utterance in morphemes. The authors concluded that when using the four

previously mentioned analysis techniques, language samples from preschool children should contain at least 175 utterances in order to obtain acceptable temporal reliability levels. While this study did not directly address temporal reliability of DSS, the study served to illustrate that utterance length did have an effect on the language analysis procedures used, corroborating the conclusions of Johnson and Tomblin (1975).

Muma (1998) assessed seven normally developing children's repertoires in grammatical categories including subject nominals, object nominals, auxiliaries, verbals, and grammatical operations. Muma calculated accuracy rates to be as low as 40% and 55% when comparing 50- and 100-utterance language samples, respectively, to 400-utterance samples. Ten 400-utterance conversational language samples were gathered from seven children age 2;2 to 5;2. In order to obtain a more accurate representation of a child's grammatical repertoire, Muma recommended using 200- to 300-utterance language samples. These findings suggest implications for reliability of syntactic measures such as DSS.

Prevalence of DSS

A survey by Hux et al. (1993) of 239 speech-language pathologists found that of the standardized language analysis procedures, 31% of those surveyed reported frequent use of DSS, making DSS the most commonly used standardized procedure, and the only standardized procedure used with regularity. The survey reported that of those clinicians who indicated a preferred language sample analysis procedure, 15% chose DSS, ranking DSS within the top three most preferred procedures.

Kemp and Klee's (1997) survey found that 35% of the clincians who reported using language sample analysis reported using DSS to analyse those samples.

Strengths and Limitations of DSS

DSS has been carefully evaluated over the years; its strengths and limitations are well documented (Hughes et al., 1992).

Time requirements and complexity of DSS analysis have made DSS difficult to implement for some clinicians (Fristoe, 1979). Collecting and transcribing a language sample is often a time consuming task. The DSS procedure may require a substantial amount of time as well, depending upon factors including the clinician's skill and experience in DSS (Fristoe, 1979).

While DSS was originally designed to assess only those with standard American English dialect, Nelson (1976) created a version of DSS for speakers of Black non-mainstream English, and Toronto (1976) created a version of DSS for Spanish speakers. Some experts have criticized these adaptations. Vaughn-Cooke (1983) stated that DSS and attempted adaptations are inappropriate for any individuals other than those who speak standard American English. Vaughn-Cooke criticized Nelson's adaptation, explaining that the user must choose to either award a comparative point value for features accepted by Black, non-mainstream dialect, or withhold the point value so that the child is not credited for features that are similar to premature forms of standard English. Vaughn-Cooke reported that Nelson's attempt to modify the test fails to treat other dialects with equality. Thus while the adaptations allow wider use of DSS, their validity has been questioned.

Liles and Watt (1984) found that DSS was useful in discriminating between normally developing children and those with language impairment. Twelve language impaired boys were matched according to mean length of utterance (MLU) scores to 12

normally developing boys. Although DSS scores did not significantly differ between the two groups, performance in four of the eight grammatical categories did vary. When compared to MLU matched peers, mean scores of the language impaired children were significantly higher in secondary verbs and conjunctions, and significantly lower in *wh*-questions and primary verbs. In addition, the language impaired group received significantly fewer sentence points compared to the MLU matched peers. Liles and Watt concluded that the difference in performance in various categories demonstrates that DSS can assist clinicians in discriminating between normally developing and language impaired individuals.

Hughes et al. (1992) corroborated Liles and Watt's finding when they found that DSS scores confirmed the classification of 30 of 31 children who had previously been classified as language impaired through observation and non-standardized analysis of language samples; only one DSS score incorrectly categorized a child as having no language impairment. In addition, analysis of DSS results not only assists clinicians in identifying children with language disorders, but also provides information that clinicians can use in selecting treatment goals and measuring progress over time (Hughes et al., 1992).

Although DSS provides useful information for assessment and planning treatment, scoring language samples using DSS often takes a considerable amount of time and practice in order to become competent (Lively, 1984). Lively (1984) described errors that DSS learners frequently make when scoring DSS language samples. For example, because some children connect many utterances in their dialogue with the conjunction *and,* Lee (1974) specified that only two independent clauses can be given

credit when connected by *and* in order to avoid overestimating a child's ability to use coordinating conjunctions. While mistakes are common among beginning DSS scorers, Lively noted that accuracy in scoring DSS often increases dramatically when practiced. Lively recommended that DSS learners obtain assistance from experienced DSS scorers, so error patterns can be identified and corrected.

While DSS may be difficult to score initially, software is available to assist clinicians in improving accuracy. Hughes, Fey, Kertoy, and Nelson (1994) compared the accuracy of students learning to perform DSS when two groups of students were assigned to learn DSS by one of two methods. The first method was a traditional classroom-based method, which consisted of listening to lectures, practicing, and receiving feedback from an instructor. The second method consisted of using a computer assisted software program, which presented lessons and provided corrective feedback as well. Fifty-five speech-language pathology graduate students who had previously listened to a lecture and had been assigned to read a portion of Lee's (1974) text were randomly assigned to one of two groups. The students took a pretest, and following completion of five practice exercises, they took a post test. Hughes et al. (1994) found no significant differences between the groups, suggesting that automated DSS instruction may be efficient and beneficial. Ninety-three percent of students performed at or above an 80% accuracy level, suggested by Hughes et al. (1994) to be "acceptable and proficient" (p. 93).

Klee (1985) described some of the limitations of DSS. Klee cautioned that order of acquisition of grammatical forms in each of the eight DSS categories may not be in agreement with current research; thus, a clinician may use results of DSS to select treatment goals under the false premise that point values in DSS categories follow

developmental emergence. In addition, Klee urged clinicians to avoid strongly relying on a single score to replace careful analysis of the language sample because much information is lost when a language sample is summarized by a numeric score.

Hughes et al. (1992) encouraged the analysis of DSS results beyond the arrival of a DSS score as well. The researchers explained that clinicians must interpret DSS scores with caution because a single score can represent many different combinations of abilities in grammatical form categories. In addition, while normative data is available for comparison of a child's DSS score to same age peers, the child being assessed may be dissimilar to the normative group in terms of ethnicity, dialect spoken, socio-economic status, etc., making the norm-referenced score invalid (Hughes et al., 1992); DSS is thus more valuable as an assessment tool when it is used as an organizational tool for linguistic analysis instead of being used solely for its norm-referencing capabilities.

Automated DSS

Computerized DSS programs have been designed which could make DSS scoring more efficient and accurate (Channell, 2003); however, computerized DSS analysis continues to be infrequently used. Hux et al. (1993) found that only 3% of those surveyed use automated language analysis. Kemp and Klee (1997) found that while lack of time was reported as the most common reason for not performing language sample analysis, only 8% of those surveyed use automated language analysis. However, computers may be more available to clinicians than was the case a decade ago. Automated language analysis could make language sample analysis procedures, such as DSS, more efficient, and thus more feasible to use.

DSS Computer Program. Hixson (1983) was the first to use computer software for a partial DSS analysis. Klee and Sahlie (1986) evaluated Hixson's DSS Computer Program, DSSCP, and explained its goal, stating, "The objective of this program is to automate tallying and computing the points assigned to sentence constituents, and thus reduce the amount of time required to conduct a language sample analysis" (p. 232). In order to assign points to grammatical categories, DSSCP matches dictionary items with items from the child's language sample. Some grammatical items are required to be entered using a specific format in order to score the item; therefore, the user must know which items require specific formatting, and which items are readily recognized by the program. The authors stated that learning to use DSSCP could be accomplished in several hours if an individual was already efficient and knowledgeable in DSS. Published information concerning the accuracy of DSSCP does not exist (Channell, 2003).

Child Language Analysis. Child Language Analysis (CLAN; MacWhinney 1991, 1996, 2006) performs a variety of searching features, and also includes programs for analyzing aspects of language, including DSS, as well as Mean Length Utterance (MLU; Brown, 1973), and type token ratio (TTR; Templin, 1957). Language samples must be entered using a standard format called Codes for Human Analysis of Transcripts (CHAT; MacWhinney, 1996). After the language sample is entered using CHAT format, it must undergo morphological analysis performed by the "MOR" program (MacWhinney, 2006). Following morphological analysis, the language sample is coded for parts of speech using the "POST" program before DSS analysis can be performed. The clinician can choose between an automatic or an interactive mode for DSS analysis. A table, which lists each utterance and the point values assigned in each of the eight grammatical forms

categories, is displayed for both the automatic and interactive modes of the DSS analysis; however, when the automatic mode is chosen, the computer does not calculate a total DSS score. When the interactive mode is used, the clinician is required to award sentence points according to Lee's (1974) rules (MacWhinney, 2006). Attempt marks of incorrect, un-scored grammatical features may be noted as well. Scored grammatical forms and incorrect attempts are depicted in a table along with the total DSS score (MacWhinney, 2006).

MacWhinney (2006) noted limitations of automated DSS analysis using CLAN, stating that the program was unable to analyze the following three grammatical forms:

1. The pronominal use of "one" (e.g. "One should eat when one is hungry."), scored as p7.
2. Distinction between non-complementing infinitive forms (e.g. "I learned to dance"), scored as s3, and infinitival complement forms (e.g. "I have to listen."), scored as s5.
3. Wh- questions that contain an embedded clause but do not contain a conjunction (e.g. "Where is the shoe you lost?") as opposed to Wh- questions that do contain a conjunction (e.g. "Where is the shoe that you lost?").

At this time, CLAN does not offer fully automated DSS analysis. Because sentence points are awarded based on pragmatic and semantic accuracy, in addition to grammatical accuracy, they must be awarded manually. No data have been published regarding the accuracy of the CLAN DSS analysis software.

Computerized Profiling. Computerized Profiling (CP; Long, 1986; Long & Fey, 1993; Long et al., 2000) is another program that performs automated DSS analysis, as

well as analysis of other linguistic domains including prosody, lexical semantics, and phonology. Klee & Sahlie (1987) described one of the goals of CP, stating, "One of the author's objectives in developing this program was to encourage greater clinical use of language sample analysis by alleviating some of the time demands required for the task" (pp. 87-88).

To use CP, files must first be formatted according to rules similar to those used in Systematic Analysis of Language Transcripts (SALT; Miller and Chapman, 2000). The user must then select how the program should handle stereotypical utterances. Next, utterances containing *'s* are presented along with the program's attempt at classifying the grammatical feature as either marking possessive or a contraction; the user must indicate agreement or rejection of the classification (Channell, 2003). The utterances, along with the information supplied by the user, are saved as a "CORPUS" file before being run through Language Assessment Remediation Screening Profile (LARSP; Crystal, 1982; Crystal, Garman, & Fletcher, 1989). LARSP performs complex grammatical analysis which provides information that serves as a basis for DSS analysis (Channell, 2003). CP allows for manual correction of errors made in the automated analysis. In fact, Klee and Sahlie (1987) affirmed that those who use CP must be familiar with the linguistic analysis procedures that are performed because users must be able to recognize and correct errors that the program makes. Long's CP software (1986) comes with a 200 page manual which explains how to use the software program with the expectation that the user has a prior understanding of the grammatical analysis procedures, such as LARSP and DSS (Klee & Sahlie, 1987).

Long's 1986 original CP program was found to be lacking in both efficiency and

accuracy (Klee & Sahlie, 1987). Klee and Sahlie (1987) stated that correcting computer generated LARSP output errors took more time than manual scoring. However, the authors acknowledged that the program assists clinicians in efficiently locating specific grammatical features (e.g. SVO clause structures), which could save clinicians time when analyzing language sample in further detail once a general language analysis had been performed.

Another limitation of Long's 1986 program was that it was unable to analyze samples containing more than 125 utterances in a single CORPUS file; however, the program had a feature allowing for combining of CORPUS files (Klee & Sahlie, 1987).

Improvements in CP have been made over the years. Long and Fey's (1993) CP system allows for analysis of 1,000 sentences (Gregg & Andrews, 1995). In addition, a new module was added for analyzing a child's assertiveness and responsiveness in conversation called Conversational Acts Profile (CAP; Fey 1986). Further adaptations include a feature that allows for more efficient manual editing of the LARSP output, as well as help files which provide some information about scoring LARSP and DSS (Gregg & Andrews, 1995).

Gregg and Andrews (1995) stated that accuracy of automated DSS increases when the automated LARSP output is manually edited; however, the possibility exists of running DSS without manually correcting the output. Error patterns, such as including minor responses as sentences, commonly occur when LARSP is not manually edited. Not only should the LARSP output be examined before running the automated DSS program, but DSS output should be examined as well. Gregg and Andrews explained that the automated DSS output must be examined even if the LARSP output was manually edited

(Gregg & Andrews, 1995). Long and Fey's (1993) version of CP was predicted to be more efficient than manual scoring once the user becomes familiar with the program (Gregg & Andrews, 1995).

In 1999, a program by Channell and Johnson called GramCats became a part of CP. The GramCats program uses calculated probabilities to classify ("tag") words into grammatical categories (Channell & Johnson, 1999). These grammatical categories include copula, conjunctions, determiners, adjectives, nouns, pronouns, adverbs, verbs, and auxiliaries. The program contains information about the grammatical tag option(s) associated with each word and information regarding the probability of various tag sequences. For example, a noun is more likely than a verb to follow a particle. GramCats does not grammatically classify word combinations, which Crystal et al. (1989) referred to as analysis at the clause level and sentence level (Long & Channell, 2001).

Channell and Johnson (1999) tested the accuracy of their GramCats program. Automated grammatical tagging was compared to manual tagging in conversational language samples of 30 normally developing children age 2;6 to 7;11. Each language sample contained approximately 200 utterances. GramCats' average accuracy for tagging individual words was 95.1%. A significant inverse relationship existed between the child's age and the tagging accuracy of GramCats; thus, samples from younger children were tagged with higher accuracy than those from older children. The authors noted that accuracy decreased when the program attempted to tag words that introduced a subordinate clause, or when the program was required to distinguish between either the auxiliary or main verb forms of the words *be*, *have*, or *get*. While GramCats' word by word analysis reached a high accuracy level for tagging individual words, the accuracy of

tagging whole utterances was lower, averaging 78%. The authors suggested that improvements be made to increase the accuracy of GramCats before the program is used without reviewing and manually correcting the program's output errors.

Long (2001) conducted a study to test the efficiency and accuracy of computerized verses manual analysis of language and phonology using the CP software. Participants included 256 students and practicing clinicians from the United States and Australia. The students and clinicians were asked to participate only in analysis procedures for which they had previously received university-level training; one such procedure included DSS. Manual and computerized DSS was performed on two language samples. Performing manual DSS on the two language samples took clinicians an average of 56.2 minutes and 75.0 minutes. On average, manual DSS took approximately twice as long as computerized DSS. Although numerical results comparing accuracy of manual to computerized DSS were not reported, overall accuracy of computerized grammatical analysis procedures were reported to be higher than overall accuracy of manual grammatical analysis procedures (Long, 2001).

Long and Channell (2001) tested the accuracy of four automated CP language analysis procedures, including DSS, MLU, LARSP, and Index of Productive Syntax (IPSyn; Scarborough, 1990). Sixty-nine language samples from typically developing and language impaired children age 2;6 to 7;10 were used for the study. Several different dialects were represented as the language samples came from four groups of children from the United States, Canada, and Australia. To test the accuracy, fully automated outputs were compared to manually corrected outputs. Long & Channell found that the fully automated and manually corrected outputs of DSS were highly correlated (r = .92).

The percentage accuracy for automated DSS across the four groups of children was 89.8%. Similar to comparing interrater reliability between two clinicians, Long & Channell suggested that when comparing agreement between automated computer software to manual coding, the software could be considered acceptable if its correlation to manual coding is greater than .85. Agreement levels greater than .90 would be considered good, while agreement levels greater than .95 would be considered excellent. Using these criteria, the researchers classified DSS and IPSyn as good, LARSP as acceptable, and MLU as excellent.

Agreement between manual coding and fully automated coding for DSS was found to be similar to agreement between clinicians doing manual coding only, suggesting that fully automated DSS may provide results that would be considered reliable (Long & Channell, 2001). However, the authors suggest that DSS should continue to be reviewed. Long and Channell tested whether or not using fully automated DSS would change classification of a child as either normally developing or language impaired. Using the 10^{th} percentile as the cutoff, 7 of 56 children would be incorrectly classified if fully automated results were used. Because inaccuracy of automated DSS output affects clinical decisions at times, the authors recommended that users review automated outputs.

Channell (2003) examined the accuracy of fully automated DSS using CP under difficult circumstances, using language samples from older children with a majority of the children being classified as language impaired. A total of 9,084 utterances meeting the DSS criteria of containing a subject + predicate were used. The language samples used had been previously gathered from former studies, one by Fujiki, Brinton, and

Sonnenberg (1990), and one by Collingridge (1998). Utterances from a total of 48 children age 5;6 to 11;2 living in either the Reno, Nevada area or in the Jordan School District area of Salt Lake County, Utah were used in the study. To test accuracy, Channell compared manual DSS analysis, which was considered to be 100% accurate, to CP's automated DSS analysis. Channell (2003) found the average accuracy of CP to be 78%. Accuracy in grammatical categories ranged from 0% to 98%. The program met the 85% automated accuracy level suggested as acceptable (Long & Channell, 2001) in 11 of the 36 DSS cells. The overall accuracy was just below the 80% level suggested as "acceptable" by Hughes et al. (1994); therefore, Channell (2003) suggested that users continue to edit results following automated DSS analysis.

Point Difference. Sagae, Lavie, and MacWhinney (2005) conducted a study evaluating a software program called Grammatic Relations (GR), which automatically performs Index of Productive Syntax (IPSyn; Scarborough, 1990). IPSyn is a method for analyzing children's syntactic complexity according to the presence of 56 linguistic structures in a 100 utterance language sample (Scarborough, 1990). The GR program was compared to CP (Long, Fey, & Channell, 2004) which not only performs automated DSS, but performs IPSyn as well. Comparison of manually computed to automatically computed IPSyn scores of 23 transcribed language samples from children between the ages of 2;0 and 3;0, and 18 language samples from children between the ages of 8;0 and 9;0 was made. Accuracy of the program was analyzed in two ways. The first was by finding point difference, defined as the "(unsigned) difference between scores obtained manually and automatically" (Sagae et al., 2005, p. 201). The second was by finding *Point-to-Point Accuracy*, which is calculated by dividing the number of decisions that the

automated program made which complied with manual analysis by the total number of decisions. GR was found to be more accurate than CP in performing IPSyn. GR's average point-to-point accuracy was 92.8% while the average point difference was 3.3.

While several automated DSS analysis programs exist at this time, there continues to be room for improved accuracy. In addition, the assessment of accuracy in terms of point difference (Sagae et al., 2005) has not yet been applied to automated DSS programs. The purpose of this study was to test the accuracy of a recently released automated DSS program, DSSA, in terms of point to point agreement as well as in terms of point difference in order to find if accuracy of DSSA is sufficiently high to replace manual scoring.

Method

Participants

This study used a total of 17,990 utterances from five corpora of language samples; one corpus consisted of multiple language samples taken from a single child, and four corpora consisted of language samples each taken from different children. A total of 118 language samples taken from 99 children were used. A total of 98 language samples obtained from 98 children between the ages of 3 and 11 years comprised the between-child corpora. Participants included children classified as typically developing as well as those classified as language impaired. The between-child language samples had been previously obtained from four groups. The single child corpus included 20 languages samples collected periodically from a typically developing child. Although the language samples used in this study were originally collected for different purposes, the language samples were considered adequate for the current study based on the sound language sampling procedures employed and the presence of a broad range of language abilities represented in the samples.

Reno subjects. Thirty of the language samples were collected by Fujiki et al. (1990) for their study of conversational repair from children attending a school in Reno, Nevada. Fujiki et al. divided the children into 3 subgroups. The first subgroup (Reno-LI) consisted of 10 children with language impairments ranging in age from 7;6 (years; months) to 11;1 (M = 9;1). These children scored at or below 1 *SD* on two formal receptive language tests and on two expressive language tests. Receptive language tests used included the Peabody Picture Vocabulary Test–Revised (Dunn & Dunn, 1981), the Test for Auditory Comprehension of Language–Revised (Carrow-Woolfolk, 1985), the

Test of Language Development–Primary (Newcomer & Hammill, 1982), and the processing subtest of the Clinical Evaluation of Language Functions Screening Test (Semel & Wiig, 1980). The expressive language tests used include production subtests of the Clinical Evaluation of Language Functions Screening Test (Semel & Wiig, 1980), production subtests of the Clinical Evaluation of Language Functions–Diagnostic Battery (Semel-Mintz & Wiig, 1982), and the oral vocabulary and sentence imitation subtests of the Test of Language Development–Primary (Newcomer & Hammill, 1982).

In Fujiki et al.'s (1990) study, the first subgroup was matched to a second subgroup (Reno-LA) of typically developing children who had similar language abilities as the first. This second subgroup included 10 children ranging in age from 5;6 to 8;4 (M = 6;9). None of the children belonging to the second subgroup had a history of hearing impairment, speech or language impairment, mental retardation, behavioral problems, or academic problems.

The third subgroup (Reno-CA) included 10 typically developing children who were similar in age to the first subgroup. This third subgroup met the same set of criteria as the second subgroup. Ages ranged from 7;6 to 11;2 (M = 9;0).

Jordan subjects. Eighteen of the language samples were collected by Collingridge (1998) from Jordan School District in Salt Lake County, Utah to complete a pilot study comparing DSS scores taken by a clinician in on-line transcription to scores the clinician later calculated in reviewing the samples. Language samples came from children ages 6;2 to 10;9 who were language impaired. Three of the 21 samples originally collected were inadequate for DSS analysis as they contained fewer than 50 complete sentences, defined as containing a subject and a verb.

Wasatch subjects. Twenty of the language samples were collected in the Wasatch School District located in Utah for a study that compared on-line transcription to later transcription (Nichols, 2002). Twenty children, each diagnosed as language impaired, ranged in age from 3 to 9 years. All of the participants spoke English as their primary language, and had normal hearing, as verified by a hearing screening performed earlier in the year. Nine of the participants attended a self-contained language preschool while the remaining 11 children received language services outside of their classroom provided by their elementary school.

Provo subjects. Thirty of the language samples were collected by three graduate students for various research reasons from 30 children ranging in age from 2;6 to 7;11 who lived in Provo, Utah. The language samples were used in studies by Channell and Johnson (1999) and by Seal (2001). The children were typically developing, spoke English as their primary language, and passed a pure-tone, bilateral hearing screening at 15 dB HL.

Adam. Twenty language samples were collected from a typically developing single subject residing in Boston, Massachusetts, referred to as "Adam". These samples were originally collected as part of a longitudinal study of preschool children's language development (Brown, 1973). Language samples were collected periodically when the child was between the ages of 3;8 and 5;2 (Brown, 1973). At least 100 consecutive utterances from each sample were DSS coded for use in the present study.

Procedure

DSS was performed on each of the 118 language samples both manually and by using the DSSA software to compare results and find the accuracy of the DSSA software.

Manual DSS scoring. Procedures established by Lee (1974) were previously used to manually analyze utterances in 94 of the 118 samples. Twenty-four language samples were manually scored at the time of this study. The procedure for using DSS required the clinician to gather a language sample before scoring at least 50 utterances according to presence of grammatical forms. Words or phrases were classified into one of eight categories of grammatical forms, which include (a) indefinite pronouns or noun modifiers, (b) personal pronouns, (c) main verbs, (d) secondary verbs, (e) negatives, (f) conjunctions, (g) interrogative reversals in questions, and (h) *wh*- questions.

These grammatical forms were then awarded points based on difficulty level, with forms that typically develop at an older age receiving more points than forms that typically develop at a younger age (Lee, 1974). For example, the “main verbs” category contains grammatical forms in six of the eight difficulty levels, including m1, m2, m4, m6, m7, and m8. An additional "sentence" point was awarded to each sentence that complied with all adult grammatical rules. The sum of the points was divided by the number of utterances, resulting in the DSS score, which represents the average grammatical complexity of spoken sentences (Lee, 1974).

The inter-rater reliability of manual analysis was found when two speech-language pathologists scored 10% of the samples. Inter-rater reliability was found to be high, ranging from 95% to 97%.

DSSA scoring. Before running the DSSA software program, language samples were entered according to the following rules:

1. Type only one utterance per line, each ending with a period, question mark, exclamation mark (for imperatives), or comma.

2. Put all words in lower case except proper nouns and the pronoun *I.*.
3. Put mazes, repetitions, interjections, and anything else not to be analyzed in parentheses.
4. To skip the analysis of a whole utterance, put a non-alphanumeric character at the beginning of the utterance.

Data Analysis

Accuracy was inferred from percent agreement between DSSA, and manual coding, with manual coding used as the standard of correctness. To calculate DSSA's accuracy, DSSA coding results were compared to manual coding results using a utility program.

The number of agreements and disagreements between manual coding and DSSA coding as well as accuracy of DSSA for each grammatical form was found when all language samples from the between-child corpora and Adam corpus were evaluated. Disagreements consisted of misses or intrusions. Misses were defined as items that were scored manually but did not receive a score from DSSA. Intrusions consisted of items that were not scored manually but received a score from DSSA. Mean accuracy for the between-child corpora as well as mean accuracy for each of the five groups was reported. The mean point difference was found as well by averaging the differences in absolute values between the automated scores and the manual scores for each sample.

Results

The overall accuracy of DSSA for the between-child corpora was found to be 85.99% (*SD* = 5.05). Accuracy results for the between-child groups are displayed in Table 1. As Table 1 illustrates, the mean DSSA accuracy scores of groups comprised of children with language impairments were generally lower (around 84%) while the mean DSSA accuracy scores of groups comprised of normally developing children were generally higher (around 88%).

A one-way analysis of variance revealed between-group differences in accuracy scores to be statistically significant, $F(5,92) = 10.73$, $p < .01$. A correlation between accuracy levels and manual DSS scores was moderately high, $r = .59$, and statistically significant, $p < .01$, indicating that the automated DSS analysis of samples from children with higher DSS scores tended to be more accurate. When language impaired corpora were removed, a statistically significant correlation remained, $r = .60$, $p < .01$. Similarly, accuracy tended to increase in samples with higher DSS scores from language impaired children as well. A statistically significant correlation was found between accuracy levels and manual DSS scores, $r = .39$, $p < .01$. Post-hoc comparisons using the Bonferroni procedure (at an alpha level of .05) revealed the Reno-CA group accuracy to be significantly higher than all groups except the Reno-LA group, and the Reno-LA group to be significantly higher than the remaining groups.

Mean accuracy of the Adam corpus was found to be lower (*M* = 82.70%, *SD* = 3.67) than mean accuracy for all between-child corpora. When per cell agreement of grammatical forms of the Adam corpus was compared to the between-child corpora,

Table 1

Mean Accuracy and Standard Deviation of DSSA Software

Group	*M*	*SD*
Reno-LI	84.80	6.60
Reno-LA	89.70	1.70
Reno-CA	93.10	0.74
Jordan	83.61	4.05
Wasatch	83.25	4.28
Provo	86.03	4.20

it was observed that Adam samples scored more than 10 percentage points lower in accuracy in the following grammatical categories: i4, p7, m2, m6, m7, s4, and n7.

The number of agreements, misses, intrusions, and accuracy of DSSA for each value in each grammatical form category is displayed in Table 2, along with examples of grammatical forms. These values were compiled from the between child corpora and the Adam corpus. The grammatical form, p3, in the personal pronoun category was scored most accurately, 99%, followed by p1 of the personal pronoun category and i1 of the indefinite pronoun, noun modifier category, 98%. The Interrogative reversal category contained least accurate scores, with r8 at 0%, and r1 at 5%. Earlier developing grammatical forms were generally scored more accurately by DSSA compared to later developing grammatical forms; however, the negatives category did not fit this trend. In addition, because scores in the secondary verbs and interrogative reversals categories were extremely low, a trend could not be detected.

The overall mean point difference of the between-child corpora was found to be .74 (SD = .30). The mean point differences of the between-child groups are displayed in Table 3. The mean point difference for the Adam corpus was .91 (SD = .36).

Table 2

Number of Agreements, Misses, Intrusions, and Accuracy of DSSA for Each DSS Cell

Cell	Agree	Miss	Intrude	Total	%	Examples
Indefinite pronouns, noun modifiers						
i1	7,340	56	109	7,505	98	*it, this, that*
i3	3,261	47	183	3491	93	*all, someone, lot(s), more*
i4	35	0	9	44	80	*nothing, nobody, none, no one*
i7	422	30	44	496	85	*few, many, anyone, each, everything*
Personal pronouns						
p1	8,798	78	136	9,012	98	*I, me, my, mine, you, your(s)*
p2	4,917	51	77	5,045	97	*he, him, his, she, her(s)*
p3	3,749	9	38	3,796	99	*we, us, our, they, them, these, those*
p5	33	4	0	37	89	*myself, herself, itself, themselves*
p6	598	51	79	728	82	*who, what, which, that, how many*
p7	27	7	4	38	71	*one, oneself, whoever, whatever*
Main verbs						
m1	9,775	261	1,817	11,853	82	She *is dancing.* We *sing.* It *is* red.
m2	6,995	306	938	8,239	85	I *am* tall. We *ate.* She *listened.*
m4	1,865	276	36	2,177	86	I *don't know*. I *can skip.* It *will fly.*
m6	680	204	29	913	74	I *did help*. I *should go.* I *might do* it.
m7	395	143	88	626	63	I *have eaten.* I *got scared*.
m8	19	37	3	59	32	She *could be pretending*.
Secondary verbs						
s2	830	375	6	1,211	69	I want *to read.* Let's go *play.*
s3	33	157	8	198	17	It is fun *to watch.* We came *to help.*
s4	189	39	528	756	25	It was the boy *wearing* the shirt.
s5	1,161	197	368	1,726	67	I want you *to read.* I like *to color.*
s7	9	14	1	24	38	I have *to get going*. I like *to be held.*
s8	112	93	26	231	48	*Swimming* is scary. I started crying.
Negatives						
n1	80	22	1	103	78	This is *not* fun. It's *not* a dent.
n4	778	80	20	878	89	I *can't* sing. I *don't like* icecream.
n5	93	9	9	111	84	She *isn't* nice. I *won't* listen.
n7	512	6	122	640	80	They *aren't* going. He *hasn't* called.
Conjunctions						
c3	2,309	69	30	2,408	96	*and*
c5	757	21	59	837	90	*so, so that, and so, or, if*
c6	223	8	3	234	95	*because*
c8	1,285	92	142	1,519	85	He is *like* me. I'm bigger *than* you.

Interrogative reversals						
r1	25	466	7	498	5	*Is* it yours? *Is* she nice?
r4	9	67	0	76	12	*Is* he hiding? *Were* you waiting?
r6	318	356	7	681	47	*Can* I come? He spilled, *didn't* he?
r8	0	16	0	16	0	*Have* we met? *Has* he *been playing?*
Wh- questions						
w2	663	39	11	713	93	*who, what, where, how many*
w5	121	24	23	168	72	*when, how*
w7	97	21	1	119	82	*why, what if, how come, how about*
w8	13	1	7	21	62	*whose, which*

Table 3

Between-child Corpora Mean Point Difference of Developmental Sentence Scores

Group	*M*	*SD*
Reno-LI	.63	.28
Reno-LA	.78	.19
Reno-CA	.82	.12
Jordan	.71	.39
Wasatch	.68	.33
Provo	.80	.30

Note. Point difference is the absolute value of the difference between manually and automatically computed scores (Sagae et al., 2005).

Discussion

The overall accuracy of DSSA was moderately high. In addition, DSS scores obtained using DSSA differed on average by less than one point compared to scores obtained manually, indicating that existing DSS norms could likely be applied to DSSA. Although lower DSS scores and the presence of a language impairment were found to be correlated with lower accuracy of DSSA, the accuracy levels obtained should give clinicians confidence in using DSSA as a tool for language sample analysis.

When the between child corpus was examined, accuracy was found to be approximately 86%. A significant positive correlation was found between children's language scores and accuracy of DSSA, indicating that the program tended to more accurately score language samples of children who achieved higher DSS scores. This trend was present when both normally developing and language impaired groups were examined separately, although the trend was stronger in the normally developing group.

Yet at the same time, it was observed that earlier developing grammatical forms tended to be scored with more accuracy. This finding corresponds with previous research. Channell (2003) found that when language samples from older children, as well as children with language impairment were analyzed using CP's automated DSS program, the accuracy decreased when compared to a previous study by Long and Channell (2001), which used language samples from younger children. In addition, Channell and Johnson (1999) found that their GramCats program, a program for classifying words into parts of speech categories, was more accurate when analyzing samples from younger children.

Because GramCats serves as an underlying program for both DSSA and CP, one might predict that DSSA would also more accurately score younger children's language

samples; however, this does not appear to be the case, as increased DSS score was correlated with higher accuracy. Perhaps the higher DSS scores masked the underlying accuracy of grammatical form use in general, resulting in higher predictability for the DSSA software in scoring samples. Therefore, perhaps younger children who consistently use correct earlier developing grammatical forms would earn higher DSS scores compared to older children who use more complex DSS forms but are not credited because of low accuracy of DSSA in scoring these forms. Further research is needed to evaluate this aspect.

The present study corroborated previous findings that DSS scores from automated analyses were highly correlated with scores from manual analyses. The obtained *r* value was .98, slightly higher than the level (*M* = .92) presented in Long and Channell (2001) and the level (.97) obtained by Channell (2003).

Accuracy of the Adam corpus was found to be approximately 83%, while mean point difference was approximately .9. Reasons for low accuracy compared to the other corpora may include the fact that Adam was younger at the time his language samples were collected compared to children in the between-child corpora, thus his language was less complex. Because DSSA was found to more accurately score more complex language, one might reason that Adam's less complex linguistic performance contributed to lower accuracy results. In addition, the program may have been unable to correctly code or score frequent over-regularization errors made by Adam. For example, he often inserted "is" inappropriately (e.g. It's does not work). Although less likely, a final possibility may be that the lower accuracy scores were a reflection of the program's bias toward regional dialects. Perhaps the fact that language samples analyzed, with the

exception of Adam, came from children residing in either Utah or Nevada had an influencing effect on the accuracy.

Point difference, introduced by Sagae et al. (2005), was found to be fairly low, indicating that minimal difference existed between scores obtained automatically compared to scores obtained manually. While children's overall DSS scores typically range between 0 and 12, DSSA erred by less than one point in estimating the child's overall DSS score. Improving point difference by adding a correction value to the calculated DSSA score without overestimating a child's DSS score was found to be impossible. While DSSA generally underestimated children's DSS scores, it was found that the program overestimated 8 of the 118 children's scores. Although adding a correction value to the DSSA scores could improve overall mean point difference, sensitivity of DSSA in assisting to identify children with language impairment may decrease, resulting in failure to identify children who in fact had a language impairment. However, when using DSSA, clinicians should keep in mind that stipulations regarding use of DSS also apply to DSSA. Therefore, if clinicians use DSSA according to DSS guidelines established by Lee (1974), failure to detect a child with a language disorder should be limited because DSS was not intended to be used alone to detect the presence or absence of a language disorder.

While the overall accuracy of DSSA was moderately high, agreement in each grammatical form category varied. If 90% agreement between DSSA and manual analysis was considered high, 80% agreement was considered moderate, and at or below 70% agreement was considered low, then 9 of the 38 grammatical forms were scored with high accuracy, 16 were scored with moderate accuracy, and 13 were scored with low

accuracy. Accuracy in the secondary verbs and interrogative reversals categories were particularly low. Clinicians must use caution when interpreting results of grammatical form categories, particularly the categories that have been found to be especially problematic for the DSSA program.

When interpreting these results, limitations of this study should be considered. Limitations include the size and diversity of the sample population. Although this study did use subjects both with and without language disorders, the subjects, with the exception of Adam, resided in Utah or Nevada. In addition, multiple clinicians manually scored language samples. Although inter-rater reliability between two clinicians ranged between 95% and 97%, only one clinician remained constant for this comparison; therefore, inter-rater reliability may have been lower if further comparisons between clinicians were made. Finally, because DSSA is an automated version of DSS, one must keep in mind that limitations of DSS, which have been previously discussed, also apply to DSSA.

Continued improvement of DSSA could take one of two courses. The first approach would be to refine DSSA as a new language sample analysis program, grounded in DSS principles, but completely separate from DSS. Developing a separate program would take a great deal of time and research. For example the program would need its own set of norms, grading criteria, etc. The new program could attempt to improve aspects for which DSS has been criticized. For example, more grammatical forms such as adverbs could be analyzed. Further research would focus on testing psychometric qualities of this program independent from other programs.

The second approach would be to create a form of automated DSS analysis that more closely resembles manual DSS. Because many aspects of DSS have not been researched, there is much to learn about DSS as well as about how these aspects are affected by using an automated form of DSS. Information is limited concerning the psychometric aspects of DSS. One study (Gavin & Giles, 1996), although not addressing DSS directly, estimated the temporal reliability of language sample procedures as a function of sample size. Other than Koenigsknecht's work in the original Lee (1974) book, only one study has investigated the psychometric characteristics of DSS, including estimating reliability as a function of sample size (Johnson & Tomblin, 1975); therefore further research is needed. This research would lead to improved understanding of the nature of DSSA since it is patterned after DSS. Studies would focus on investigating how automated analysis alters aspects of manual analysis.

One may ask, under what circumstances is DSSA useful in clinical practice? Previous studies have found that automated DSS programs have offered improved efficiency compared to manual scoring (Long, 2001). Although efficiency of DSSA has not been analyzed at this time, improved efficiency compared to manual scoring is likely, as it can code DSS at over 100 sentences per second, a rate exceeding most if not all clinicians. Because lack of time was sited as the most common reason for not using language sample analysis (Hux et al., 1993; Kemp & Klee, 1997), DSSA may provide the efficiency needed for clinicians with heavy caseloads to perform language sample analysis.

In addition, as is true with automated IPSyn (Sagae et al., 2005), automated DSS analysis adds some degree of standardization to manual analysis. Using DSSA,

clinicians' results would coincide perfectly, with the exception that transcription differences may occur when clinicians are presented with the same language sample. Providing uniformity in DSS analysis allows clinicians to more confidently accept colleague's DSS analyses.

One aspect clinicians should consider when using DSSA for assessment purposes is the proximity of the child's score to the norm-referenced criteria for classification as disordered. For example, if the child scores near the cut-off score distinguishing between normally developing and disordered, the clinician should consider manually editing the automated analysis. Scoring within one point of the norm-referenced cut-off score may be a recommended guideline. However, if the child's score is well within the normal range or well within the disordered range, then manual editing would be less critical. While manual editing of automated scoring may be important for assessment purposes, comparing a child's score to normative data is unnecessary when evaluating progress over time. Automated analysis provides a method for efficient comparison of the child's performance over time.

The accuracy level found in this study should give clinicians confidence that if used appropriately, and if supplemented at times with manual editing, automated analysis can replace manual analysis. As suggested by Long and Channell (2001), one can treat the analysis of the program as that of any other colleague. Therefore, the accuracy of the program would be comparable to the inter-rater reliability between two clinicians. DSSA's accuracy level of approximately 86% would be considered "acceptable" according to criteria levels proposed by Hughes et al. (1994) and Long and Channell (2001), allowing DSSA to serve as a useful tool in clinical practice. However, when

using DSSA results to assist in selecting goals, clinicians should use caution, especially when considering grammatical forms which have reportedly low accuracy levels. Clinicians who take the time to understand the clinically relevant strengths and limitation of DSSA may find DSSA to be an effective method for language sample analysis.

While improvements could be made, DSSA continues to function as a useful tool. Although the program is less accurate at scoring language samples from individuals with language disorders compared to normally developing individuals, the 84% accuracy attained by DSSA in scoring language samples of individuals with language disorders is higher than that of CP software by Channell (2003) even though many of the samples (i.e. the Reno and Jordan sets) were the same. In addition, the approximately 86% accuracy level of DSSA is higher than accuracy levels obtained in the past by other DSS software programs. This study should strengthen clinicians' confidence in using DSSA as a clinical tool for language analysis.

References

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

Carrow-Woolfolk, E. (1985). *Test for auditory comprehension of language* (Rev. ed.). Allen, TX: DLM Teaching Resources.

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

Channell, R. W. (2006). DSSA (Version 1.0) [Computerized software] Provo, UT: Brigham Young University.

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

Collingridge, J. D. (1998). *Comparison of DSS Scores from On-Line and Subsequent Language Sample Transcriptions.* Unpublished master's thesis, Brigham Young University, Provo, UT.

Crystal, D. (1982). *Profiling Linguistic Disability.* London: Edward Arnold.

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

Dunn, L. M., & Dunn, L. M. (1981). *Peabody picture vocabulary test* (Rev. ed.). Circle Pines, MN: American Guidance Service.

Fey, M. E. (1986). *Language intervention with young children.* Boston: College-Hill Press.

Fristoe, M. (1979). Developmental sentence analysis. In F. L. Darley (Ed.), *Evaluation of appraisal techniques in speech and language pathology* (pp. 15-17). Reading, MA: Addison Wesley.

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.

Gallagher, T. M. (1983). Pre-Assessment: A procedure for accommodating language use variability. In T. M. Gallagher & C. A. Prutting (Eds.), *Pragmatic assessment and intervention issues in language* (pp. 1-15). San Diego, CA: Singular.

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

Gregg, E. M., & Andrews, V. (1995). Review of Computerized Profiling. *Child Language Teaching and Therapy, 11,* 209-216.

Hixson, P. (1983). *DSS Computer Program*. Omaha, NE: Computer Language Analysis.

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*, 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.

Hux, K., Morris-Friehe, M., & Sanger, D. D. (1993). Language sampling practices: A survey of nine states. *Language, Speech, and Hearing Services in Schools, 24,* 84-91.

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.

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,* 61-176.

Klee, T. (1985). Clinical language sampling: Analyzing the analyses. *Child Language Teaching and Therapy, 1*, 182-198.

Klee, T., & Sahlie, E. (1986). Review of DSS computer program. *Child Language Teaching and Therapy, 2*, 97-100.

Klee, T., & Sahlie, E. (1987). Review of Computerized Profiling. *Child Language Teaching and Therapy, 3*, 87-93.

Lee, L. L. (1974). *Developmental sentence analysis: A grammatical assessment procedure for speech and language clinicians.* Evanston, IL: Northwestern University Press.

Lee, L. L., & Canter, S. M. (1971). Developmental Sentence Scoring: A clinical procedure for estimating syntactic development in children's spontaneous speech. *Journal of Speech and Hearing Disorders, 36,* 315-340.

Liles, B. Z., & Watt, J. H. (1984). On the meaning of "language delay". *Folia Phoniatric, 36,* 40-48.

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

Long, S. H. (1986). Computerized Profiling (Version 1.0) [Computer software]. Arcata, CA: Author.

Long, S. H. (2001). About time: A comparison of computerized and manual procedures for grammatical and phonological analysis. *Clinical Linguistics & Phonetics, 15*, 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. (1993). Computerized Profiling (Version 7.0) [Computer software]. The Psychological Corporation.

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.

Long, S.H., Fey, M. E., & Channell, R. W. (2004). Computerized Profiling (CP) (Version 9.6.0) [Computer software]. Cleveland, OH: Department of Communication Sciences, Case Western Reserve University.

Long, S. H., & Masterson, J. J. (1993). Computer technology: Use in language analysis. *Asha*, *35*, 40-41, 51.

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

MacWhinney, B. (1996). The CHILDES System. *American Journal of Speech-Language Pathology, 5,* 5-14.

MacWhinney, B. (2006). *CLAN Manual*. Retrieved June 8, 2006, from http://childes.psy.cmu.edu/manuals/CLAN.pdf

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.

Muma, J. R. (1998). *Effective speech-language pathology: A cognitive socialization approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Nelson, N. (1976). Dialect differences in language samples gathered from Black preschoolers: Interview effects and measurement procedures. Paper presented at the American Speech and Hearing Association Convention, Houston, TX.

Newcomer, P. L., & Hammill, D. D. (1982). *Test of language development-primary.* Austin, TX: Pro-ed.

Nichols, C. B. (2002). *Live transcription of clinical child language samples.* Unpublished master's thesis, Brigham Young University, Provo, UT.

Sagae, K., Lavie, A., & MacWhinney, B. (2005). Automatic measurement of syntactic development in child language. Proceedings of the 43rd meeting of the Association for Computational Linguistics, Ann Arbor, Michigan, 197-204.

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

Seal, A. (2001). *Scoring sentences developmentally: An analog of developmental sentence scoring.* Unpublished master's thesis, Brigham Young University, Provo, UT.

Semel, E. M., & Wiig, E. H. (1980). *Clinical evaluation of language functions screening test.* Columbus, OH: Charles E. Merrill.

Semel-Mintz, E. M., & Wiig, E. H. (1982). *Clinical evaluation of language functions diagnostic battery.* Columbus, OH: Charles E. Merrill.

Stephens, I., Dallman, W., & Montgomery, A. (1988, November). Developmental sentence scoring through age nine. Paper presented at the annual convention of the American Speech-Language Hearing Association, Boston, MA.

Templin, M. C. (1957). Certain language skills in children: Their development and interrelationships. Minneapolis: The University of Minnesota Press.

Toronto, A. S. (1976). Developmental assessment of Spanish grammar. *Journal of Speech and Hearing Disorders, 41,* 150-171.

Vaughn-Cooke, F. B. (1983). Improving language assessment in minority children. *ASHA*, 25, 39-34.

Wilson, K. S., Blackmon, R. C., Hall, R. E., & Elcholtz, G. E. (1991). Methods of language assessment: A survey of California public school clinicians. *Language, Speech, & Hearing Services in Schools, 22,* 236-241.