2016-05-01

A Comparison of Manual and Automated Grammatical Precoding on the Accuracy of Automated Developmental Sentence Scoring

Sarah Elizabeth Janis
Brigham Young University - Provo

A Comparison of Manual and Automated Grammatical Precoding on the Accuracy of Automated Developmental Sentence Scoring

Sarah Elizabeth Bennett Janis

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

Master of Science

Ron W. Channell, Chair
Christopher Dromey
Shawn Nissen

Department of Communication Disorders

Brigham Young University

May 2016

ABSTRACT

A Comparison of Manual and Automated Grammatical Precoding on the Accuracy of Automated Developmental Sentence Scoring

Sarah Elizabeth Bennett Janis
Department of Communication Disorders, BYU
Master of Science

Developmental Sentence Scoring (DSS) is a standardized language sample analysis procedure that evaluates and scores a child's use of standard American-English grammatical rules within complete sentences. Automated DSS programs have the potential to increase the efficiency and reduce the amount of time required for DSS analysis. The present study examines the accuracy of one automated DSS software program, DSSA 2.0, compared to manual DSS scoring on previously collected language samples from 30 children between the ages of 2;5 and 7;11. Additionally, this study seeks to determine the source of error in the automated score by comparing DSSA 2.0 analysis given manually versus automatedly assigned grammatical tag input. The overall accuracy of DSSA 2.0 was 86%; the accuracy of individual grammatical category-point value scores varied greatly. No statistically significant difference was found between the two DSSA 2.0 input conditions (manual vs. automated tags) suggesting that the underlying grammatical tagging is not the primary source of error in DSSA 2.0 analysis.

Keywords: Developmental Sentence Scoring, automated language sample analysis, automated Developmental Sentence Scoring

ACKNOWLEDGEMENTS

I would like to thank Dr. Channell for the guidance and wisdom he provided throughout this project and for being a patient and caring mentor. I would like to thank my husband for the endless amounts of support and encouragement he provided and my parents for their constant support. I would also like to thank Melissa L. Barber, Lori Taylor Banta, and Elizabeth Chamberlain Mitchell for the use of language samples they collected.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF APPENDICES

Appendix Page

DESCRIPTION OF THESIS CONTENT AND STRUCTURE

This thesis, *A Comparison of Manual and Automated Grammatical Precoding on the Accuracy of Automated Developmental Sentence Scoring*, is part of a larger research project, and all or part of the data from this thesis may be published as part of articles listing the thesis author as a co-author. The thesis itself is to be submitted to a peer-reviewed journal in speech-language pathology. An annotated bibliography is presented in Appendix A.

Introduction

Language sample analyses are valuable tools for clinicians and researchers in speech language pathology; however, the time and training required for these analyses often leads to limited usage of this tool (Long, 2001). To save time, automated versions of widely used analyses such as Developmental Sentence Scoring (DSS; Lee, 1974) continue to be developed and improved, but require careful analysis of accuracy and efficiency. The present study reports on the evaluation of one recently developed clinical tool for automated DSS analysis (DSSA 2.0; Channell, 2015).

Language Sample Analysis

Language sample analysis (LSA) is a method of childhood language assessment that seeks to systematically assess, describe, and understand a child's expressive language abilities as observed in a naturalistic setting, most often in interactive conversation. Language sample analyses thus provide data that are more representative of the child's true linguistic abilities than the elicited language assessed in standardized tests. There are various procedures for conducting language sample analysis, including standardized and nonstandardized procedures, but almost all involve four steps: recording the conversation, transcription, analysis, and interpretation. The comprehensive information gained through language sample analysis is valuable for diagnosis of language disorders as well as treatment planning (Klee, 1985).

Due to its clinical usefulness, language sample analysis is widely used among speech language pathologists (SLP). Kemp and Klee (1997) conducted a survey of a representative sample of United States-based SLPs working with preschool children to assess clinical practices with regard to language sample analysis. Of the 253 respondents, 85% reported using language sample analysis despite only 8% reporting that language sample analysis is mandated by their

state. Respondents reported using language sample analysis for diagnosis, intervention, post-intervention, and/or screening of language disorders (Kemp & Klee, 1997).

A similar study conducted by Westerveld and Claessen (2014) surveyed SLPs from Australia to determine clinician practices and opinions regarding language samples, including the purpose of language sampling, elicitation methods, transcription, and analysis. Of the 257 respondents, 90.8% reported routine language sample collection and analysis for a variety of purposes. Among the 8.2% of respondents who reported not collecting language samples, time constraints, lack of training, and lack of computer hardware/software were reported as the primary reasons. Additionally, 87% of respondents reported always or often using an informal LSA procedure while only 37% reported always or often completing a detailed analysis of the language samples; time was reported as the primary barrier to more detailed sample analysis (Westerveld & Claessen, 2014). These findings were, overall, consistent with the findings from Kemp and Klee's (1997) study.

Despite the benefits of language sample analysis, there are disadvantages, including the amount of expertise required, the difficulty of eliciting a representative sample, and the lack of procedural uniformity for collection and elicitation. One of the greatest disadvantages of language sample analysis is the amount of time required (Hux, Morris-Friehe, & Sanger, 1993). Of the respondents to the Kemp and Klee (1997) survey who reported not using language sample analysis, 86% cited a lack of time, making this the most common reason not to use language sample analysis.

Computerized Language Sample Analysis

One option for decreasing the time required for language analysis procedures is computerized analysis. Long (2001) compared the length of time required to complete

phonological and grammatical analyses manually versus by computer. Long found that computerized analysis significantly reduced the amount of time required to complete otherwise lengthy analyses. Computerized phonological analyses were at least 11 times faster than the manual analyses, with an average between 17 and 35 times faster. Computerized grammatical analyses were between one and five times faster than the comparative manual analyses (Long 2001), even given the time needed to correct the automated analysis. This study demonstrated that computerizing language analysis procedures could significantly reduce the amount of time required, thus making routine language analysis a more realistic clinical possibility for SLPs.

A probabilistic automated grammatical tagging program called GramCats (Channell 1998) was found to have a word-by-word accuracy of between 92.9% and 97.4% with a mean of 95.1% when compared with manual tagging, with manual tagging considered the gold standard, and a whole-utterance accuracy of between 60.5% and 90.3% with a mean of 77.7% (Channell & Johnson, 1999). Although the whole-utterance accuracy was lower than desirable, automated word-by-word accuracy compared to manual accuracy was almost as high as interjudge human reliability (Channell & Johnson, 1999). This study demonstrated that automated analyses have the potential to achieve levels of accuracy comparable to human-conducted manual analyses while significantly reducing the amount of time required to complete such analyses.

Developmental Sentence Scoring

Developmental Sentence Scoring has been one of the most commonly used standardized language sample analysis procedures since it was developed by Laura Lee in the early 1970s. The DSS procedure aims to evaluate and score a child's use of standard American-English grammatical rules within complete sentences (Lee, 1974). A spontaneous language sample consisting of at least fifty sentences is obtained in a naturalistic setting. DSS examines this

sample to scale a child's grammatical development in 8 areas: (a) indefinite pronoun or noun modifier, (b) personal pronoun, (c) main verb, (d) secondary verb, (e) negative, (f) conjunction, (g) interrogative reversal in questions, and (h) wh- questions. Within each of these eight grammatical areas, the presence of one or more particular grammatical structures in a sentence earns a point value score ranging from one to eight points, with higher point scores being awarded to more developmentally advanced forms. For example, the use of a second-person singular pronoun such as *you* earns two points in the personal pronoun area. A sentence may have scores in several grammatical areas; for example, the sentence *where is she going?* would score two points in the wh- questions area, one point in the main verb area, two points in the personal pronoun area, and four points in the interrogative reversal category, for a total of nine points. The person doing the DSS analysis could then summarize these point scores as "*w2*, *m1*, *p2*, *r4*" for this sentence.

Together with the category-point value scores, an additional point, called the *sentence point*, is awarded to a sentence that meets all adult standard rules; the sentence point is forfeited for any mistakes, including mistakes in features beyond the eight scored grammatical forms. The points awarded to each utterance are summed, and this total is divided by the number of sentences analyzed to produce the overall DSS score. Additionally, attempt marks are awarded for structures that are attempted but lack some feature or requirement of standard English; these marks do not affect the overall DSS score but do provide additional information regarding the child's current grammatical system and developmental gains that may not be revealed in the score (Lee, 1974).

One of the most distinctive features of DSS has been its normative data, which allow an individual child's score to be compared with typically developing chronological age peers (Lee,

1974). Data were collected from 200 white children age 2;0 to 6;11 with five males and five females in each three-month interval. Since the normative group is not representative of the general population of American children, it is only appropriate to reference the norms with white, middle-class, monolingual children who speak standard American English (Fristoe, 1979). Additionally, as norms older than 10 years are considered to be outdated (Salvia & Ysseldyke, 2010), the normative DSS scores should be viewed with caution. However, even without referencing normative data, the numerical DSS score can be used for comparisons across time or between children (Hughes, Fey, & Long, 1992).

The information provided by DSS is clinically useful in many ways. The procedure was not designed as a diagnostic tool and, according to the developer, is better suited for tracking progress and determining when to discharge a child from therapy (Lee, 1974). However, DSS has been found to discriminate between normally developing children and those with language impairment and can be useful in making diagnostic judgments when used in addition to other measures (Hughes et al., 1992). Lee (1974) notes that, while DSS does not assess all possible grammatical structures, it can direct the clinician's attention to specific language areas for further and continuing investigation. Additionally, analysis of attempt marks, point values, sentence point errors, and frequency of occurrence of forms for each category can inform goal selection and guide intervention (Fristoe, 1979, Hughes et al., 1992).

DSS Usage

For many years, DSS was a widely used clinical tool. Kemp and Klee's (1997) survey of SLP practices in preschool language assessment found that DSS was the most frequently used standardized language sample analysis procedure, with 35% of clinicians reporting use of DSS. A similar survey of SLPs from nine Midwestern states found that DSS was the only standardized

language analysis procedure used with regularity; in this study, 31% of the 239 respondents reported using DSS (Hux et al., 1993). Despite the continued usefulness of DSS (Hughes et al., 1992), however, clinical usage of DSS appears to have declined in recent years due to the outdated normative data and the amount of time required to complete the analysis.

Despite the decline in clinical usage, DSS is still a widely used tool in research. Smith, DeThorne, Logan, Channell, and Petrill (2014) evaluated the impact of prematurity on language skills using both discourse-level language and standardized tests, including DSS. In a study of the role of input in children's use of third person singular marker, Leonard, Fey, Deevy, and Bredin-Oja (2014) used DSS as one of the evaluations to determine children's placement in the specific language impairment group.

Automated DSS Analysis Software

According to Fristoe (1979), the greatest disadvantage of DSS is the amount of time required to complete the procedure. Automated DSS analysis software has the potential to reduce this disadvantage, thus making the routine use of DSS in both clinical and research settings more feasible. Several software programs for automated DSS analysis have been developed with varying levels of success.

One such program is Child Language Analysis (CLAN; MacWhinney, 2000). CLAN is able to compute grammatical analyses including Mean Length of Utterance (MLU), type token ratio (TTR), DSS, and Index of Productive Syntax (IPSyn). In order to complete DSS analysis using CLAN software, files must be in Codes for Human Analysis of Transcripts (CHAT) format and the user must run the sample through a morphological analysis program (MOR; MacWhinney, 2000) and code the sample for parts of speech using a part of speech tagging (POST) program. The DSS program has both an automatic and interactive mode (MacWhinney,

2000). However, no data exist at this time as to the accuracy of the CLAN automated DSS analysis.

A second program with the capability to conduct automated DSS analysis is Computerized Profiling (CP), initially developed by Steven H. Long in 1986. Disadvantages of the initial version included restrictions on maximum corpus size, word truncation, and misanalysis of multiple embedded clauses (Klee and Sahlie, 1987). Later versions of CP integrated a probabilistic automated grammatical tagging program, GramCats (Channell 1998), which helped reduce some of the disadvantages present in the early versions. Long and Channell (2001) completed an automated DSS analysis using the updated version of CP on 69 language samples collected from children of various ages, regional dialects, diagnostic categories, and levels of linguistic development. The results of the automated analysis were compared to manual analyses of the same samples to provide a percentage accuracy for the automated software. The CP DSS analysis was found to have an accuracy of 89.8% with a positive correlation between the accuracy of the calculation and the size of the DSS score (Long & Channell, 2001).

Channell (2003) conducted an in-depth analysis of the accuracy of CP's automated DSS analysis on 48 language samples collected from school-age children, 28 of whom had language impairment. The overall accuracy of automated DSS scoring compared to manual scoring was found to be 78.2%; the lower level of accuracy compared to the Long and Channell (2001) study was judged to be the result of the increased linguistic complexity of the samples used in this study. The automated and manual scores were highly correlated with the automated scores being significantly higher than the manual scores. Channell (2003) also examined the agreement at the category and point-levels by creating cells for each DSS grammatical form category and point level (see Appendix B); for example, Channell referred to a point value of one in the personal

pronoun category as cell *p1* and a point value of seven in the negatives category as *n7*. Channell then assessed the percentages of correct tagging, false negatives (misses), and false positives (intrusions) for each of the 38 cells. The level of agreement at the point-level ranged from 10% to above 90%, with lower point value cells generally having a higher coding accuracy. There was no clear pattern for misses or intrusions. Some cells had far more misses than intrusions while some had more intrusions than misses; most had roughly equal amounts of misses and intrusions. These studies demonstrate that automated DSS analysis software has the potential to reach clinically acceptable levels of accuracy. Hughes, Fey, Kertoy, and Nelson (1994) suggested 80% accuracy as an acceptable level for effective clinical use of DSS. The CP DSS exceeded this level in the Long and Channell (2001) study but was unable to uphold the same level of accuracy on the more developmentally complex samples assessed in the Channell (2003) study.

A third automated DSS program is a locally developed software program, called DSSA (Channell, 2006). The accuracy of the initial software version was examined using 118 language samples collected from 99 children between the ages of 3 and 11 (Judson, 2006). Participants included both typically developing and language impaired children. Developmental Sentence Scoring was conducted on each language sample both manually and with DSSA. Accuracy of the DSSA software was determined from percent agreement between DSSA and manual coding with the manual coding assumed to be accurate. The *Point Difference* method (Sagae, Lavie, & MacWhinney, 2005) was used to compare automated and manual scores; the mean point difference was determined by averaging the differences in absolute values between the automated and manual scores. The overall accuracy of DSSA for the between-child corpora was found to be 85.99% with a standard deviation of 5.05 and the accuracy of the single child corpora was found to be 82.70% with a standard deviation of 3.67 (Judson, 2006). DSSA was

less accurate with language samples with lower manual DSS scores and for children with language impairment; the mean for children with language impairment was around 84% compared to 88% accuracy for typically developing children. Judson (2006) provided data on the number of agreements, misses, intrusions, and accuracy of DSSA for each point value in each grammatical form category. Category *p3*, which includes the forms that earn a score of three in the person pronoun category, was the most accurate at 99% accuracy while the interrogative reversal category had the least accurate scores, with the point value of eight points (*r8*) at 0% and point value of one (*r1*) at 5%. Earlier developing forms were generally scored more accurately than later developing forms. However, the negatives category did not fit this trend and no trend could be determined in the secondary verbs and interrogative reversals categories. The overall mean point difference of the between-child corpora was found to be .74 (*SD* = .30) and the mean point difference for the single-child corpus was .91 (*SD* = .36; Judson, 2006).

The findings of Judson's (2006) study correspond with previous studies in terms of overall accuracy of automated DSS analysis. The overall accuracy was moderately high; however, accuracy in each grammatical form category varied.

Goals of the Current Study

The availability of more advanced software has allowed for modifications to DSSA, including improvements in the DSS extraction rules and the use of a more complex probability model. This study is part of a larger project to evaluate the overall and point-level cell accuracy of DSSA given improvements in the underlying software; accuracy was defined as the level of agreement between DSSA and manual scoring, with manual scoring assumed to be 100% accurate. Additionally, this study will examine the underlying cause of errors in DSSA scoring,

to try to determine whether the errors are the result of incorrect auto-tagging of grammatical categories or the result of errors in the rules for DSS extraction.

Method

Participants

The language samples used in the present study were previously collected by three speech language pathology graduate students in 1989 for various research purposes. Samples were collected from 30 children living in a BYU on-campus family housing complex in Provo, Utah. Children ranged in age from 2;5 to 7;11 (years;months) with three children in each 6-month interval from 2;5 to 6;11 and three children between 7;0 and 7;11. Each child was reported by his or her parents to be typically developing with no speech or language delay, spoke English as the primary language, and passed a bilateral hearing screening at 15 dB HL. Each of the three graduate students collected a conversational language sample for one child within every age interval. Samples were collected in the child's apartment with only the child and graduate student researcher present; a variety of props was used to elicit conversation. Each sample consists of at least 200 intelligible child utterances; any utterance with one or more unintelligible words was removed from the sample. All adult utterances were removed, to allow for uninterrupted analysis of child language.

These samples were also used by Channell and Johnson (1999) and in a graduate student study evaluating the beta version of the software used in the present study (DSSA 1.0; Judson, 2006), allowing for a direct comparison of the accuracy of past and present versions of the DSSA software.

Software

Two distinct software programs were used in this study, one to code the language samples according to grammatical category and one to extract DSS from grammatically tagged texts.

GramCats. The first software program was an updated version of the GramCats software as described and evaluated by Channell and Johnson (1999). The version used in the present study took information from two probability sources to determine and code the grammatical category of words in running text. The first probability source used relative tag likelihood to determine the probability of a word being used as a particular part of speech, independent of context. This information was pulled from a dictionary that contained a predetermined set of words, the grammatical tag options for each word, and the relative frequencies of each tag option. If only a single tag option existed for a word, the tag was assigned; words with multiple tag options were considered ambiguously tagged (Channell & Johnson, 1999).

The second probability source used a trigram Markov approach to determine the presence of a tag based on the probability of a particular tag coming after the preceding two tags, independent of the words that were assigned those grammatical tags. This aspect of the software was an improvement to the initial GramCats software, which only used data from one preceding tag.

When an ambiguous tag was encountered, the program noted the ambiguous tag and continued to move through the utterance until a word with only a single tag option or a punctuation mark was found. The sequence of ambiguously tagged words bound by an unambiguously tagged word or punctuation mark on each end was called a span. For each span, the transitional tag probabilities for each possible tag pair were obtained. The probability of each

option was the probability that a tag fits its word multiplied by the probability that a tag comes after the preceding tags in the span. The option whose probabilities multiplied out to the highest product was considered the most likely tag sequence for the span and the utterance was tagged accordingly (Channell & Johnson, 1999).

DSSA 2.0. The second software, DSSA 2.0 (Channell, 2015), examined a grammatically tagged text sample for specific patterns of words and grammatical tags. The software then matched these patterns to an internal set of DSS extraction rules to select a DSS category and point value cell. Before running DSSA 2.0, language samples were put in the format used by the Systematic Analysis of Language Transcripts (SALT; Miller, Andriacchi, & Nockerts, 2011) software.

Procedure

Developmental Sentence Scoring was performed on each of the 30 language samples three separate times, once manually and twice using the DSSA 2.0 software.

Prior to automated DSS analysis, the samples were individually coded for grammatical category information. The first round of DSSA analysis was completed following manual tagging of grammatical categories (A-DSS-M). The second round of DSSA analysis was completed following the automated tagging of grammatical categories with the GramCats software (A-DSS-A). Finally, each sample was previously manually analyzed (M-DSS) by the three graduate students who collected the data and awarded a numerical score according to the DSS scoring guidelines established in Lee (1974); interrater reliability was established by having a second clinician analyze 10% of the samples; the level of agreement was found to be 97% (Seal, 2002).

Results

The DSS score was computed three different ways for each of the 30 samples: (a) manual scoring (M-DSS), (b) automated DSS with manual coding for grammatical categories (A-DSS-M), and (c) automated DSS with automated coding for grammatical categories (A-DSS-A). Table 1 provides data for each participant, including age in months, the sample length, and DSS score for each method of calculation. For the majority of the participants (22/30), the DSS scores varied by less than 0.5 points across calculation methods. Participant 8's score was the most consistent, with a 0.02 point variation across calculation methods. Participant 6's score was the least consistent, with a 1.20 point variation.

The mean DSS score was found to be 8.43 (SD = 2.36) for the M-DSS group, 8.29 (SD = 2.33) for the A-DSS-M group, and 8.23 (SD = 2.23) for the A-DSS-A group. A univariate repeated measures ANOVA revealed a statistically significant difference among the DSS scores, $F(2,58) = 5.23$, $p < .01$. A post-hoc Tukey HSD test found that the manual scores (M-DSS) were significantly different than the fully automated DSS score (A-DSS-A) at $p < .01$ but did not differ significantly from the A-DSS-M scores. Additionally, there was no significant difference between the A-DSS-A and A-DSS-M scores. The overall accuracy of A-DSS-M was 86.78% and the overall accuracy of A-DSS-A was 86.32%.

Table 1

DSS Scores from Manual Analysis (M-DSS) and Automated Analysis Using Manually-Tagged (A-DSS-M) or Auto-Tagged (A-DSS-A) Input

ID	Age in Months	Number of Sentences	M-DSS	A-DSS-M	A-DSS-A
1	30	141	5.71	6.21	6.45
2	30	117	5.79	5.64	5.68
3	33	129	5.96	5.76	5.82
4	35	168	5.93	5.95	5.98
5	37	144	6.22	6.18	6.15
6	39	101	3.67	4.87	4.87
7	45	186	7.46	7.50	7.48
8	45	185	7.10	7.09	7.08
9	46	152	7.04	6.20	6.18
10	53	148	8.82	8.86	9.02
11	56	161	10.11	10.02	10.03
12	59	138	10.54	10.53	10.28
13	59	182	9.09	8.68	8.65
14	62	162	8.25	8.18	8.22
15	62	168	6.84	6.25	6.38
16	64	161	7.32	7.16	7.14
17	65	135	7.41	7.30	7.45
18	65	187	11.50	11.61	11.51
19	66	163	8.36	8.24	8.08
20	68	151	7.86	7.43	7.40
21	69	177	11.29	11.38	10.82
22	72	149	9.17	8.38	8.26
23	75	190	9.28	8.80	8.67
24	77	195	12.29	12.37	11.90
25	79	160	6.96	6.59	6.57
26	79	167	8.42	7.74	7.78
27	84	149	8.38	8.11	8.03
28	91	160	8.92	8.57	8.39
29	94	195	13.81	13.88	13.72
30	95	189	13.41	13.12	13.03

Table 2 and Table 3 provide data on the accuracy of each DSS cell for the two automated DSS calculations, A-DSS-M and A-DSS-A; a cell is defined as a single point value within a particular grammatical category, for example, a point value of one in the personal pronouns category is labeled as cell p1. Table 2 describes the A-DSS-M calculation method compared to scoring DSS manually (M-DSS). There was a wide range of cell accuracy, from 0% (r8) to 98.86% (p3). Cells with 90% or higher accuracy included the following point values: one and three in the indefinite pronouns column (*i1*, *i3*); one, two, and three in the personal pronouns column (p1, p2, p3); three and six in the conjunctions column (*c3*, *c6*); and two in the wh-questions column (w2). Cells with 80% to 89% accuracy included the following point values: four and seven in the indefinite pronouns column (*i4*, *i7*); five in the personal pronouns column (*p5*); one, two, and four in the main verbs column (m1, *m2*, *m4*); four and five in the negatives column (*n4*, *n5*); and five in the conjunctions column (*c5*). Cells with 70% to 79% accuracy included the following point values: six in the personal pronouns column (*p6*); six in the main verbs column (*m6*); five in the secondary verbs column (*s5*); one and seven in the negatives column (*n1*, n7); and one in the interrogative reversals column (r1). Other cells had accuracy levels under 70%, with the the point values of four in the secondary verbs column (*s4*) and six and eight in the interrogative reversals column (*r6*, r8) having accuracy levels below 20%. Generally, the lower point value cells within each grammatical category were coded with higher accuracy than the higher point value cells. The indefinite pronoun and personal pronoun categories were coded with the greatest accuracy while the interrogative reversal category was coded with the least accuracy.

Table 2

Cell Accuracy Levels of Automated DSS Based on Manually Tagged Input (A-DSS-M)

Cell	N	Agree	Miss	Intrusion	Percent Correct
i1	2064	2026	14	24	98.16
i3	999	945	15	39	94.59
i4	8	7	0	1	87.50
i7	143	128	3	12	89.51
p1	2887	2812	18	57	97.40
p2	911	885	16	10	97.15
p3	1320	1305	4	11	98.86
p5	13	11	2	0	84.62
p6	117	84	22	11	71.79
p7	15	9	5	1	60.00
m1	2996	2466	137	393	82.31
m2	2251	1936	77	238	86.01
m4	710	618	81	11	87.04
m6	272	209	57	6	76.84
m7	146	78	22	46	53.42
m8	28	9	18	1	32.14
s2	459	307	142	10	66.88
s3	62	14	46	2	22.58
s4	176	13	5	158	7.39
s5	502	355	30	117	70.72
s7	14	7	6	1	50.00
s8	80	36	39	5	45.00
n1	46	33	13	0	71.74
n4	272	240	24	8	88.24
n5	33	27	5	1	81.82
n7	232	184	2	46	79.31

c3	601	579	8	14	96.34
c5	271	234	5	32	86.35
c6	83	82	0	1	98.80
c8	330	152	147	31	46.06
r1	144	110	27	7	76.39
r4	28	4	24	0	14.29
r6	197	94	96	7	47.72
r8	7	0	7	0	0.00
w2	211	196	9	6	92.89
w5	73	44	1	28	60.27
w7	29	16	13	0	55.17
w8	3	2	0	1	66.67

Note. The cell codes refer to the intersection of the point values and grammatical category columns of the DSS Scoring Chart (see Appendix B). I = indefinite pronouns or noun modifiers, p = personal pronouns, m = main verbs, s = secondary verbs, n = negatives, c = conjunctions, r = interrogative reversals, w = wh-questions.

Table 3

Cell Accuracy Levels of Automated DSS Based on Automatedly Tagged Input (A-DSS-A)

Cell	N	Agree	Miss	Intrusion	Percent Correct
i1	2066	2025	15	26	98.02
i3	1029	948	12	69	92.13
i4	7	6	1	0	85.71
i7	142	119	12	11	83.80
p1	2885	2811	19	55	97.44
p2	911	885	16	10	97.15
p3	1320	1305	4	11	98.86
p5	13	11	2	0	84.62
p6	126	88	18	20	69.84
p7	15	9	5	1	60.00
m1	3083	2500	103	480	81.09
m2	2242	1898	115	229	84.66
m4	712	616	83	13	86.52
m6	273	209	57	7	76.56
m7	124	64	36	24	51.61
m8	28	8	19	1	28.57
s2	459	308	141	10	67.10
s3	62	13	47	2	20.97
s4	146	12	6	128	8.22
s5	507	352	33	122	69.43
s7	13	6	7	0	46.15
s8	80	30	45	5	37.50
n1	46	33	13	0	71.74
n4	272	241	23	8	88.60
n5	33	27	5	1	81.82
n7	231	184	2	45	79.65

c3	601	579	8	14	96.34
c5	252	235	4	13	93.25
c6	83	82	0	1	98.80
c8	348	161	138	49	46.26
r1	152	109	28	15	71.71
r4	28	3	25	0	10.71
r6	195	94	96	5	48.21
r8	7	0	7	0	0.00
w2	213	186	19	8	87.32
w5	56	44	1	11	78.57
w7	29	17	12	0	58.62
w8	2	2	0	0	100.00

Note. The cell codes refer to the intersection of the point values and grammatical category columns of the DSS Scoring Chart (see Appendix B). I = indefinite pronouns or noun modifiers, p = personal pronouns, m = main verbs, s = secondary verbs, n = negatives, c = conjunctions, r = interrogative reversals, w = wh-questions.

Table 2 also shows the disagreements between M-DSS and A-DSS-M scoring divided into misses (false negatives) and intrusions (false positives). Some cells, such as the point values of six in the personal pronouns column (p6), four and six in the main verbs column (m4, m6), two, three, and eight in the secondary verbs column (*s2*, *s3*, *s8*), four in the negatives column (n4), eight in the conjunctions column (*c8*), and one and six in the interrogative reversals column (r1, r6) had at least twice as many misses than intrusions while other point values, such as one in the indefinite pronouns column (i3), one in the personal pronouns column (p1), one, two, and seven in the main verbs column (m1, m2, *m7*), four and five in the secondary verbs column (s4, s5), seven in the negatives column (n7), and five in the conjunctions column (c5) had at least twice as many intrusions than misses. The remaining cells either had a fairly equal number of misses and intrusions or had too few errors to allow a pattern to be seen.

Table 3 provides data on the accuracy of each DSS cell for the A-DSS-A calculation method compared to manual scoring (M-DSS). Cell accuracy with this calculation method ranged from 0% (r8) to 100% (*w8*). The point value of three in the personal pronouns column (p3) had the second highest accuracy level at 98.86%, which was the same level of accuracy for this cell found in Table 2. Individual cell accuracy was highly correlated between Table 2 and Table 3. In comparison to Table 2, the majority of the cells in the 90% to 100% accuracy range remained the same (i1, i3, p1, p2, p3, c3, and c6); the point value of two in the wh-questions column (w2) was no longer in this range, and five in the conjunctions column (c5) and eight in the wh-questions column (w8) achieved greater than 90% accuracy in Table 3 but not Table 2. Cells in the 80% to 89% accuracy remained the same (four and seven in indefinite pronouns, five in personal pronouns, one, two, and four in main verbs, and four and five in negatives; i4, i7, p5, m1, m2, m4, n4, n5) with the exception of the removal of five in conjunctions (c5) and addition

of two in wh-questions (w2). The point values of six in the main verbs (m6) column, one and seven in the negatives column (n1, n7), and one in the interrogative reversals column (r1) remained in the 70% to 79% accuracy range along with five in the wh-questions column (*w5*); six in the personal pronouns column (p6) and five in the secondary verbs column (s5) dropped just below 70% accuracy. The remaining cells had accuracy levels under 70%, with the point value cells of four in secondary verbs (s4) and four and eight in interrogative reversals (r4, r8) having accuracy levels below 20%.

Of the 38 cells, seven varied by more than two percentage points in the Percent Correct column between Table 2 and Table 3 (three and four in indefinite pronouns, eight in main verbs, seven in secondary verbs, one and four in interrogative reversals, and seven in wh-questions; i3, i4, *m8*, *s7*, r1, r4, *w7*), three varied by more than five percentage points (one in indefinite pronouns, eight in secondary verbs, five in conjunctions, and two in wh-questions; i7, s8, c5, w2), and two varied by 10 percentage points (five and eight in wh-questions; w5 and w8). Of these cells, the point values of five, seven, and eight in the wh-questions column (w5, w7, w8) had higher values on Table 3 than Table 2; all others had higher values on Table 2. The remaining 26 cells varied by less than two percentage points between the two automated DSS calculation methods (Table 2 and Table 3).

Table 2 and Table 3 displayed similar patterns of misses and intrusions. Point value cells of four and six in the main verbs column (m4, m6), two, three, and eight in secondary verbs column (s2, s3, s8), four in the negatives column (n4), eight in the conjunctions column (c8), and six in the interrogative reversals column (r6) continued to have at least twice as many misses than intrusions, along with four in the interrogative reversals column (r4); six in the personal pronouns column (p6) and one in the interrogative reversals column (r1) no longer did. The point

value cells of three in indefinite pronouns (i3), one in the personal pronouns column (p1), one and two in the main verbs column (m1, m2), four and five in the secondary verbs column (s4, s5), and seven in the negatives column (n7) continued to have more intrusions than misses; seven in the main verbs column (m7) and five in the conjunctions column (c5) no longer displayed this pattern in Table 3.

Discussion

The focus of the present study was to compare DSS scores that were calculated manually (M-DSS) to DSS scores calculated automatedly using both manually-assigned grammatical tags (A-DSS-M) and automatedly-assigned grammatical tags (A-DSS-A). A statistically significant difference was found between the M-DSS and A-DSS-A scores, but not between the M-DSS and A-DSS-M scores or between the A-DSS-A and A-DSS-M scores. The DSS scores for individual participants were numerically similar across calculation methods, with the variation across methods ranging from 0.02 points to 1.20 points and the majority of participants displaying less than 0.5 points of variation in DSS score across calculation methods. The overall accuracy of DSSA 2.0 was moderately high, with 86.78% accuracy for A-DSS-M and 86.32% for A-DSS-A. The accuracy of individual grammatical category-point value cells, however, varied greatly, ranging from 0% to 98.88% accuracy for A-DSS-M and 0% to 100% accuracy for A-DSS-A. Earlier developing forms appeared to be scored more accurately than later developing forms; this trend was consistent across grammatical categories.

Long and Channell (2001) examined the accuracy of several automated grammatical analysis techniques, including DSS, using Computerized Profiling (CP) software compared to manual scoring, which was assumed to be 100% accurate. The reported percentage accuracy for automated DSS analysis with CP was 89.8%, which is higher than the overall accuracy found for

either method of automated DSSA 2.0 analysis evaluated in the present study. The lower level of accuracy found in the current study could be the result of differences in the sampling population: the samples evaluated in Long and Channell (2001) were primarily from preschoolers, and it is expected that preschoolers would use earlier grammatical forms, which were scored more accurately by both CP and DSSA 2.0, therefore resulting in increased overall accuracy.

Additionally, Long and Channell (2001) hypothesized that many of the automated DSS errors were due to errors in the underlying grammatical tagging. The present study evaluated this claim by comparing automated DSS completed with manually and automatedly assigned grammatical tags. Although the mean accuracy of the A-DSS-M scores was slightly higher than the A-DSS-A scores, this difference was not statistically significant, suggesting that errors in the automated DSS score were primarily the result of the DSS extraction rules rather than errors in the grammatical tagging. This finding is important to note, both for future improvements to DSSA software and in comparing DSSA 2.0 to the previous version (DSSA 1.0). The changes in the DSSA 2.0 software compared to the earlier version primarily targeted improved automated grammatical tagging rather improvements in the DSS extraction rules, which explains the minimal improvement in the overall accuracy of DSSA 2.0 (86.32% for the fully automated analysis) compared to the 86.03% accuracy for the same samples calculated with DSSA 1.0 (Judson, 2006).

Channell (2003) provided a more detailed analysis of CP's automated DSS, including category-point value cell accuracy along with the overall accuracy, as was done in the present study. Channell (2003) found an overall accuracy level of 78.2% (*SD* = 4.4). The increase in accuracy of DSSA 2.0 compared to Channell's (2003) data could be the result of improvements in the scoring software, differences in sample population, or a combination of both factors.

Channell analyzed samples from typically developing children and children with language impairment (LI) whereas the present study only examined samples from children reported as being typically developing. Multiple studies (Channell, 2003; Channell & Long, 2001) previously found that automated DSS was less accurate for children with LI, which might also help to explain the increase in overall accuracy in the present study compared to Channell (2003).

Channell (2003) also provided information on individual cell accuracy. Using the Long and Channell (2001) guidelines of considering automated accuracy greater than 85% as acceptable, greater than 90% as good, and greater than 95% as excellent, more cells from a wider range of grammatical categories fell into the excellent range for both the A-DSS-M and A-DSS-A calculation methods compared to Channell's (2003) CP data, including the point value cells of one in the indefinite pronouns column (i1), three and six in the conjunctions column (c3, c6), and eight in the wh- questions column (w8), along with one, two, and three in the personal pronouns column (p1, p2, p3). Furthermore, a greater number of cells fell into the acceptable range for A-DSS-M (four and seven in indefinite pronouns, two in main verbs, four in negatives, and five in conjunctions; i4, i7, m2, n4, and c5), although there was no overlap between the acceptable cells in the two studies. The number of acceptable cells with A-DSS-A remained the same (four in indefinite pronouns, four in main verbs, four in negatives, and two in wh-questions; i4, m4, n4, w2), although only one point value cell, two in the wh-questions column (w2), was found acceptable in both the A-DSS-A and CP data. Of the cells that fell in the acceptable range with CP analysis, three of four were in the good or excellent range with A-DSS-M and two of four were in the good or excellent range with A-DSS-A, with one remaining acceptable. This demonstrates an improvement in the accuracy of the DSSA 2.0 software compared to CP for

these specific cells in addition to the overall increase in accuracy. It is important to note, however, that some cells, such as the point values of six in the interrogative reversals column (r6) and five in the negatives column (n5), were significantly less accurate with DSSA 2.0 than with CP's DSS analysis. Additionally, fewer cells fell into the good range for both DSSA calculation methods (point values of three in the indefinite pronouns column and two in the wh-questions column for A-DSS-M and point values of one in the indefinite pronouns column and five in the conjunctions column for A-DSS-A) than with CP.

It is also significant to note that certain grammatical categories, such as indefinite pronouns and personal pronouns, appeared to be coded with high accuracy overall while other categories, such as interrogative reversal, demonstrated lower overall levels of accuracy. There are several possible explanations for the difference in coding accuracy between various point value cells and grammatical categories.

One possible explanation is a flaw or flaws in the underlying program design. Some of the DSS cells are infrequent in spontaneous language. Because the DSS extraction rules were created based on input data from child language samples, there is sparse data to support automated analysis for these cells, resulting in a less accurate automated analysis. Additionally, some categories and/or cells can be achieved by a wide number of grammatical constructions, making it difficult to consider all the possibilities as the basis for DSS extraction rules. For example, the salient features of the interrogative reversals category are spread further across the sentence, necessitating a greater number of extraction rules to account for all possible programs. Therefore, the lower accuracy of the interrogative reversal category can be explained by the increased computational complexity of establishing sufficient DSS extraction rules.

Furthermore, in her report on common DSS scoring errors, Lively (1984) points out that certain DSS categories are more difficult for human analysts to score correctly. For example, Lively noticed that her students of DSS had the most difficulty scoring the main verbs category correctly whereas there were few errors in scoring the wh-question category. In their comparison of computer-assisted instruction versus classroom instruction, Hughes, Fey, Kertoy, and Nelson (1994) similarly found that certain DSS categories were more difficult for students of DSS to master. Although DSSA does not display the same pattern of ease or difficulty of scoring particular categories as manual scorers, the fact that human analysts demonstrate increased difficulty with certain categories reveals that it would be a mistake to assume an even level of scoring difficulty across grammatical categories. Therefore, it is to be expected that automated DSS analyses, such as DSSA, would demonstrate increased difficulty with some grammatical categories. However, it is recommended that clinicians exercise caution when interpreting the results of the grammatical categories and cells that achieved low levels of accuracy with DSSA 2.0.

When interpreting these results, limitations of this study should be considered. These limitations include the size and diversity of the sample population. The language analyzed in this study came from a relatively small and homogenous population. All participants were residents of the same city: Provo, Utah. The accuracy of DSSA could be better understood by analyzing samples from more children and from a wider range of U.S. regions. Additionally, the samples assessed in this study were collected by speech-language pathology graduate students in 1989. These samples were used in the present study to allow for more direct comparison of DSSA accuracy across software versions; however, there is a possibility that trends and patterns in child language have shifted since 1989. Therefore, it would be beneficial to replicate the present study

with new, recently collected samples for increased accuracy and relevance to current clinical and research studies.

Additionally, all participants of the present study were reported as typically developing. Previous research by Long and Channell (2001) and Channell (2003) demonstrated that automated DSS is less accurate when scoring language samples from children with LI. Future studies could examine the accuracy of DSSA 2.0 on LI samples to provide a more complete picture of the strengths and weaknesses of this software program. Furthermore, future studies could seek to examine the cause of the reduced accuracy of DSSA 2.0 on LI samples as well as diminish this gap, since DSS is most often used to analyze the language of children with LI. Although the present study demonstrated that DSSA 2.0 can achieve moderately high levels of accuracy with typical samples, DSSA would need to reach this same level of accuracy with disordered language samples to maximize clinical usefulness.

Furthermore, the samples analyzed in the present study and in previous studies have all been conversational language samples. Expository and narrative-type language samples are also of great value in providing a deeper understanding of a child's language abilities. It would be of interest, therefore, to assess the accuracy of DSSA on these types of language samples. Finally, it is important to recognize that the limitations of DSS, as previously discussed, also apply to DSSA.

Based on the results of this study, DSSA 2.0 could be a useful tool for both clinicians and researchers. Although the accuracy of DSSA was moderately high, it is recommended that clinicians and researchers continue to manually edit the DSSA generated score given the statistically significant difference between the manual and fully automated scores. The lack of a significant difference between the DSSA score given input from manually and automatedly

assigned grammatical tags suggests that manual correction of DSSA auto tagging would have minimal impact on the final DSS score. Therefore, time and effort would be better spent on correcting the final DSS score than on the grammatical tags. It is also recommended that clinicians and researchers closely examine the DSS scores for the grammatical categories and point-value cells that demonstrated low levels of accuracy when scored with DSSA.

The results of this study showed that automated DSS has the potential to significantly reduce the amount of time required for DSS analysis while achieving moderately high levels of accuracy, thereby reducing one of the most significant disadvantages of DSS: the amount of time required to complete DSS scoring. Another significant disadvantage of DSS is the age of the normative data, which has not been updated since it was first published in 1975. Although DSS is still widely used in research, the lack of current normative data limits clinical applicability and usage of DSS. Future research to update the normative data for manual DSS or to establish of new normative data for automated DSS scores could increase the usage of DSS in clinical practice as a beneficial, standardized tool for language assessment.

While improvements to DSSA could still be made, this study confirmed previous findings that DSSA is capable of achieving moderately high levels of accuracy and provided important information regarding the source of errors in DSSA analysis. Although it was previously hypothesized that the auto tagging completed by the GramCats software may have been the primary source of error (Channell, 2003; Long & Channell, 2001), no significant difference was found between the DSSA analysis completed on samples that were manually tagged versus those that were tagged automatedly. Therefore, it can be concluded that the primary source of DSSA error is not the grammatical tagging, but in the DSS extraction rules.

This information is valuable both for directing future improvements to the DSSA software and for guiding where manual correction is most efficient and beneficial.

References

Channell, R. W. (1998). *GramCats*. Version 1.0 (MS-DOS) [Computer program]. Provo, UT: Department of Audiology and Speech-Language Pathology. Brigham Young University.

Channell, R. W. (2003). Automated Developmental Sentence Scoring using Computerized Profiling software. *American Journal of Speech-Language Pathology*, *12*, 369-375. doi: 10.1044/1058-0360(2003/082)

Channell, R. W. (2006). *DSSA*. Version 1.0 [Computer program]. Provo, UT: Department of Communication Disorders, Brigham Young University.

Channell, R. W. (2015). *DSSA*. Version 2.0 [Computer program]. Provo, UT: Department of Communication Disorders, 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.

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.

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.

Judson, C. A. (2006). *Accuracy of automated Developmental Sentence Scoring software* (Unpublished master's thesis). Brigham Young University, Provo, UT.

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.

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

Klee, T., & Sahlie, E. (1987). Review of Computerized Profiling. *Child Language Teaching and Therapy, 1,* 87–93.

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

Lively, M. A. (1984). Developmental Sentence Scoring: Common scoring errors. *Language, Speech, and Hearing Services in Schools, 15*, 154-168. doi:10.1044/0161-1461.1503.154

Leonard, L. B., Fey, M. E., Deevy, P., & Bredin-Oja, S. L. (2014). Input sources of third person singular –s inconsistency in children with and without specific language impairment. *Journal of Child Language*, *42*, 786-820. doi: 10.1017/S0305000914000397

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.

MacWhinney, B. (2000). *CLAN Manual*. Retrieved July 10, 2015 from http://childes.psy.cmu.edu/pdf/clan.zip

Miller, J. F., Andriacchi, K., & Nockerts, A. (2011). *Assessing language production using SALT software.* Middleton, WI: SALT Software.

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.

Salvia, J., & Ysseldyke, J. (2010). *Assessment in special and inclusive education* (12th ed.). Belmont, CA: Wadsworth.

Seal, A. (2002*). Scoring Sentences Developmentally: An analog of Developmental Sentence Scoring* (Unpublished master's thesis). Brigham Young University, Provo, UT.

Smith, J. M., DeThorne, L. S., Logan, A. R. L., Channell, R. W., & Petrill, S. A. (2014). Impact of prematurity on language skills at school age. *Journal of Speech, Language, and Hearing Research*, *57*, 901-916. doi: 10.1044/1092-4388(2013/12-0347)

Westerveld, M. F., & Claessen, M. (2014). Clinician survey of language sampling practices in Australia. *International Journal of Speech-Language Pathology. 16*(3), 242-249. doi: 10.3109/17549507.2013.871336

Appendix A: Annotated Bibliography

Channell, R. W. (2003). Automated Developmental Sentence Scoring using Computerized Profiling software. *American Journal of Speech-Language Pathology, 12*, 369-375. doi: 10.1044/1058-0360(2003/082)

Focus: Developmental Sentence Scoring is a well-known method of analyzing language samples and is valued for its ability to provide a standard score. In this study, Channell performed an in-depth analysis of the accuracy of automated Developmental Sentence Scoring (DSS) using Computerized Profiling (CP).

Participants: Language samples from 48 school-aged children, 28 of whom had language impairment, were scored by the CP software and manually. The children's ages ranged from 6;2 to 11;1.

Procedure: Channell examined the level of agreement between manual and automated DSS analyses, which he considered the accuracy of the automated DSS coding. He found the accuracy to be 78.2%.

Results: The CP-computed and manually-computed scores were highly correlated, although the CP scores were significantly higher than the manual scores. Channell examined the per-category and point-level levels of agreement by creating cells for each category and point level and provided percentages of correct tagging, false negatives (misses), and false positives (intrusions). The point-level level of agreement ranged from 10% to above 90%, with lower point value cells generally having a higher coding accuracy. There was no clear pattern for misses or intrusions. Some cells had far more misses than intrusions while some had more intrusions than misses; most had roughly equal amounts of misses and intrusions.

Discussion: According to this study, the accuracy of automated DSS is slightly below the acceptable level of 80%. The accuracy of CP-computed DSS was lower in this study than in the previous study conduced by Long and Channell (2001). This discrepancy is due to the increased complexity of the samples used in this study. The per-cell accuracy data provides guidance as to which automated analysis cells need more careful manual checking as well as direction for what areas of CP need further improvement.

Relevance for my study: One aim of my study is to evaluate the accuracy of a similar automated DSS analysis software. Thus, the accuracy of the program I am analyzing can be compared to the CP software. Additionally, my study will employ the same basic procedure that was employed in this study.

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

Focus: Probabilistic automated grammatical tagging uses information about the relative frequencies of possible tags a word may have (relative tag likelihood) and the probability that one particular tag will follow another tag (tag transition probabilities). This study examined the accuracy of a probabilistic automated grammatical tagging program, GramCats, compared to manual tagging on naturalistic child language samples.

Participants: Thirty naturalistic language samples were collected from typically developing children ranging in age from 2;6 to 7;11 as the child interacted with a graduate student.

Procedure: Thirty language samples were tagged manually and by the GramCats software. The tagged files were compared to calculate word-by-word, whole-utterance, and overall percentage of agreement.

Results: Word-by-word accuracy ranged from 92.9% to 97.4%, with a mean of 95.1%. Whole-utterance accuracy ranged from 60.5% to 90.3%, with a mean of 77.7%. Both word-by-word and whole-utterance accuracy had a significant negative correlation with age. Accuracy varied among tags and grammatical categories.

Discussion: This study found that probabilistic methods of automated tagging have the same level of overall word-by-word accuracy for naturalistic language samples from typically developing children as for adult written text. Additionally, the automated word-by-word reliability was almost as high as the reliability between human analysts, although the types of errors were different. However, the whole-utterance accuracy was considerably lower, suggesting that the program needs further improvement.

Relevance for my study: The results of this study show that automated grammatical tagging software has the potential to achieve similar levels of reliability as human analysts. Although the software used in my study has yet to achieve this level of accuracy on samples from children with language impairment, this study suggests that it may be possible. Additionally, GramCats is a component of the DSSA 2.0 software that I am using in my study.

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.

Focus: This chapter is a review of Developmental Sentence Analysis (DSA), including Developmental Sentence Scoring (DSS). Fristoe provides a summary of the purpose of DSA and how to administer, score, and interpret DSA, as well as an evaluation of the test adequacy.

Results: The purpose of Developmental Sentence Analysis (DSA) is to provide a detailed quantified and scored evaluation of a child's spontaneous speech. It allows a clinician to select developmentally appropriate goals and measure a child's progress. DSA has two components. Developmental Sentence Types (DST) evaluates whether a child has the elements of basic sentence development and is based on 100 utterances. Developmental Sentence Scoring (DSS) examines the development of certain grammatical structures within complete sentences and is based on 50 utterances. DSA is criterion-referenced, with the criterion being adult language structures. A major disadvantage of DSA is the amount of time required to complete the procedure. Additionally, DSA focuses more on surface structure than on underlying meaning. DSA does, however, provide a standard way of evaluating a child's expressive language and is useful in tracking progress and guiding intervention. Normative data for DSS is provided; data was collected from 200 white children age 2;0 to 6;11, with five males and five females in each three-month interval. However, the normative group was not randomly selected and is not representative of the general population of American children. Therefore, it is only appropriate to reference the norms with white, middle-class, monolingual children who speak standard American English.

Discussion: Although it has its limitations, DSA, when completed correctly, is a useful tool in measuring progress over time, assessing a child's use of specific syntactic structures, and, to some degree, determining whether or not a child is typical in his or her syntactic development.

Relevance for my study: The main disadvantage of DSA, according to Fristoe, is the amount of time required to complete the procedure. The software program I am evaluating has the potential to greatly reduce the amount of time required for DSS, thus reducing this disadvantage.

Additionally, the values and usefulness of DSA described in this article support the relevance of DSS in clinical practice, even today.

Hughes, D. L., Fey, M. E., Kertoy, M. K., Nelson, N. W. (1994). Computer-assisted instruction for learing Developmental Sentence Scoring: An exeperimental comparison. *American Journal of Speech-Language Pathology, 3*, 89-95.

Focus: One clinical application of computers is the development of interactive computer programs, known as computer-assisted instruction (CAI) to facilitate learning of grammatical analysis procedures. One grammatical analysis procedure, Developmental Sentence Scoring (DSS), has traditionally been taught via a classroom model, including reading, in-class lectures, and practice exercises. Hughes and Low developed a computerized DSS tutorial, the use of which would reduce the amount of instructor time required for students to master DSS. This study examined the efficacy of the CAI (Version 2.0) versus the more traditional classroom approach to learning DSS.

Participants: Fifty-five graduate students in speech-language pathology programs from three universities participated in the study. All participants were enrolled in classes on language disorders. Some participants had been introduced to DSS, but none had previously attempted to score sentences with this procedure.

Procedure: Participants were assigned to read Chapter 4 in Lee's (1974) text and attend a 2-hour introductory lecture after which they took a pretest. Participants were then assigned to one of two treatment groups for DSS instruction, the classroom-based tutorial (CBT) group and computer-assisted instruction (CAI). Instruction took place over a five-week period, after which

participants took a post-test to determine and compare the effectiveness of the two teaching methods.

Results: Participant performance was analyzed in a two-way ANOVA with instruction as the between-subjects factor and pre-/post- test as the within factor. There was no main effect for instruction method, $F(1, 53) = .69$, $p = .41$, showing that the two teaching methods did not differ in their effects on student performance. There was a significant effect for time, $F(1, 53) = 133.59$, $p = .0001$, showing that students' scores significantly improved as a result of instruction. Across both groups, 93% of the students achieved acceptable skill for effective clinical use of DSS, 80% accuracy. Additionally, there were significant reductions in the amount of time required for instructors and modest reductions for students with the CAI method compared to the CBT method.

Discussion: This study suggests that both methods of instruction were effective and did not significantly differ in terms of student post-test scores. Additionally, students performed at near-ceiling levels after instruction. Advantages of the CAI method include immediate feedback, convenience to learners, and time savings for instructors. Overall, the authors concluded that a combination of a CAI and CBT approach may be most effective.

Relevance for my study: The results of this study demonstrate that different methods of DSS learning can be effective and that students provided with extended practice and feedback can reach acceptable levels of scoring accuracy; this is important to note as I learn how to use the DSS system. Additionally, the authors suggest 80% accuracy as an acceptable level for effective clinical use. Thus, the program I am evaluating should reach at least this level of accuracy.

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.

Focus: In this article, the authors present reasons that DSS is still useful and relevant, despite being 20 years old at the time of the article's publication. The authors also provide suggestions for making new and more efficient uses of DSS, both clinically and in research.

Procedure: The authors begin by explaining what DSS is and why and when it is useful. They also discuss some of the difficulties associated with DSS.

Results: DSS is a measure of spoken syntax. Scores, which range from one- to eight- points, are awarded for eight grammatical categories according to developmental order. Normative data from 200 children ranging in age from 2:0 to 6:11 allows clinicians to compare their client's performance with other children. The three primary features of DSS that make it useful are: (a) the numeric variable that allow for comparison across time or between children, (b) the developmental data, and (c) DSS's organizational framework, which helps clinicians ask and answer clinical questions. It is important to note that DSS scores alone do not provide complete information on a child's ability; however, DSS is useful in making diagnostic judgments, especially when quantitative measures are required. Three different studies have found that DSS can discriminate between typically developing and language-impaired children. Children with language impairment tend to score significantly lower and appear to perform poorest on the main verb, sentence point, secondary verbs, negatives, and conjunctions categories. Analysis of attempt marks, point values, sentence point errors, and frequency of occurrence of forms for each category is useful in guiding goal selection and treatment planning. Additionally, DSS is useful in measuring intervention efficacy. The authors of this study conducted a comparative analysis that demonstrated that language-impaired children who received language treatment made

observable DSS gains compared to language-impaired children who did not receive treatment. The authors conclude by acknowledging some criticisms of DSS and discussing a few DSS rules that they believe are counter-intuitive or likely to skew the results.

Relevance for my study: My study is evaluating a program for computerized DSS analysis. This program has the potential to dramatically reduce the time required for DSS analysis, thus making it a more realistic clinical tool. It is important to recognize that, despite its age and flaws, DSS is still a valuable clinical tool for many reasons.

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.

Focus: Experts have long agreed that language sampling is an important aspect of spoken language evaluations. Barriers to implementing language sampling in clinical practice include the amount of time required, the difficulty of eliciting a representative language sample, and the lack of procedural uniformity. Despite these barriers, the use of language sampling is relatively wide-spread. The purpose of this study was to survey school-based speech language pathologist's (SLP's) methods of language sample collection and analysis.

Participants: Surveys were sent to 500 school-based SLPs from 10 Midwestern states, 50 from each state. Participants from each state were randomly selected from personnel lists from the State Department of Education or the state professional organization for SLPs.

Procedure: A 51-item survey, assessing SLP's background information, practices, and attitudes regarding language sampling procedures was filled out by 239 SLPs from 9 Midwestern states.

Results: Two-thirds of the respondents reported receiving language sample analysis training in college. Half of those also reported participation in continuing education courses related to

language sampling. Over 82% of the respondents reported that language sample analysis was not required by their state or district, although approximately half of these encouraged it. Language sample analysis was most likely to be carried out with younger children (i.e., elementary school) and with more severe cases. Fifty-one percent of the respondents gathered language samples from one setting only, usually the clinician's treatment room. Pictures, manipulatives, familiar subjects, and story books were the most common materials used for elicitation. Conversation (82%) was the most common elicitation technique, followed by descriptions (66%), story retelling (54%), and others. As a whole, respondents preferred using non-standardized language sample analysis procedures. Only 63% had a "most preferred" procedure; 49% of these were self-designed. The only standardized language sample analysis procedure used with regularity was Developmental Sentence Scoring (DSS) at 31%. Fifteen percent of the respondents who stated a preference selected DSS. Only 3% of respondents reported using a computer-assisted analysis. The most common uses of language sample analysis were supplementation of standardized procedures (80%) and assistance in treatment planning (77%). The majority of respondents viewed language sample analysis as reliable and helpful in distinguishing between language-delayed and typically developing students.

Discussion: The results of the survey suggest that speech-language pathologists are aware of the beneficial information gained through language sample analysis and regularly implement these procedures, even when they are not required to. As a whole, collection procedures followed the guidelines and suggestions given by researchers, including sample size and elicitation tasks and materials. The broad tendency of clinicians to use self-designed and non-standardized procedures may be a result of the narrow focus of many standardized procedures compared to the wide range of child language behaviors. However, the quality, efficiency, and adequacy of the use of these

self-designed procedures are questionable. Restrictions on time, finances, and resources may make in-depth analysis difficult for school-based clinicians to carry out regularly.

Relevance for my study: This study shows that school-based clinicians regularly use language sample analyses in clinical practice. Furthermore, DSS was the most commonly used standardized procedure, suggesting that a software program that would make DSS analysis faster and easier for the clinician would be useful and well received. Additionally, the increased speed and easiness of use of an automated DSS may encourage more clinicians to adopt this procedure, rather than self-designed procedures of questionable quality.

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.

Focus: The Developmental Sentence Scoring (DSS) protocol recommends a 50 utterance sample size for analysis. However, the principles of statistics state that reliability increases as sample size increases. A study of DSS reliability conducted by Lee and Koenigsknecht (1972) found reliability to be lower than ideal given that DSS is frequently used for making important clinical and research decisions. Johnson and Tomblin conducted a study to evaluate how DSS reliability was influenced by sample size.

Participants: Language samples were collected from 50 children, ages 4;8 to 5;8. The children, who were selected from the University of Iowa Institute for Child Development Preschool, were all monolingual and had normal hearing.

Procedure: Language samples were collected from 50 children. Samples were collected in an individual setting and were elicited through use of two sets of questions, two types of picture stimuli, and a number of common household tools. Fifty complete, consecutive, different, and

intelligible sentences were transcribed for each child. Twenty-five sentences from each corpus of 50 were randomly selected for DSS scoring. Each 25-sentence sample was divided into five response segments of five sentences. The total and component scores were recorded for each segment. Reliability measures and standard errors of measurement were calculated for sample sizes of five to 250 sentences (one to 50 response segments).

Results: The reliability of both the total and the component scores increased with sample size. The increase was fairly dramatic with each 25-sentence sample size increase up to 150 sentences (from 0.60 for 25 sentences to 0.90 for 150 sentences), after which the increase in reliability was much smaller (.001 for each 25-sentence increase). The reliability value for a 50 sentence sample was 0.75.

Discussion: The results of this study suggest that a larger sample is required to obtain equivalent reliability values with other measures, such as MLR (see Darley and Moll 1960). The standard error of measurement values provided in this study can be used to help clinicians determine an appropriate sample size for DSS to achieve desired accuracy. Using these values, the authors determined that a sample size of approximately 175 sentences is necessary to achieve even limited reductions in error. However, the authors also recognize that collecting such a large sample may not be possible in many instances. Therefore, the authors caution that clinicians recognize the degree of error potentially present in DSS and not base clinical decisions solely on this measure.

Relevance for my study: Although DSS has its merits, it is important to recognize its limitations as well. Some of the sample I am using in my study are smaller than the 175 sentences recommended by these authors. Therefore, according to the results of this study, it is possible that the DSS scores are not entirely representative of each child's true ability. Additionally, it is

important that I promote awareness of these findings and recommend that DSS scores should not be the solitary factor in clinical decision making.

Judson, C. A. (2006). *Accuracy of automated Developmental Sentence Scoring software* (Unpublished masster's thesis). Brigham Young University, Provo, UT.

Focus: Computerized language analysis has the potential to significantly reduce the amount of time required to complete language analyses. As such, several different automated analysis programs have been developed. One program is a recently developed automated Developmental Sentence Scoring (DSS) software, DSSA. This study examined the accuracy of DSSA compared to manual DSS scoring.

Participants: 118 language samples were collected from 99 children between the ages of 3 and 11. Subjects included both typically developing and language impaired children. The samples in this study were taken from five corpora. One corpus consisted of 20 samples collected from one typically developing child, "Adam" (Brown, 1973), thirty were previously collected from Reno (Fujiki et al., 1990), eighteen were collected from Jordan School District in Salt Lake County, Utah (Collinridge, 1998), twenty were collected from Wasatch School District in Utah (Nichols, 2002), and thirty were collected from Provo, Utah (Channell & Johnson, 1999; Seal, 2001).

Procedure: DSS was conducted on each language sample both manually and using DSSA software. Accuracy was inferred from percent agreement between DSSA and manual coding; the manual coding was assumed to be 100% accurate. The *Point Difference* method (Sagae, Lavie, & MacWhinney, 2005) was used to compare automated and manual scores; the mean point difference was determined by averaging the differences in absolute values between the automated and manual scores.

Results: The overall accuracy of DSSA for the between-child corpora was found to be 85.99% with a standard deviation of 5.05. DSSA was less accurate with language samples with lower manual DSS scores and children with language impairment; the mean for children with language was around 84% compared to 88% accuracy for typically developing children. Mean accuracy of the Adam corpora was found to be 82.70% with a standard deviation of 3.67. The author provides data on the number of agreements, misses, intrusions, and accuracy of DSSA for each point value in each grammatical form category. P3 in the person pronoun category was the most accurate, at 99% accuracy while the interrogative reversal category had the least accurate scores, with r8 at 0% and r1 at 5%. Earlier developing forms were generally scored more accurately than later developing forms. However, the negatives category did not fit this trend and no trend could be determined in the secondary verbs and interrogative reversals categories. The overall mean point difference of the between-child corpora was found to be .74 (*SD* = .30) and the mean point difference for the Adam corpus was .91 (*SD* = .36; Judson, 2006).

Discussion: The overall accuracy of DSSA compared to manual scoring was moderately high; however, accuracy of individual grammatical form categories and scores was variable. This suggests that manual correction of certain forms may be necessary.

Relevance for my study: The results of this study suggest that automated DSS procedures are in need of further improvement to reach an acceptable level of accuracy. My study is assessing whether the accuracy of an updated version of DSSA has increased to acceptable levels given improvements in the underlying software as well as examining the source of error in DSSA analysis.

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.

Focus: Analysis of clinical language samples allows clinicians to examine a child's language in the context of naturalistic interactions. There are many different methods and types of analyses. This study examined the clinical practices of speech-language pathologists in the United States in regards to language sample analyses.

Participants: Information was collected from a representative sample of United States-based speech language pathologists working with pre-school children. Participants were selected from the 1993-94 Directory of the American Speech-Language Hearing Association (ASHA); all respondents held the ASHA Certificate of Clinical Competence in Speech-Language Pathology (CCC-SLP). Of the 500 speech-language pathologists who received a survey, 253 (51%) responded.

Procedure: A 22-question survey was sent by mail to 500 randomly-selected participants with a cover letter indicating that the purpose of the study was to explore current practices in preschool language assessments. The survey asked about the participant's professional qualifications, work experience, current caseload, use of standardized tests, and use of language sample analysis.

Results: Only 8% of respondents reported that language sample analysis was mandated by their state; however, 85% reported using language sample analysis, suggesting that the majority use language sample analysis by choice. Of those who reported using language sample analysis, 92% use it for diagnosis, 77% as part of intervention, 64% in the post-intervention stage, and 44% for screening. The participants who reported not using language sample analysis cited a lack of time as the most common reason (86%). Other reasons included lack of computer resources (40%),

lack of training and expertise (16% each), and financial constraints (15%). The most common transcription practices included using real-time transcription and basing analysis on a fixed number of utterances, with 50 utterances being the most common length. Forty-eight percent of clinicians who use language sample analysis reported using non-standardized procedures. Developmental sentence scoring (DSS) was the most frequently used standardized procedure; 35% of clinicians used DSS. Only 8% of participants reported using a computer program for analysis.

Discussion: The high percentage of clinicians who reported using language sample analysis, even when the state did not mandate it, suggests that language sample analysis is viewed as an important part of the assessment process by most clinicians. According to the results of this survey, the most common procedure for language sample analysis is live transcription based on 50 utterances and analyzed with a nonstandard procedure.

Relevance for my study: The results of this study suggest that the majority of clinicians view language sample analysis as extremely important; however, many clinicians do not have enough time for these analyses. The automated DSS program that I am evaluating has the potential to reduce the amount of time required for a language analysis, thus making the routine analysis of language samples more feasible. Additionally, DSS was the most frequently used standardized procedure, suggesting that there would be a demand for automated DSS.

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

Focus: Language sample analysis (LSA) is a method of childhood language assessment that seeks to systematically describe a child's expressive language abilities in an interactive

conversation. This article provides a rationale for the use of language sample analyses in a clinical setting and evaluates 14 different methods of language sample analysis.

Procedure: Klee first presents reasons why LSA is a valuable and necessary clinical tool. He explains the four steps of language sample analysis and provides a review of the main branches of linguistic analysis and the clinical procedures used to assess each area. Klee then discusses and compares three procedures – DSS, LARSP, and SALT – in detail. The article concludes with a discussion of numerical indices in language sample analysis.

Results: Language sample analysis is the most direct method of assessing, describing, and understanding linguistic behavior. Because LSA evaluates spontaneous language, it has the advantage of being representative of the client's true abilities whereas elicited language may not be representative. Additionally, the comprehensive level of detail inherent in LSA is useful in establishing appropriate treatment goals. The four steps of language sample analysis are (1) recording the conversation, (2) transcription, (3) analysis, and (4) interpretation. Phonological analyses, which examine the client's sound system, include five procedures: PROPH: Profile of Phonology; NPA: Natural Process Analysis; PA: Phonological Analyses; LINGQUEST 2: Phonological Analysis; and PROP: Prosody Profile. Semantic analyses include lexical semantics and relational semantics. Systems for evaluating semantics include PRISM-L: Profile in Semantics; LINGQUEST 1; SALT; Prism-G: Profile in Semantics—Grammar; and C/F A: Content/Form Analysis. Grammatical analyses include LARSP: Language Assessment, Remediation and Screening Procedure; ASS: Assigning Structural Stage; SALT: Semantic Analysis of Language Transcripts; LSAT: Language Sampling, Analysis and Training; LINGQUEST 1: Language Sample Analysis; and DSS: Developmental Scoring. These procedures vary in terms of theoretical basis, the level of grammatical coverage and detail, and

which areas of grammar are analyzed. There are disadvantages inherent in numerical indices measuring length (i.e., MLU) and vocabulary (i.e., type-token ratio); thus, these alone cannot be trusted as accurate measures of language abilities.

Discussion: According to Klee, one advantage of DSS is its usefulness in establishing a baseline from which clinical progress can be measured. A disadvantage of DSS is that the developmental ordering within some of the grammatical categories no longer reflects current research, which may lead therapists to select developmentally inappropriate goals. Klee notes that analysis of a language sample using DSS, LARSP, or SALT may provide a different clinical impression of the child based on which procedure was used. Compared to LARSP, which is a multilevel analysis, DSS is a phrase level analysis guided by lexical items. It does not distinguish between lexical items that function on different syntactic levels, for example, *and* used as a phrase level conjunction versus a clause coordinator.

Relevance for my study: Klee provides a rationale for using language sample analyses clinically. The goal of the automated analysis I am evaluating in my study is to reduce the time required to complete a language sample analysis, and, thus, make these procedures more realistic for clinicians to complete. Additionally, Klee briefly discusses some of the advantages and disadvantages of DSS. It is important to be aware of DSS's merits and limitations so that it can be used correctly within the clinical setting.

Klee, T., & Sahlie, E. (1986). Review of DSS Computer Program. *Child Language Teaching and Therapy, 2,* 231–235.

Focus: This article is a review of DSS Computer Program (DSSCP), which was developed by Peter K. Hixson in 1983.

Procedure: Klee and Sahlie describe DSSCP, including the objectives, the target population, and program flexibility. They then provide a critique of the program.

Results: DSSCP was designed as the first automated Developmental Sentence Scoring (DSS) program. The program aims to reduce the time required to conduct a language sample analysis by automating tallying and computing the points assigned to sentence constituents. DSSCP segments utterances into units of analysis, which are then compared to entries in the program's 'dictionary' in order to assign segments into the DSS grammatical categories and associated scores. Scores are calculated for each sentence and averaged over 50 utterances to produce a developmental sentence score. DSSCP also computes an Attempt Score to reflect what the average sentence score would have been if all forms were produced correctly, and an Error Score, which is the difference between the obtained score and the attempt score. DSSCP does require special codes in the language sample transcription, such as a < preceding irregular past tense verb forms. DSSCP is intended to analyze language samples for children age 2 to 6. Users of DSSCP need to have a thorough understanding of DSS. Even for clinicians who are proficient at DSS, it takes several hours of training to become familiar with the transcription format. One difficulty with DSSCP is the outdated nature of the developmental sequence inherent to DSS. There are also some disadvantages of the actual DSSCP program: the user needs to be cautious when inputting the transcript to avoid over-scoring, the transcription used for DSSCP is significantly different than the original spoken utterance, and the program does not provide error messages.

Relevance for my study: DSSCP was the first attempt at a computerized version of DSS and thus can be compared to the program I am using in my study.

Klee, T., & Sahlie, E. (1987). Review of Computerized Profiling. *Child Language Teaching and Therapy, 1,* 87–93.

Focus: This article is a review of Computerized Profiling (CP) Version 1.0, which was developed by Stephen H. Long in 1986.

Procedure: Klee and Sahlie describe CP, including the objectives, how the program works, the target population, and program flexibility. They then provide a critique of the program.

Results: Computerized profiling integrates several clinical language analyses into a single program to reduce the amount of time required for language sample analysis in order to encourage greater clinical use. CP contains CORPUS for creating and editing a transcript file; DSS for lexico-grammatical analysis; LARSP for grammatical analysis; PROP for prosodic analysis; PRISM-L for lexical semantic analysis; and PROPH for phonological analysis. LARSP must be completed before DSS or PROP may be completed. Language transcripts are entered into CP using the CORPUS model. This file can then be used for LARSP analysis. CP can, in principle, be used to analyze the language production of children or adults. CP is easy to learn; however, there are several disadvantages. These include restrictions on maximum corpus size, word truncation, and misanalysis of multiply embedded clauses. Clinicians who use CP need to be knowledgably about the the linguistic analyses in order to make manual corrections, which are regularly needed. Unfortunately, making the manual corrections to the computer analysis takes longer than just doing the analyses by hand.

Relevance for my study: My study will analyze a separate automated DSS software, DSSA, the results of which can be compared against the results of CP's DSS analysis.

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

Introduction: Lee begins by providing a brief summary of psycholinguistics research examining grammatical development as well as the applicability of psycholinguistics to clinical practice. Lee explains that Developmental Sentence Analysis, including Developmental Sentence Scoring (DSS), is a method for making a quantified and scored evaluation of a child's use of standard American-English grammatical rules, which allows clinicians to select appropriate goals and track progress. Lee explains that Developmental Sentence Analysis is only appropriate for children learning standard American-English. The introduction concludes with an explanation that the book does not separate the various etiologies of language-learning problems because every child needs to learn the same language forms, regardless of the etiology of the learning difficulty.

Chapter 1: Chapter 1 provides background information on grammatical structure, including the basic concepts and terminology of grammatical analysis. This chapter does not strictly follow a traditional or a psycholinguistic view; rather, it combines elements of both while striving to use familiar terminology. The basic sentence, which is one of the early landmarks in child language development, is explained using both transformational generative grammar and case grammar. Clinical problems in basic sentence formulation are also addressed. Elaboration of the basic sentence, including elaboration of the noun phrase, development of pluralization, and elaboration of the verb are discussed, followed by modifications of the basic sentence. For each modification, the grammatical rules and clinical problems with that modification are discussed. Lee concludes with a discussion of the mechanisms for combining basic sentences, including conjunctions, conjunctions with deletions, infinitives, participles and gerunds.

Chapter 2: Chapter 2 provides a detailed description of the process of taking a language sample in a clinical setting. First, Lee provides instructions on how to talk to children who are learning language, including suggestions for how to help produce samples from children with language problems. Lee then discusses the transcription process, including corpus selection for both parts of Developmental Sentence Analysis – Developmental Sentence Types (DST) and Developmental Sentence Scoring (DSS) – and scoring procedures.

Chapter 4: Chapter 4 explains the third (and final) version of Developmental Sentence Scoring (DSS). Lee first explains the scoring system, including scores and attempt marks, the sentence point, and details of the eight grammatical forms being scored. These 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- questions. Scores range from one- to eight-points for each category, with higher scores being awarded to more developmentally advanced forms. Lee then explains how to derive and evaluate the Developmental Sentence Score, including norms. She notes, however, that the DSS score alone is not a sufficient diagnostic tool. DSS is better suited to track progress during therapy.

Chapter 5: Chapter 5 discusses Developmental Sentence Analysis, a procedure that includes classifying and scoring a 100 utterance speech sample using Developmental Sentence Type (DST) and Developmental Sentence Scoring (DSS) to gain information about which grammatical rules the child has demonstrated. Although the analysis does not provide an exact account of all the grammatical structure that a child can use, it can direct the clinician's attention to specific language areas for further and continuing investigation and inform treatment goals. Lee provides and compares examples of detailed grammatical analyses for three children and how teaching goals can be derived from this information.

Relevance to My Study: These chapters are relevant to my study for several reasons. Because my study is examining an automated DSS program, it is important to understand the theory and background behind DSS. I will manually score all of the language samples as a point of comparison with the automated analysis; thus, it is important that I have a thorough understanding of how to complete DSS scoring. Additionally, this book explains the value of DSS analysis and why it is useful as a clinical tool.

Leonard, L. B., Fey, M. E., Deevy, P., & Bredin-Oja, S. L. (2014). Input sources of third person singular –s inconsistency in children with and without specific language impairment. *Journal of Child Language, 42,* 786-820. doi: 10.1017/S0305000914000397

Focus: As children progress through the preschool years, the proportion of correctly used tense/agreement morphemes gradually increases for most children. However, children with language impairment have a prolonged period of inconsistency and lower levels of correct productions. There have been multiple theories as to the reason for this inconsistency. Previous studies suggested that novel verbs were produced primarily in the manner in which they were heard, especially for children with language impairment.

Participants: Sixty children, thirty with specific language impairment (SLI) and thirty typically developing younger children were selected for this study. The children with SLI ranged in age from 3;9 to 5;8 and were selected due to a score below 87 on the *Structured Photographic Expressive Language Test – Preschool 2* or a score below the 10th percentile on Developmental Sentence Scoring. The typically developing children ranged in age from 2;5 to 3;1. Mean Length of Utterance (MLU) was statistically similar for both groups.

Procedure: Subjects were given an 18-item probe to assess their use of third person singular -s to confirm that the children were inconsistent in their use of third person singular *–s* and that there was a difference between the two groups. Fifteen children from each group heard five novel verbs and five familiar verbs in the finite frame and a different set of five novel and five familiar verbs in the frame *Let's watch the NV?*. The fifteen other children from each group heard five novel and five familiar verbs in the finite form and five novel and five familiar verbs in the frame *Does the N V?* For each frame type, half the novel verbs were matched with the remaining five novel verbs on length, phonotactic probability, and neighborhood density; these were also matched to the familiar verbs. One novel verb and one familiar verb were presented to the child during each session. The session began with the child playing with toy characters by performing a variety of novel actions and familiar actions. The child was instructed to remember a funny name for a novel action, which the experimenter modeled in the assigned frame three times. This was followed by the experimenter modeling the familiar verb three times. The child was presented with the three presentations in nine sets for a total of 27 exposures to each verb. For half of the children, the first five sessions employed non-finite presentation frames and the next five sessions employed finite presentation frames. The other half of the subjects received the opposite order. At the end of each session, the subjects' use of the two verbs were assessed. Three items assessed the child's use of the verb in a context with an obligatory third person singular *–s* and three items assessed a non-finite context; note that the verbs would have been heard in only one of these contexts during the exposure period. The probe followed a sentence completion format, with the examiner making the toy perform an action and providing a sentence for the child to complete appropriately. For both contexts (the third person singular *–s* and the non-finite), the first probe item used the same subject as the exposure period while the second

and third items were used with a novel subject. The subjects' verb productions were then scored. In the established scoring system, an attempt score was awarded if the production contained no more than one phoneme error that deviated from known developmental errors; more than one developmental error was allowed. Productions that corresponded to a real verb that was a description of the action and productions that matched a previously trained novel verb were judged non-attempts. Productions with an [s] or [z] at the end of a production considered an attempt at the appropriate verb were considered attempts at third person singular *–s*. Correct responses of the finite probe items were productions of the appropriate verb with a third person singular *–s* inflection.

Results: *Finite verb probes.* A mixed model ANOVA found a highly significant effect for exposure condition. Verbs heard exclusively in the finite form were more likely to be produced correctly on the finite verb probe than verbs heard exclusively in the non-finite form. No significant effect was found for participant group, sentence frame, or verb familiarity. The interaction of exposure condition x verb familiarity was significant; both familiar and novel verbs heard in finite form were more accurate than non-finite form. However, this difference was much greater for the novel verbs. Additionally, novel verbs heard in finite form were more accurate than familiar verbs heard in finite form. All other interactions were non-significant. Both groups were as likely to produce the third person singular *-s* correctly when the novel verb followed an unfamiliar subject as with a familiar subject. *Non-finite verb probes.* A mixed-model ANOVA found a main effect for participant group, verb type, and exposure condition. The typically developing children performing more accurately than the children with SLI, familiar verbs were more accurate than novel verbs, and verbs heard in non-finite form were more accurate than those heard in finite form. The participant group x exposure condition interaction,

exposure condition x verb familiarity interaction, and participant group x exposure condition x sentence frame interaction were significant as well.

Discussion: This study addressed four questions. The first question – whether there were significant input effects and if the effects were stronger in children with SLI – was found to be true in that there were strong input effects for both groups and the input effects were stronger for children with SLI, but only on the non-finite probes. The second question asked if children with SLI were prone to over-apply third person singular *-s* on sentences containing finiteness dependencies, which was found to be true. This suggests that children's tense and agreement errors may be based in difficulties recognizing the larger verb context. Additionally, the authors concluded that errors on non-finite probes were an over-application of the third person singular *-s* form, possibly due to a failure to recognize that the finiteness is tied to the earlier verb. The third question was whether input effects were seen for all sentence frames. Significant input effects were observed for both types of finite exposure frames and both non-finite frames. For children with SLI, one type of finite frame had a greater input effect than the other finite frame. The final question asked in input effects were stronger for novel verbs than familiar ones and if inflected novel verbs could be used with novel subjects. As predicted, novel verbs did exert a stronger input effect than familiar verbs, especially for finite forms. Overall, this study found that the children's use of finite or non-finite forms could be traced to the input; this was stronger with the SLI group compared to to typically developing group.

Relevance for my study: One of the measures used to evaluate children's language abilities and qualify subjects for the SLI group was Developmental Sentence Scoring (DSS). This demonstrates that, although the clinical use of DSS may have declined, it is still used in the research setting. Therefore, an automated DSS program would be beneficial for researchers.

Lively, M. A. (1984). Developmental Sentence Scoring: Common scoring errors. *Language, Speech, and Hearing Services in Schools, 15*, 154-168. doi:10.1044/0161-1461.1503.154

Focus: For many years, Developmental Sentence Scoring (DSS) was one of the most frequently used methods for evaluating preschool children's language. However, the DSS procedure requires much study and practice to learn. This article describes several scoring errors that are commonly seen among those learning DSS in order to help clinicians recognize potential problems, facilitate learning of DSS, and reduce scoring errors.

Participants: Student clinicians working with the author were observed as they learned DSS.

Procedure: Lively observed student clinicians learning DSS and noted several "problem areas" with frequent, consistent scoring errors. She describes and provides examples of each of these areas, including common mistakes and how to score each area correctly.

Results: The most common problems in each scoring area were as follows.

• Determining an appropriate 50-response language sample: Common errors included listing an utterance multiple times and having difficulty determining what counted as a "complete" sentence.

• Awarding the sentence point: These errors seemed to reflect students' lack of knowledge of grammatical rules.

• Using attempt marks and incomplete designations: Students often awarded a score to incorrect structures, rather than an attempt mark. Students also awarded a sentence point to utterances with attempt marks.

• Indefinite pronouns and noun modifiers: The two most common errors included scoring adverbs as indefinite pronouns or noun modifiers and forgetting to score words in this category.

• Personal pronouns: Errors often related to confusion between wh-pronouns and wh-conjunctions.

• Main verbs: This was the area in which errors were most common. Errors included misuse of attempt marks, mis-scoring of inflected verbs, mis-scoring of auxiliary *do* and modal verbs, difficulties with passive constructions, and errors with compound sentences. Lively hypothesized that many of these errors occurred because students were scoring the sentence one word at a time, rather than identifying the main verb phrase.

• Secondary verbs: In many instances, students failed to notice that a secondary verb was present, leading to errors in scoring. Errors also occurred in scoring infinitives.

• Negatives: The most common errors were in discriminating what scored 1 point and what scored 7 points in this category.

• Conjunctions: Mistakes included confusion between wh-conjunctions and wh-pronouns, not scoring conjunctions that introduce an independent clause at the beginning of a sentence, and the special rules for dividing sentences with multiple *and*s.

• Interrogative reversals: Students occasionally forget to score the interrogative reversal when scoring wh-questions.

• Wh-questions: Errors in this category were rare.

Relevance for my study: The samples evaluated in my study require manual as well as automated DSS scoring. This article made me aware of errors that many DSS learners make, and, thus, errors that I am likely to make or observe in the manual scoring. I will need to pay close attention to these potential errors as I score to insure that my DSS scoring is accurate. Additionally, a software program capable of fully-automated DSS analysis, such as the program

I am evaluating, could greatly reduce the frequency of these errors, if brought to a sufficient level of accuracy.

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

Focus: Although language sample analysis is an important part of a clinical evaluation, it takes a considerable amount of time to complete. The time required to complete the analysis depends greatly on the complexity of the method of analysis, the length and complexity/severity of the sample, and the knowledge and experience of the clinician. It has been theorized that computerized analyses could greatly reduce the amount of time required. This study examined the time-efficiency of computerized versus manual analysis for several analysis procedures.

Participants: Two hundred and fifty-six students and practicing speech-language pathologists from the United States and Australia participated in the study. All participants had received university-level instruction on the analysis procedures they completed.

Procedure: Phonological analyses were conducted on three samples ranging in size, complexity, and lexical variability. Grammatical analyses were conducted on three other samples, also varying in size, complexity, utterance variability, and suitability for different grammatical analyses. Multiple participants analyzed each sample; each participant analyzed the samples by hand and by computer. The Computerized Profiling (CP) software was used for the computer analyses. Participants carefully tracked the amount of time spent on each analysis. The phonological analyses included (a) type-token ratio (TTR), (b) variability analysis, (c) homonymy analysis, (d) word shape analysis, (e) vowel inventory, (f) consonant inventory, (g) vowel target analysis, (h) consonant target analysis, (i) percentage consonants correct (PCC), and

(j) phonological process analysis. Grammatical analyses included (a) MLU and descriptive statistics, (b) number of syntactic types (NST), (c) LARSP, (d) Developmental Sentence Score (DSS), and (e) Index of Productive Syntax (IPSyn).

Results: The length of time required to complete all ten phonological analyses was considerable, although it varied according to the type of sample being analyzed. Sample P1, which was the simplest, took, on average, just over 3 hours to complete. Sample P2, which was the longest and most complex, took almost 10 hours, on average. There was variation across samples within each procedure as well as variation between procedures. Phonological process analysis was the procedure that required the most time to complete. Computerized phonological analysis was at least 11 times faster than the comparative manual analysis; the average was between 17 and 35 times faster. Computerized grammatical analysis was also more time-efficient than manual grammatical analysis, although significantly less so than for the phonological analysis. The average efficiency ratio was between one and five. The simpler analysis procedures (i.e., MLU, NST, and descriptive statistics) were generally completed more quickly than the elaborate measure (i.e., LARSP, DSS, and IPSyn).

Discussion: The quantification of the amount of time required for clinical language sampling completed in this study demonstrated that manual language sampling, as intended for treatment planning, will not be regularly possible for most clinicians, despite the importance of and need for language sample analysis. Additionally, the procedures that require the least amount of time to complete are also the least useful at informing treatment goals while the analysis procedures that provide the best information for treatment planning require the greatest amounts of time. Computer analysis software greatly reduced the time required for all analysis procedures. For efficient clinicians, analysis software should make language sample analyses a realistic practice.

Relevance for my study: This study effectively demonstrates the time-saving capabilities of computerized grammatical analysis, including for DSS analysis. These findings support the basis of my study – that developing an accurate automated DSS analysis software could save clinicians significant amounts of time and make language sample analyses a more realistic procedure.

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

Focus: Many computer programs have been developed to conduct language analyses faster than could be done by hand. However, these programs have restricted clinical application due to limitations in complex analyses, such as at the clause and sentence levels. Long and Channell compared the ability of four different automated language analysis procedures – MLU, LARSP, DSS, and IPSyn – to provide clinically significant analyses.

Participants: Language samples were collected from 69 children, aged 39 to 94 months. Participants represented a range of dialects, linguistic levels, and diagnostic categories, which included fluency, Specific Language Impairment (SLI), Specific Expressive Language Impairment (SELI), and typically developing.

Procedure: Sixty-nine language samples were analyzed by four different procedures. Each procedure was used in two conditions, first with the software program Computerized Profiling (CP) alone and second with human judges reviewing and correcting the CP codes. The two conditions were compared to yield an accuracy score for the fully automated analyses.

Results: Overall, the automated analyses tended to yield higher scores than the manually corrected scores. The simplest measure, MLU, had the highest accuracy (99.4%), followed by IPSyn (95.8%) and DSS (89.8%). LARSP did not yield a comparable accuracy summary score.

Point-by-point comparisons found a negative correlation between the size of the corrected IPSyn score and the accuracy of the automated calculation and a positive correlation between the size of the corrected DSS score and the accuracy of the automated calculation. The positive DSS correlation was due to the fact that the most accurately scored elements occurred more often as utterance length and complexity increased. LARSP had low accuracy for subordinate clause structures but reasonable accuracy at the word, phrase, and clause levels.

Discussion: After comparing the automated analyses scores with available reference data for each procedure, the authors concluded that fully automated MLU and LARSP yield acceptably accurate data, automated IPSyn scores should be manually reviewed and edited when the score is within 6 points of the cutoff score, and automated DSS should always be manually reviewed.

Relevance for my study: The results of this study suggest that automated DSS procedures are in need of further improvement to reach an acceptable level of accuracy. My study is assessing whether the accuracy of fully automated DSSA analyses has increased to acceptable levels given improvements in the underlying software.

MacWhinney, B. (2000). *CLAN Manual*. Retrieved July 10, 2015 from http://childes.psy.cmu.edu/pdf/clan.zip

Focus: This manual describes Computerized Language Analysis (CLAN), including the automated DSS analysis. MacWhinney first details the CHAT conventions that must be followed for DSS to run correctly and inclusion criteria for the 50 sentence corpus. The user must complete a morphological analysis using the MOR program prior to running the DSS analysis. Additionally, the DSS program has both an automatic and an interactive mode. He then briefly explains how the program works.

Relevance for my study: CLAN is an alternate software program that can conduct DSS analysis against which the program I am assessing can be compared.

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.

Focus: In recent years, modern neuro-linguistic programming (NLP) techniques have been used for parsing child language transcripts for syntactic analysis; this provides a valuable tool in automated syntactic analysis. Sagae, MacWhinney, and Lavie (2004) developed a syntactic annotation scheme to identify syntactic structures and analysis system that takes sentences and produces a labeled dependency structure representing the grammatical relations (GR). These annotation and analysis systems were used to complete partially automated Index of Productive Syntax (IPSyn) scoring.

Procedure: *Grammatical Relations Analysis.* The GR analysis system involves three main steps: text preprocessing, unlabeled dependency identification, and dependency labeling. In text preprocessing, samples are converted in to the CHAT transcription format. CLAN tools (MacWhinney, 2000) are used to remove disfluencies, retracings, and repetitions. Each sentence is run through the MOR morphological analyzer (MacWhinney, 2000) and the POST part-of-speech tagger (Parissee & Le Normand, 2000). In the second step, text is parsed to obtain the unlabeled dependencies. This can be completed by processing constituent trees with a set of rules to determine the lexical heads of constituents. After the unlabeled dependencies have ben identified, they are labeled with GR labels in step three. Automated systems are employed in steps two and three, with the systems first receiving training with practice items. The classifier is

91.4% accurate on labeling dependencies on the same 2,018 words used to evaluate unlabeled accuracy. The overall labeled dependency accuracy, including unlabeled dependencies obtained with the Charniak parser and the labels obtained with the classifier, was 86.9%.

Automated IPSyn. This syntactic analysis of grammatical relations in transcripts allows for fully automated IPSyn computations with a level of reliability comparable to human scoring. The author's automated IPSyn program, which draws on part-of-speech (POS) and morphological analysis as well as GR information, was evaluated compared to Computerized Profiling (CP; Long, Fey, & Channell, 2004), which relies solely on POS and morphological analysis. The automated IPSyn program was evaluated in two ways. The first was *Point Difference*, which is calculated by taking the difference between scores obtained manually and automatically. The second evaluation method is *Point-to-Point Accuracy*, which is calculated by counting how many decisions (identifying the presence of absence of language structure) were made correctly and dividing by the total number of decisions. Point Difference demonstrates how close the automatically and manually produced scores are while Point-to-Point Accuracy reflects the overall reliability of each scoring decision. Two sets of transcripts were obtained from two different child language research groups. The first set (A) contained 18 transcripts from children ranging between two and three years of age; these were scored manually. The second set (B) contained 23 transcripts from children ranging between eight and nine years of age; these were initially automatically scored by CP and then manually corrected by researchers. All samples were also scored using the author's automated IPSyn program for comparison.

Results: The scores computed automatically from the author's automated IPSyn program (GR) were very similar to the manually computed scores (HUMAN). The average point difference between the GR and HUMAN scores was 3.3 with a minimum point difference score of zero and

a maximum of 12. There was no clear trend on whether the automated scores were higher or lower than the manual scores. The average point difference between HUMAN and the CP scores was 8.3 with a minimum of zero and a maximum of 21. Additionally, GR was more accurate with older children, who have more syntactically complex utterances, than CP. The mean point-to-point accuracy between GR and HUMAN over the 41 transcripts was 92.8% with the lowest agreement on a transcript falling at 88.5%. In the original IPSyn reliability study (Scarborough, 1990), point-to-point measurements among human scorers was 94% with a minimum agreement of 90%. An error analysis found that four of the 56 IPSyn structures account for almost half of the system errors; these include items S11 (propositional complements), V15 (copula, modal, or aux for emphasis or ellipsis), S16 (relative clauses), and S14 (bitransitive predicates).

Discussion: The GR automated IPSyn calculation examined in this study is similar to manual scoring in both point difference and point-to-point accuracy. Additionally, the authors demonstrated that the GR analysis is superior in terms of accuracy to the other automated IPSyn program available, CP. This demonstrated the value of automated GR assignment to child language research. However, improvements could be made to the identification of specific GRs.

Relevance for my study: In this study, Sagae, Lavie, and MacWhinney introduce the point difference method for comparing manually and automatically computed scores. This method of scoring was used to evaluate the accuracy of the beta version of the software I am evaluating.

Smith, J. M., DeThorne, L. S., Logan, A. R. L., Channell, R. W., & Petrill, S. A. (2014). Impact of prematurity on language skills at school age. *Journal of Speech, Language, and Hearing Research, 57,* 901-916. doi: 10.1044/1092-4388(2013/12-0347)

Focus: Various studies have suggested a link between premature birth and language ability. However, most of these studies focus on standardized test scores. This study examined the comparative language abilities of a group of twins born prematurely versus a control group of twins born full-term using both discourse-level samples and standardized test data.

Participants: Data was drawn from the Western Reserve Reading Project (WRRP), which includes 368 same-sex twin pairs, primarily from Ohio. Participation in the study began in kindergarten or first grade. Children were selected from the premature group based on very low birth weight (less than 1,500 g) or prematurity (born at 32 weeks gestation or earlier), which included a group of 57 children (38 girls and 19 boys). A control group of children born at 37+ weeks gestation was matched for gender, age, highest level of education for the primary caregiver, and race/ethnicity was selected.

Procedure: Families were visited in their homes annually by a pair of WRRP examiners at the average ages of 7, 8, and 10 years old. Evaluators collected conversational language samples, narrative language samples, and standardized tests. Semantic measures include number of different words (NDW)/number of total words (NTW), which were calculated in SALT, word-frequency analysis, morphologically complex words, and literate language elements of adverbs and metacognitive verbs. Syntactic measures include mean length of utterance in C-units, conjunction analyses, elaborated noun phrases (ENPs), developmental sentence scoring (DSS), and conversion of frequency counts to density measures. Norm-referenced measures included the short form of the Stanford-Binet Intelligence Scale, the Oral Narration portion of the TNL, and selected subtests of the Clinical Evaluation of Language Fundamentals – Fourth Edition (CELF-4).

Results: At all three visits, the premature group produced all target structures less frequently than the control group; in some cases, the differences were small. Additionally, the control group scored better than the premature group on the standardized test results, although both sets of scores were within the normal range. The effect of child gender, breastfeeding duration, and parental education were analyzed; however, only parental education showed a significant effect. This effect was most pronounced at Year 3.

Discussion: This study found that school-age children born prematurely perform less well on standardized tests than full-term peers. However, the means for the premature children were in the lower end of normal, rather than below. Language sample measures did not demonstrate a statistical significance between groups; however, the premature group demonstrated consistently lower scores. Additionally, where every parent had at least a high school diploma, the decreased language ability was not due to lower levels of parental education. There was no significant effect for either gender or breastfeeding on language or cognitive outcomes.

Relevance for my study: One of the measures used to evaluate children's language abilities in this study was Developmental Sentence Scoring (DSS). This demonstrates that, although the clinical use of DSS may have declined, it is still used in the research setting. Therefore, an automated DSS program would be beneficial for researchers.

Westerveld, M. F., & Claessen, M. (2014). Clinician survey of language sampling practices in Australia. *International Journal of Speech-Language Pathology. 16*(3), 242-249. doi: 10.3109/17549507.2013.871336

Focus: Spontaneous language samples are an important element of clinical practice; sample analysis can confirm and complement standardized test results and assist in assessment,

intervention planning, and outcome measurement. Speech-language pathologists (SLPs) from around Australia participated in an online survey to evaluate clinicians' practices and opinions regarding: a) the purpose of language sample elicitation, b) elicitation methods, c) transcription, and d) analysis.

Participants: Two hundred and fifty-seven SLPs from around Australia responded to the survey; 80.8% of the respondents were members of Speech Pathology Australia.

Procedure: Survey questions were designed based on research questions, a review of literature, consultation with clinical SLPs, and previous surveys based by Hux et al. (1993) and Kemp and Klee (1997). The final survey consisted of 29 questions in four sections covering demographic information, assessment measures, spontaneous language sampling, and language sample analysis. All members of Speech Pathology Australia (SPA) who worked with children, aged 0–18 were invited to participate through an Association email. A message was also posted on the phonological therapy listserve in order to recruit SLPs who were not members of Speech Pathology Australia. In addition, authors and colleagues from each state personally emailed clinicians to make them aware of the survey and encourage them to pass the invitation on to potential participants.

Results: 97.3% of respondents reported using standardized assessments when assessing children with suspected language impairment; 90.8% collected spontaneous language samples. Language samples were elicited for a range of purposes, including screening (68.8%), diagnosis (78.8%), remediation (61.5%), and post-intervention (54.6%). Among the 8.2% of respondents who reported not collecting language samples, time constraints, lack of training, and lack of computer hardware/software were reported as the main reasons for not collecting language samples. The majority of respondents reported using an informal procedure (87%) with 62% reporting usage

of standardized tests. 56% of respondents indicated recording language samples and 49% reported transcribing the sample in real-time. The reported typical length in utterance of the language samples was between 0 and 500 utterances with an average between 16 and 23 utterances.

Discussion: The results of this study were consistent with previous research (Kemp & Klee, 1997) in confirming that SLPs value the importance of collecting spontaneous language samples. The authors did note, however, that since participation in the survey was by open invitation, the sample may potentially be biased towards LSA. Additionally, the results show that the majority of the respondents use informal procedures or standardized tests to collect language samples. Clinicians also varied their elicitation procedures depending on their clients' age.

Relevance for my study: This study confirms the importance of language sample analysis in clinical practice and the widespread usage of language sample analysis among practicing clinician, suggesting that an automated language analysis, such as the one being evaluated in my study, may be well received.

Appendix B: DSS Scoring Chart (from Lee, 1974)

SCORE	INDEFINITE PRONOUNS OR NOUN MODIFIERS	PERSONAL PRONOUNS	MAIN VERBS	SECONDARY VERBS
1	it, this, that	1st and 2nd person: I, me, my, mine, you, your(s)	A. Uninflected verb: *I see you.* B. copula, is or 's: *It's red.* C. is + verb + ing: He *is coming.*	
2		3rd person: he, him, his, she, her, hers	A. -s and -ed: *plays, played* B. irregular past: *are, saw* C. Copula: *am, are, was, were* D. Auxiliary *am, are, was, were*	Five early-developing infinitives: I wan*na see* (want *to see*) I'm gon*na see* (going *to see*) I got*ta see* (got *to see*) Lemme [to] see (let me [*to*] *see*) Let's [to] play (let [us *to*] *play*)
3	A. no, some, more, all, lot(s), one(s), two (etc.), other(s), another B. something, somebody, someone	A. Plurals: we, us, our(s), they, them, their B. these, those		Non-complementing infinitives: I stopped *to play.* I'm afraid *to look.* It's hard *to do* that.
4	nothing, nobody, none, no one		A. can, will, may + verb: *may go* B. Obligatory do + verb: *don't go* C. Emphatic do + verb: I *do see.*	Participle, present or past: I see a boy *running.* I found the toy *broken.*
5		Reflexives: myself, yourself, himself, herself, itself, themselves		A. Early infinitival complements with differing subjects in kernels: I want you *to come.* Let him [*to*] *see.* B. Later infinitival complements: I had *to go.* I told him *to go.* I tried *to go.* He ought *to go.* C. Obligatory deletions: Make it [*to*] go. I'd better [*to*] go. D. Infinitive with wh-word: I know what *to get.* I know how *to do* it.
6		A. Wh-pronouns: who, which, whose, whom, what, that, how many, how much I know *who* came. That's *what* I said. B. Wh-word + infinitive: I know *what* to do. I know *who(m)* to take.	A. could, would, should, might + verb: *might come, could be* B. Obligatory does, did + verb C. Emphatic does, did + verb	
7	A. any, anything, anybody, anyone B. every, everything, everybody, everyone C. both, few, many, each, several, most, least, much, next, first, last, second (etc.)	(his) own, one, oneself, whichever, whoever, whatever Take *whatever* you like.	A. Passive with *get*, any tense Passive with *be*, any tense B. must, shall + verb: *must come* C. have + verb + en: *I've eaten* D. have got: I'*ve got* it.	Passive infinitival complement: With *get:* I have *to get dressed.* I don't want *to get hurt.* With *be:* I want *to be pulled.* It's going *to be locked.*
8			A. have been + verb + ing had been + verb + ing B. modal + have + verb + en: *may have eaten* C. modal + be + verb + ing: *could be playing* D. Other auxiliary combinations: *should have been sleeping*	Gerund: *Swinging* is fun. I like *fishing.* He started *laughing.*

NEGATIVES	CONJUNCTIONS	INTERROGATIVE REVERSALS	WH-QUESTIONS
it, this, that + copula or auxiliary is, 's, + not: It's *not* mine. This is *not* a dog. That is *not* moving.		Reversal of copula: *Isn't it* red? *Were they* there?	
			A. who, what, what + noun: *Who* am I? *What* is he eating? *What book* are you reading? B. where, how many, how much, what . . . do, what . . . for *Where* did it go? *How much* do you want? *What* is he *doing?* *What* is a hammer *for?*
	and		
can't, don't		Reversal of auxiliary be: *Is he* coming? *Isn't he* coming? *Was he* going? *Wasn't he* going?	
isn't, won't	A. but B. so, and so, so that C. or, if		when, how, how + adjective *When* shall I come? *How* do you do it? *How big* is it?
	because	A. Obligatory do, does, did: *Do they* run? *Does it* bite? *Didn't it* hurt? B. Reversal of modal: *Can you* play? *Won't it* hurt? *Shall I* sit down? C. Tag question: It's fun, *isn't it?* It isn't fun, *is it?*	
All other negatives: A. Uncontracted negatives: I can *not* go. He has *not* gone. B. Pronoun-auxiliary or pronoun-copula contraction: I'm *not* coming. He's *not* here. C. Auxiliary-negative or copula-negative contraction: He was*n't* going. He has*n't* Been seen. It couldn*'t* be mine. They are*n't* big.			why, what if, how come how about + gerund *Why* are you crying? *What if* I won't do it? *How come* he is crying? *How about* coming with me?
	A. where, when, how, while, whether (or not), till, until, unless, since, before, after, for, as, as + adjective + as, as if, like, that, than I know *where* you are. Don't come *till* I call. B. Obligatory deletions: I run faster *than* you [run]. I'm *as big as* a man [is big]. It looks *like* a dog [looks] C. Elliptical deletions (score 0): That's *why* [I took it]. I know *how* [I can do it]. D. Wh-words + infinitive: I know *how* to do it. I know *where* to go.	A. Reversal of auxiliary have: *Has he* seen you? B. Reversal with two or three auxiliaries: *Has he been* eating? *Couldn't he have* waited? *Could he have been* crying? *Wouldn't he have been* going?	whose, which, which + noun *Whose* car is that? *Which book* do you want?