2015-05-01

Precoding and the Accuracy of Automated Analysis of Child Language Samples

Rachel Christine Winiecke
Brigham Young University - Provo

Precoding and the Accuracy of Automated Analysis of Child Language Samples

Rachel Christine Winiecke

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
Kristine Tanner
Shawn L. Nissen

Department of Communication Disorders

Brigham Young University

April 2015

ABSTRACT

Precoding and the Accuracy of Automated
Analysis of Child Language Samples

Rachel Christine Winiecke
Department of Communication Disorders, BYU
Master of Science

Language sample analysis is accepted as the gold standard in child language assessment. Unfortunately it is often viewed as too time consuming for the practicing clinician. Over the last 15 years a great deal of research has been invested in the automated analysis of child language samples to make the process more time efficient. One step in the analysis process may be precoding the sample, as is used in the Systematic Analysis of Language Transcripts (SALT) software. However, a claim has been made (MacWhinney, 2008) that such precoding in fact leads to lower accuracy because of manual coding errors. No data on this issue have been published. The current research measured the accuracy of language samples analyzed with and without SALT precoding. This study also compared the accuracy of current software to an older version called GramCats (Channell & Johnson 1999). The results presented support the use of precoding schemes such as SALT and suggest that the accuracy of automated analysis has improved over time.

Keywords: language sample, automated analysis, tagging, language software

ACKNOWLEDGEMENTS

I would like to thank the all the supportive staff and students here at Brigham Young University, especially the members of my thesis committee: Dr. Nissen and Dr. Tanner. Thank you for all your advice and positivity. I would especially like to thank Dr. Channell for his kindness and unfailing support. This thesis would not have been possible without his mentoring and teaching. Lastly I would like to thank my family for supporting me through every step up to this point.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

DESCRIPTION OF THESIS STRUCTURE

This thesis is part of a larger research project, and portions of this thesis may be published as part of articles listing the thesis author as a co-author. The body of this thesis is written as a manuscript suitable for submission to a peer-reviewed journal in speech-language pathology. An annotated bibliography is presented in the Appendix.

Introduction

For over 40 years, the analysis of samples of children's language has been a valuable assessment tool in speech-language pathology. These samples are transcribed recordings of the linguistic interaction of a child with the clinician, a parent, or a peer. The analysis of these transcripts allows the study and description of a child's abilities to formulate and express sentences, and the different varieties or methods of analysis allow clinical goals to be set and progress associated with therapy to be monitored (Ebert & Scott, 2014). However, the collection, transcription, and especially the analysis of these samples makes great demands on a clinician's time (Heilmann, Nockerts, & Miller, 2010; Long, 2001), and many clinicians lack the expertise to perform these analyses quickly and accurately (Long, 1996). Accordingly, the last 20 years have seen the development and possibly the improvement of software to automate and thus speed up various methods and procedures for the clinical analysis of samples of children's language. These clinical analyses rely on an initial probability-based analysis wherein the grammatical categories of the words in the sample are identified (Hassanali, Liu, Iglesias, Solorio, & Dollaghan, 2014; Long & Channell, 2001). However, a detailed assessment of the improvement, if any, over time in the accuracy of software for this foundational analysis has not yet been made. In addition, questions about the effects of formatting on the accuracy of the analysis have arisen but evidence regarding this issue is lacking. The present study quantifies improvement in the accuracy of automated analysis and presents data regarding the effect of formatting on this accuracy.

Language samples aim to collect a child's spontaneous, unscripted responses and productions to a conversation, a story, or another descriptive task. This unscripted language behavior is in contrast to the restricted naming, completion, and repetition used as items of

published, standardized tests of language. Language samples allow a less-biased description of the language of children speaking different varieties of English (Fiestas & Peña, 2004; Gutierrez-Clellen, Restrepo, Bedore, Peña, & Anderson, 2000; Gutierrez-Clellen & Simon-Cereijido, 2009). A variety of analysis procedures have been developed to examine different aspects of the language used by the child, and to save time, several of these procedures have been automated. These automated procedures include the Language Assessment Screening Profile (LARSP; Crystal, Garman, & Fletcher, 1989), automated by Computerized Profiling software (CP; Long, Fey, & Channell, 2006) and studied by Long and Channell (2001); Developmental Sentence Scoring (DSS; Lee, 1974) automated by CP and by the Child Language Analysis software (CLAN, MacWhinney, 2010) and studied by Long and Channell (2001) and by Channell (2003); the Index of Productive Syntax (IPSyn; Scarborough, 1990) automated by CP, CLAN, and AC-IPSyn (Hassanali et al., 2014) and studied by Hassanali et al., Long and Channell (2001) and by Sagae, Lavie, and MacWhinney (2005); and the Mean Length of Utterance (MLU; Brown, 1973), tabulated by the Systematic Analysis of Language Transcripts software (SALT; Miller, Andriacchi, & Nockerts, 2011) but automated by CLAN and CP.

In automated language sample analysis, one essential, foundational step is the coding of a sample's words according to their grammatical categories, also known as syntactic categories or parts of speech. This is done by using an algorithm called a Hidden Markov Model (HMM; Jurafsky & Martin, 2000; Manning & Schütze, 1999). A Markov model calculates the probabilities of different states in a transition (which in the present context would be the sequence of words in an utterance) and predicts the probability of some future word without looking too far into the past (Jurafsky & Martin, 2000), which in the present context would be the previous words in the utterance. The length of this look back is described as an N-gram

model, wherein a bigram bases its prediction on one step before and a trigram bases its prediction on the two steps before. A hidden Markov model is one in which the sequence of the steps is not known but is estimated using probabilities (Manning & Schütze, 1999).

To increase generality, grammatical categories rather than particular words are used as the steps in the model. This use of categories brings in a second factor: the likelihood that a word will be of that grammatical category. DeRose (1988) found that although only 11% of the words in English have more than one possible grammatical category, such as the fact that the word *can* can be an auxiliary verb, a noun, or a main verb, these grammatically ambiguous words are used so frequently as to make up 40% of the tokens in a large set of texts. So, scanning the data in a large corpus, *can* will be used as an auxiliary verb about 90% of the time, a noun about 9%, and a main verb less than 1% of the time. The best (most likely) grammatical category for a given word is thus the product of how likely that category is, given the one or two possible categories before it, and how likely it is that the word is that category. The number of possible category codings for an utterance increases exponentially. If, for example, an utterance is 25 words long, and each word has an average of two possible categories, the number of possible category codings would be 2^{25}, which equals 33,554,432. Various optimization algorithms attempt to make calculation of the correct category coding sequence for an utterance more efficient (DeRose, 1988; Jurafsky & Martin, 2000; Manning & Schütze, 1999).

Only two appraisals of the use of HMM-based software with samples of child language have been published. MacWhinney (2008) described the overall accuracy of CLAN's grammatical category coding (using the POST component) as "over 95%" (p. 186). However, data on the performance with particular categories was not detailed. An earlier appraisal had been published by Channell and Johnson (1999), who looked at samples from 30 typically

developing children ages 2;6 to 7;11 and compared the manually-coded version of each sample with the version coded by software they called GramCats. Channell and Johnson found an overall level of accuracy of 95.1%. Data on the accuracy with which each category was correctly coded was also presented; categories ranged from a high of 100% to a low of 0%. The GramCats software was subsequently integrated into the CP software. Long and Channell (2001) determined that some of the mistakes in the LARSP, DSS, and IPSyn analyses made by the CP software resulted from miscodings made by GramCats.

The 15 years since the development of the GramCats software have seen massive improvements in computer capacity and speed, and some of the constraints in place then are absent now. For example, GramCats was limited to a maximum of 640K of random-access memory (RAM) and fit on a single floppy disk. GramCats used a bigram HMM probability model because a trigram model would have required too much RAM. The probabilities used in GramCats were extracted from a corpus of about 5,000 hand-coded child utterances. The consequences of these limitations on the level of grammatical coding accuracy are unknown.

While the accuracy of the dictionaries and the algorithms used by the software are important there are other factors that may influence accuracy of the tagging. The precoding process could be an important step in increasing accuracy. One popular precoding scheme used is the SALT (Miller et al., 2011) scheme, originally created to aid in the calculation of a client's MLU (Brown, 1973). Using a slash ("/") character, the SALT scheme separates morphemes that define aspects of language reflecting grammatical development (Brown, 1973) such as plurals and possessives. A similar scheme for precoding had previously been used in coding for the analysis done by the CLAN software. Without published data to support the claim, the precoding for CLAN analysis was dropped because of claimed higher computer performance

without human coding (MacWhinney, 2008). Precoding can be a time consuming part of the analysis process. Thus far there have been no studies to encourage or discourage the use of precoding as a means of increasing the accuracy of automated grammatical analysis. The present study is interested in comparing the accuracy of samples with SALT precoding to versions of the same samples that have not been coded.

Method

This study used language samples that had been originally collected for different purposes as part of other studies.

Participants

Conversational language samples were collected by three speech-language pathology graduate students from 30 children ranging in age from 2;6 to 7;11 who lived in Provo, Utah. There were three children in each six-month age interval between 2;6 and 7;0 and three between 7;0 and 7;11, and each graduate student sampled one child in each age interval. The children were typically developing, had no history of speech or language impairment, and spoke English as their primary language. The graduate students collected the samples in the children's own apartments using a variety of toys and conversational activities. These samples had been used by Channell and Johnson (1999) and by Long and Channell (2001).

Manual Grammatical Coding

The child utterances in these samples had been grammatically coded by hand as part of the Channell and Johnson (1999) study. The inter-coder reliability of this manual grammatical word category coding was found to range from 97% to 98% when computed on a per-word basis.

The grammatical categories used in coding the files were those of Channell and Johnson (1999), which had been adapted from the set used to tag the Brown University Corpus (Francis

& Kučera, 1982). Channell and Johnson explained that these adaptations had been motivated by distinctions important for the study of child language and to allow the use of this coding to form the basis other analyses such as DSS or LARSP.

Software

The software used in the present study, gc5, is a newer version of the GramCats software. The transcription format for the samples to be analyzed is similar to that of other child language analysis software. Like GramCats or CLAN, gc5 expects each utterance in the sample to be on a separate line. Except for proper nouns and the pronoun *I*, all words are entered in lower case characters. Mazes (interjections, repetitions, revisions, etc.) are put into parentheses and otherwise ignored. Multiword proper names have the words connected (e.g., *Salt_Lake_City*). Each utterance also ends with a terminal punctuation character.

The gc5 software uses a trigram HMM to assign grammatical codes to the words in the samples. As part of this HMM, two types of probability are used. The first probability is the likelihood of the different code options for a single word. In the Brown University Corpus (Francis & Kučera, 1982) for example, the word *stand* is about nine times more likely to be a verb (*we stand in the corner*) than a noun (*the stand is in the corner*).

The second probability is the likelihood that a given code will follow the two previous grammatical codes in an utterance. For example, a code of *noun* is far more likely to immediately follow the codes *determiner* and *adjective* than is the code *modal auxiliary*. The program thus reconciles these two forms of probability to determine the most likely sequence of codes for a set of words. The use of these two forms of probability allows N-gram HMM models to work even when faced with the "messy" data implicit in unedited oral child language (Channell & Johnson, 1999).

The gc5 program has built-in dictionaries containing word code option frequencies and code-transition frequencies that were extracted from a manually coded training corpus. The training corpus contained samples from Abe (from ages 3;0 to 5;0; Kuczaj, 1976), Sarah (ages 3;2 to 4;11; Brown, 1973), 48 children (ages 2;10 to 5;7) engaged in child-child discourse (Garvey, 1979; Garvey & Hogan, 1973), and the first-, third-, and fifth-grade children from Carterette and Jones (1974). The program dictionary also has a large number of verbs and adjectives added (without probability data) which were obtained from lists on the internet. The dictionary does not contain proper nouns (which are transcribed with an initial capital letter, if the user wants them correctly coded). Unknown words are guessed as nouns; testing of earlier versions of the program suggested this strategy because the unknown words encountered were mostly nouns such as *iPod* or *wii*. Thus in total the training corpus contained about 20,000 manually coded child utterances.

Procedure

The child language samples were run through the program twice; the first time with not SALT coding and the second time with SALT coding. Each word in each child utterance in each sample was coded by gc5 as to its grammatical category, and this automated coding was compared to the manual coding of that word using a utility program. Accuracy was defined as the percentage of agreement between automated and manual coding and was calculated on a word-by-word basis and an utterance-by-utterance basis. Similarly, the percentage of agreement in each grammatical category was calculated. Pearson's correlations were used to examine relationships between accuracy levels and factors such as the length of the child's language sample or age of the child. Over all accuracy, tag accuracy, and participant-by-participant accuracy were compared between the SALT coded and non-SALT coded samples.

Results

A major goal of this research project was to discover if SALT-coded language samples affected the accuracy of tagging software. Tables 1 and 2 present the overall tag accuracy of SALT-coding and non-SALT coding; information such as frequency, percentage of confusions, and types of confusions for each individual tag are presented as well.

Overall tagging accuracy for SALT-coded files (Table 2) was slightly higher than for non-SALT-coded files (Table 1). Non-SALT-coded files had an overall accuracy of 96.54% at the per-tag level and 84.18% at the per-utterance level. SALT-coded files had an overall accuracy of 96.84% at the per-tag level and 85.4% the per-utterance level. These overall differences were statistically significant both at the per-tag, $t(29) = 3.98$; $p = .0004$, and the whole-utterance levels, $t(29) = 4.36$, $p = .0001$. If only the tags that were used fairly often (i.e., 30 or more times) are considered, for non-SALT coded files the tag accuracy ranged from 0% to 100% and for SALT-coded files the accuracy ranged from 67% to 100%.

As seen in Table 1 and 2 the grammatical categories with over 30 exemplars that had above 90% accuracy included verbs, prepositions, and nouns. The categories that were moderately accurate, that is between 80-90%, included several auxiliary verbs. Categories such as intensifiers and adverbs fell below 80% accuracy. These patterns were true for both SALT-coded and non-SALT coded files. As can be seen by comparing Table 1 and Table 2, the SALT-coded files were found to have higher accuracy on a number of categories such as possessive *'s* and auxiliary *be*. As might be expected, the grammatical categories not subject to SALT coding did not show any difference in the accuracy levels between SALT-coded and SALT-uncoded language sample files.

Table 1

Grammatical Tagging Accuracy Without SALT Coding

Tag	Description	N	%	Confusions (%)
<*	negatives *not, n't*	516	99	
<$	possessive *'s*	104	0	BCZ (99) XBZ (1)
BC	copula (be)	74	100	
BCD	copula (were)	32	100	
BCDZ	copula (was)	136	99	XBDZ (1)
BCG	copula (being)	4	100	
BCM	copula (am)	70	100	
BCN	copula (been)	3	100	
BCR	copula (is)	192	100	
BCZ	copula (are)	971	99	XBZ (1)
CC	clausal conjunct.	1390	100	
CS	sub. conjunct.	13	69	IN (23) PRL (7)
D$	possessives	770	100	
DA	articles	1658	100	
DCN	cardinal numbers	140	93	PN(7)
DD	demonst. singular	330	99	
DDS	demonst. plural	61	100	
DN	indefinite det.	326	97	DPA (1) PN (2)
DON	ordinal number	24	75	NN (17) RB (8)
DPA	predeterminer	45	100	
DWN	noun clause det.	9	56	PWN (22) PWQ (11) RBN (11)
DWQ	interrogative det.	20	40	DN (5) DWN (5) PWQ (35) RBQ (15)
DWX	exclamative det.	5	60	RBN (40)
EX	existential *there*	120	93	RB (7)
IN	preposition	1970	98	RP (1)
JJ	adjectives	852	94	NN (3) RB (2)
JJR	comparative	28	100	
JJT	superlative	6	100	
NN	singular noun	3104	99	JJ (1)
NNS	plural noun	873	96	PN (3)
NP	proper noun	664	100	
P$	proper nouns	102	86	D$ (7) VB (7)
PD	demonstr. singular	846	98	DD (1)
PDS	demonstr. plural	113	97	DDS (3)
PI	indefinite pronoun	96	100	
PL	reflexive singular	15	100	
PLS	reflexive plural	3	100	

PN	quantifiers	651	94	DCN (2) DN (2) NN (1) RB (1)
PO	object case pro.	1244	99	PZ (1)
PRL	relative pronoun	45	100	
PS	subject pronoun	3124	99	
PWN	nominal clause pro.	80	98	PWQ (2)
PWQ	interrogative pro.	159	93	PN (1) PWN (6)
PZ	3rd person pro.	1185	99	PO(1)
RB	adverb	1994	95	EX (1) JJ (1) NN (1) RQL (1)
RBN	noun clause adv.	106	84	RBQ (2) RBS (14)
RBQ	questions wh-adv.	254	87	RBS (13)
RBR	comparative adv.	23	61	JJR (17) PN (17) RBR (23) VB (2)
RBS	subordinating adj.	420	93	RB(6)
RP	verb particle	596	87	IN (12)
RQL	qualifier	328	79	DD (1) JJ (1) NN (3) PD (1) PN (5) RB (6) RBR (1) UH (4)
RQLP	post qualifier	10	90	RQL (10)
TO	infinitive marker	559	98	IN (2)
UH	interjection	12	100	
VB	verb	3661	99	NN (1)
VBD	past	1043	90	VB (8)
VBG	present participle	472	94	NN (5) RQL (1)
VBN	past participle	183	64	JJ (23) NN (3) VB (2) VBD (7)
VBZ	3rd pres. singular	436	98	NNS (1) XDZ (1)
VPO	verb + pronoun *let's*	94	100	
VTO	catenative	258	100	
XB	aux *be*	20	85	BC (15)
XBD	aux *were*	13	92	BCD (8)
XBDZ	aux *was*	43	88	BCDZ (12)
XBM	aux *am*	129	97	BCM (3)
XBN	aux *been*	11	100	
XBR	aux *are*	135	93	BCR (7)
XBZ	aux *is*	182	85	BCZ (15)
XD	aux *do*	308	93	VB (93)
XDD	aux *did*	111	86	VBD (14)
XDZ	aux *does*	89	96	VBZ (4)
XG	aux *get*	19	42	VB (58)
XGD	aux *got*	5	100	
XGG	aux *getting*	1	100	
XGZ	aux *gets*	1	0	VBZ(100)
XH	aux *have*	40	78	VB (22)
XHZ	aux *has*	17	0	BCZ (6) VBZ (6) XBZ (88)
XM	modal	532	99	
XM*	modal + neg.	143	100	

Table 2

Grammatical Tagging Accuracy With SALT Coding

Tag	Description	N	%	Confusions (%)
<*	negatives *not, n't*	514	99	
<$	possessive *'s*	104	69	BCZ (30) XBZ (1)
BC	copula be	74	100	
BCD	copula were	32	100	
BCDZ	copula was	136	100	
BCG	copula being	4	75	VBG (25)
BCM	copula am	70	100	
BCN	copula been	3	100	
BCR	copula are	192	100	
BCZ	copula is	971	99	
CC	clausal conjunct.	1390	100	
CS	subord. conjunct.	13	69	IN (23) PRL (8)
D$	possessives	770	100	
DA	articles	1657	99	
DCN	cardinal numbers	140	93	PN (7)
DD	demonst. singular	330	99	
DDS	demonst plural	61	100	
DN	indefinite det.	326	96	DPA (1) PN (2)
DON	ordinal number	24	75	NN (17) RB (8)
DPA	predeterminer	45	100	
DWN	noun clause det.	9	56	PWN (22) PWQ (11) RBN (11)
DWQ	interrogative det.	20	25	DN (5) DWN (5) PWQ (35) RBQ (5)
DWX	exclamative det.	5	60	RBN (40)
EX	existential *there*	120	93	RB (7)
IN	preposition	1969	98	RP (1)
JJ	adjectives	852	94	NN (3) RB (2)
JJR	comparative	28	100	
JJT	superlative	6	100	
NN	singular noun	3101	99	JJ (1)
NNS	plural noun	872	96	PN (3) VBZ (1)
NP	proper nouns	659	100	
P$	possessive pro.	102	86	D$ (7) VB (7)
PD	demonstr. singular	846	98	DD (1)
PDS	demonstr. plural	113	98	DDS (2)
PI	indefinite pronoun	96	100	
PL	reflexive singular	15	100	
PLS	reflexive plural	3	100	

PN	quantifiers	651	94	DCN (2) DN (2) NN(1) RB (1)
PO	object case pro.	1244	98	PZ (1)
PRL	relative pronoun	45	100	
PS	subject pronoun	3124	100	
PWN	nominal clause pro.	80	98	PWQ (2)
PWQ	interrogative pro.	159	93	PN (1) PWN (6)
PZ		1185	99	PO (1)
RB	adverb	1994	96	EX (1) NN (1)
RBN	noun clause adv.	106	84	RBQ (2) RBS (14)
RBQ	question wh-adv.	254	87	RBS (13)
RBR	comparative adj.	23	61	JJR (17) PN (7) VB (4)
RBS	subordinating adv.	420	93	RB (5)
RP	verb particle	596	87	IN (12)
RQL	qualifier	328	79	DD (1) JJ (1) NN (3) PD (1) PN (5) RB (6) RBR (1) UH(4)
RQLP	post qualifier	10	90	RQL (10)
TO	infinitive marker	559	99	IN (1)
UH	interjection	13	100	
VB	verb	3661	99	NN (1)
VBD	past	1043	90	NN (1) VB (9)
VBG	present participle	472	96	NN (4)
VBN	past participle	183	67	JJ (21) NN (3) VB (2) VBD (7)
VBZ	3rd present sing.	436	99	XDZ (1)
VPO	verb + pronoun *lets*	94	100	
VTO	catenative	258	100	
XB	aux *be*	20	90	BC (10)
XBD	aux *were*	13	92	BCD (8)
XBDZ	aux *was*	43	88	BCDZ (12)
XBM	aux *being*	129	96	BCM (4)
XBN	aux *been*	11	100	
XBR	aux *are*	135	93	BCR (7)
XBZ	aux *is*	182	86	BCZ (14)
XD	aux *do*	308	93	VB (7)
XDD	aux *did*	111	86	VBD (14)
XDZ	aux *does*	89	92	VBZ (8)
XG	aux *get*	19	42	VB (58)
XGD	aux *got*	5	100	
XGG	aux *getting*	1	100	
XGZ	aux *gets*	1	100	
XH	aux *have*	40	78	VB (22)
XHZ	aux *has*	17	88	BCZ (6) VBZ (6)
XM	modal	530	99	
XM*	modal + neg.	145	100	

Table 3 shows the accuracy of tagging for the individual 30 participants on both the SALT-coded files and the non-SALT coded files. Pearson's correlations were used to examine relationships among the number of utterances in a child's sample, the mean length of utterance (MLU; Brown, 1973) of the child, and the accuracy of tagging both SALT-coded and uncoded files at both the single-tag and the whole-utterance level. Neither the number of utterances nor the MLU were related to the accuracy of coding when looking at tag-level accuracy for either the SALT-coded or the uncoded files ($p > .05$). However, correlations existed between the number of utterances and the MLU, r (28) = .387; p = .035, and negative correlations were observed between MLU and whole utterance accuracy levels for both the SALT-coded files, r (28) = -.669; $p < .0001$ and the uncoded files, r (28) = -.651, $p <.0001$.

Discussion

The aim of this study was to examine the accuracy of SALT-coded tagging to non-SALT-coded tagging and to compare the older version of this software, which is GramCats, to the newer version, which is gc5. The present study found that overall tagging accuracy was slightly better for SALT-coded tagging. This difference in accuracy was found to be statistically significant. SALT-coding never did worse than non-SALT coded tagging, thus the idea that a computer does better without this extra coding (MacWhinney, 2008) seems to lack support. Other variables that may have affected the accuracy were explored. The relationship between accuracy and the MLU was examined. There was a significant negative correlation between whole-utterance tagging and MLU in both the SALT-coded files and the non-SALT-coded files. This suggests that as MLU increases or decreases the accuracy of the tagging on individual words does not change, but as MLU increases the accuracy of whole-utterance tagging falls. There were no correlations between the length of the sample and the accuracy observed.

Table 3

Accuracy Measures for Each Participant Sample

Samples	mlu	n_utts	GC tag%	SALT tag%	GC utt%	SALT utt%
Aaron	3.32	217	96.77	96.77	88.48	88.48
Aimee	6.22	208	96.67	97.12	77.88	80.29
Alisha	4.35	207	96.32	97.70	85.51	89.86
Amber	4.77	214	98.08	98.16	89.25	89.72
Ambree	4.59	198	96.22	96.30	81.82	82.32
Andrus	3.55	193	97.03	97.68	88.08	90.67
Ashley	3.54	196	96.18	96.42	85.20	86.22
Ashley B	3.67	189	95.56	95.56	84.13	84.66
BJ	4.51	196	97.60	97.49	88.78	88.27
Christine	4.53	192	98.45	98.65	92.19	93.23
Clarissa	7.16	212	96.85	96.95	77.36	76.89
Cody	6.92	216	96.58	97.01	77.78	80.56
Elizabeth	4.18	190	96.05	97.36	85.26	89.47
Heather B	5.40	169	96.49	96.94	83.43	85.21
Heather C	3.95	194	96.71	97.59	86.08	89.69
Jack	5.63	214	95.46	95.82	78.50	79.91
Jarom	4.75	185	97.95	98.05	90.27	90.81
Katie	4.59	199	96.05	96.14	79.40	79.90
Kevin	3.53	193	96.57	96.90	86.01	87.05
Kile	3.43	183	94.74	94.74	85.79	85.25
Michael	4.72	198	96.56	96.56	83.33	83.33
Patrick	4.07	213	96.43	96.62	84.98	85.45
Rebecca	4.17	189	95.82	95.82	82.01	82.01
Rebekah	5.09	201	97.78	97.62	87.56	87.56
Sarah	6.09	212	95.77	96.66	70.75	75.83
Scott	3.59	199	95.57	96.01	83.92	85.93
Talon	3.28	196	97.10	97.10	91.33	91.33
Tavida	4.64	203	96.59	96.77	84.73	86.21
Tiffany	5.09	221	95.95	96.14	80.09	81.00
Toinette	4.55	212	96.37	96.20	85.38	84.91

The study presented here compares most closely to the Channell and Johnson (1999) article. Since the 1999 article there have been many advances in computer capabilities making the analysis of language using software a much smoother process. The types of probabilities used in this article were the same ones that were employed in the Channell and Johnson study, in that the current study used transitional tag probabilities and relative tag probabilities. The dictionary of about 20,000 utterances was not the dictionary used by Channell and Johnson in the 1999 article, so the numerical probabilities are different. One major difference in tag selection is the process of selecting a tag for a word not included in the original dictionary. Because unknown words were manually added to the dictionary in the GramCats article, the dictionary did not have to make a decision about words that were not known. Past research on software indicates that most unknown words are nouns. The gc5 software obtains tags for words not included in the dictionary by automatically assigning them to the category of noun.

This study made use of the same samples and same tag set found in Channell and Johnson (1999), which allows us to make a stronger comparison of the newer version of the software to the older version. Comparing the non-SALT coded accuracy in this study to Channell and Johnson's (1999) we can see that the gc5 software showed an overall 1.4% increase in accuracy of the tag level, and an overall 6.5% increase at the utterance level. The accuracy increase, as mentioned before, was even slightly higher for the SALT-coded files. The SALT-coded scheme results on gc5 software scored as good or better on 78 out of 85 categories when compared to the GramCats software. Of these 8 categories, 7 categories had more than 30 exemplars. These categories included demonstrative singular, plural noun, past participle, interrogative pronoun, auxiliary *does*, modal, and subordinating adverb. Out of the 7 categories, with exemplars greater than 30, the GramCats software showed only a slight increase in

performance; between 1% and 5%. A few factors may have contributed to this, one factor being the unknown glitches in computer performance. Another may have been the changes in tagging decisions. Despite these small drops in performance, overall improvements have been made in software performance.

The results of this study should be interpreted with some caution because of the limitations to be found. This study used a specific tag scheme that may or may not extend to other tag sets. For example the SALT software uses a 25 tag scheme in its analysis process and the effect of this particular tag set is not known. The sample size was not only a limitation for the Channell and Johnson (1999) article, but remains so for this study. Three children were used at each 6 month interval going from 2;6 to 7;11 for a total of 30 participants. Ideally 30 children at each 6 month interval would be used. These samples were all taken from children who had one or more parent attending Brigham Young University. While there is no demographic data on these samples, given the make-up of the university population it is unlikely there is a variety of dialectic backgrounds. It is also important to mention that these samples were obtained 25 years ago. The results from the data of these children may not mirror young children today. Furthermore, this comparison of accuracy was done on typically developing children. We do not yet have any data to compare with children who have language impairment. Another limitation to be mentioned is that the precoding scheme used was SALT. Other precoding schemes such as those found in CLAN may not have the same results as those found here.

While direct clinical significance is not found in this specific area, as clinicians don't routinely tag each word in a child's language sample as to grammatical category, the present study makes a contribution to our collective understanding of automated language sample analysis. It is not suggested that time be used to take already automated files and recode them

into SALT-coding for higher accuracy. At that point in the analysis process, time can be better spent correcting the errors made by the computer. It can be suggested that files being transcribed for the first time be put directly into SALT-coding. As far as can be expected, adding the coding in as the clinician is transcribing does not add a great deal of time to the process. Future research does support more investigation in the accuracy of precoding schemes in automated software. Such research may include (a) SALT-coding on different tag sets such as those found in the SALT software, (b) looking at other coding schemes such as those found in the CLAN software, and (c) comparing the scheme with children who have language impairment. While there is more to explore in this area of research, this study had contributed a building block to the improvement of more clinically relevant aims. This understanding that coding schemes have the potential to increase the accuracy of automated tagging can help in the building a framework for the improvement of other programs that perform clinically useful analyses such as DSS and LARSP.

References

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

Carterette, E. C., & Jones, M. H. (1974). *Informal speech: Alphabetic and phonemic texts with statistical analyses and tables.* Berkeley, CA: University of California Press.

Channell, R. W. (2003). Automated Developmental Sentence Scoring using Computerized Profiling software. American Journal of Speech-Language Pathology, 10, 180-188. Retrieved November 3, 2014 from http://web.b.ebscohost.com

Channell, R. W., & Johnson, B. W. (1999). Automated grammatical tagging of child language samples. Journal of Speech, Language, and Hearing Research, 42, 727-734. Retrieved November 3, 2014 from http://web.b.ebscohost.com

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

DeRose, S. J. (1988). Grammatical category disambiguation by statistical optimization. Computational Linguistics, 14, 31-39.

Ebert, D. K., & Scott, C. M. (2014). Relationships between narrative language samples and norm-referenced tests scores in language assessments of school aged children. Language Speech and Hearing Services in the Schools, 45, 337-350. Retrieved November 3, 2014 from http://web.a.ebscohost.com

Fiestas, C. E., & Peña, E. D. (2004). Narrative discourse in bilingual children: Language and task effects. Language, Speech, and Hearing Services in Schools, 35, 155-168. doi:10.1044/0161-1461(2004/016)

Francis, W. N., & Kučera, H. (1982) *Frequency analysis of English usage*. Boston, MA: Houghton Mifflin.

Garvey, C. (1979). An approach to the study of children's role play. *Quarterly Newsletter of the Laboratory of Comparative Human Cognition, 1*, 69–73.

Garvey, C., & Hogan, R. (1973). Social speech and social interaction: Egocentrism revisited. *Child Development, 44*, 562-568. doi:10.1111/1467-8624.ep12115586

Gutierrez-Clellen, V. F., Restrepo, M. A., Bedore, L. M., Peña, E. D., & Anderson, R. T. (2000). Language sample analysis in Spanish-speaking children: Methodological considerations. Language, Speech, and Hearing Services in Schools, 31, 88-98. Retrieved November 3, 2014 from http://web.b.ebscohost.com

Gutierrez-Clellen, V. F., & Simon-Cereijido, G. (2009). Using language sampling in clinical assessments with bilingual children: Challenges and future directions. *Seminars in Speech and Language, 30,* 234-245. doi:10.1055/s-0029-1241722

Hassanali, K., Liu, Y., Iglesias, A., Solorio, T., & Dollaghan, C. (2014). Automatic generation of the index of productive syntax for child language transcripts. *Behavior Research Methods, 46,* 254-262. Retrieved November 3, 2014 from http://search.proquest.com.

Heilmann, J., Nockerts, A., & Miller, J. F. (2010). Language sampling: Does the length of the transcript matter? *Language, Speech & Hearing Services In Schools*, *41*, 393-404. doi:10.1044/0161-1461(2009/09-0023)

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice Hall.

Kuczaj, S. J. A. II. (1976). -Ing, -s and -ed: A study of the acquisition of certain verb inflections. Unpublished doctoral dissertation, University of Minnesota.

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

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

Long, S. H. (2001). About time: A comparison of computerized and manual procedures for grammatical and phonological analysis. *Clinical Linguistics and Phonetics, 15*, 399-426. Retrieved November 3, 2014 from http://web.b.ebscohost.com

Long, S. H., & Channell, R. W. (2001). Accuracy of four language analysis procedures performed automatically. *American Journal of Speech-Language Pathology, 10*, 180-188. Retrieved November 3, 2014 from http://web.b.ebscohost.com/

Long, S.H., Fey, M. E., & Channell, R. W. (2006). Computerized Profiling (CP; Version 9.7.0) [Computer software]. Milwaukee, WI: Department of Speech Pathology and Audiology, Marquette University. Retrieved September 17, 2007, from www.computerizedprofiling.org

MacWhinney, B. (2008). Enriching CHILDES for morphosyntactic analysis. In H. Behrens (Ed.), *Trends in corpus research: Finding structure in data,* pp. 165-198. Amsterdam, Holland: Benjamins.

MacWhinney, B. (2010). The CHILDES project: Tools for analyzing talk (electronic edition). Retrieved February 25, 2010 from http://childes.psy.cmu.edu/manuals/clan.pdf

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical language processing*. Cambridge, MA: MIT Press.

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. In K. Knight, H. Ng, & K. Oflazer (Ed.s), Proceedings of the 43rd meeting of the Association for Computational Linguistics (pp. 197-204), Ann Arbor, MI: Association for Computational Linguistics.

Scarborough, H. S. (1990). The index of productive syntax. *Applied Psycholinguistics, 11,* 1-22.

Annotated Bibliography

Channell, R. W. (2003). Automated developmental sentence scoring using computerized profiling software. *American Journal of Speech-Language Pathology, 12,* 369-375. Retrieved November 3, 2014 from http://web.b.ebscohost.com.

Introduction: Developmental sentence soring (DSS) is a clinical tool used by many clinicians to better understand the grammatical complexity of a particular child. DSS is a complex analysis that assigns points to phrases to calculate a norm-referenced score. The effectiveness of software programs like Developmental Sentence Scoring Computer Program (DSSCP) and Child Language Analysis (CLAN) to accurately do a DSS analysis is unclear. The author was interested in the ability of Computerized Profiling (CP) to do a DSS analysis.

Method: Language samples were collected from 48 children; 28 of these children had been diagnosed with language impairment. Each sample taken was both manually coded and run through the CP software. The CP software gave each word in the sample a category and a point value. The manual and automatic coding were compared using a utility program.

Results: Overall results showed higher accuracy for categories that had a lower point value. There was no clear pattern as to discrepancies between the CP software and the manual coding. The accuracy across all the samples measured 78.2% *(SD = 4.4)*. The ability of the CP software to code was a statistically significantly greater than the manual coding. For reasons unknown CP did not code 2.9% of the utterances.

Discussion: DSS scoring done by CP is lower than the desired 80% accuracy consistent with the manual tagging. The CP software is still making errors. The author speculated that there may be a need for a sample size larger than 50 utterances. Accuracy may also be dependent on the fact that coding becomes harder when the child is older. There may also be problems with the software itself. CP uses probabilities extracted from the GramCats software to grammatically tag the utterances. It is these probabilities that may be affecting the outcomes. My study will use a current version of GramCats. A more accurate analysis by GramCats or a similar program would probably enhance the accuracy of an automated DSS analysis.

Channell, R. W., & Johnson, B. W. (1999). Automated grammatical tagging of child language samples. *Journal of Speech, Language, and Hearing Research*, *42*, 727-734. Retrieved November 3, 2014 from http://web.b.ebscohost.com.

Purpose: Previous research has shown the effectiveness of automated probabilistic computer programs for analyzing language samples from adults, but the use of these programs to analyze language samples from children has not yet been studied.

Participants: Language samples from 30 typically developing children had been collected by graduate students during conversation and games in the child's own home.

Instrumentation (Computer Program): The program, called GramCats, uses two types of probability information to choose the correct tag for each word in a language sample. These

probabilities are extracted from a corpus that had already been grammatically analyzed. The relative tag probability is the likelihood of each of the possible grammatical tags for the word; this is stored in the program's dictionary. The tag transition probabilities represent the likelihood that tag B will follow tag A; these are stored in a probability matrix. The program chooses the tags for the utterance that yield the highest probability total.

Procedure: A version of each sample was manually tagged, and then the original sample was run through the program. Agreements and disagreements with the manual tagging were tabulated.

Results: Overall, tagging at the word level was 95.1 % accurate. The lowest accuracy was for tagging words as auxiliary *had*, auxiliary *get*, and auxiliary *getting* and the highest accuracy was for tagging copula *be*, copula *being*, and copula *is*. The level of tagging accuracy was almost as good as manual tagging.

Implications or relevance for the current study: My study will look at the accuracy of a computer program tagging child language samples with and without SALT-coding. (Davis, 2012).

Condouris, K., Meyer, E., & Tager-Flusberg, H. (2003). The relationship between standardized measures of language and measures of spontaneous speech in children with autism. Retrieved November 3, 2014 from http://web.b.ebscohost.com/

Introduction: The language of children is typically measured in two ways. The first is through standardized tests and the second is through a language sample analysis. It is the standardized tests that make comparisons and measurements quick and easy. Language sampling is typically longer to transcribe and analyze. It language sampling that has been noted for better analyzing what happens in informal settings. The authors of this study were interested in the impairments of language related to autism. They were particularly interested in how standardized measures versus language sampling analysis procedures compare in this population. The past research presented in this article showed that some smaller studies found comparable results for standardized tests and language sampling. The majority of studies presented showed research that favored language sample analysis in comparison to standardized tests. While past research has compared language sample analysis and standardized measures using children with language impairments no study prior to this had been done in this area with regards to children who have autism. Children with autism often present with certain behaviors that have the potential to be greatly affected by the testing environment.

Method: The participants in the study included 44 children with autism. Trained examiners confirmed the diagnosis of autism. The three standardized measures used in the study included the Peabody Picture Vocabulary Test-Third Edition (PPVT), Expressive Vocabulary Test (EVT), and the Clinical Evaluation of Language Fundamentals (CELF). The tests were administered by a speech language pathologist over two 60 minute sessions for each child. Each test was check by a trained coder. The language samples were collected in a laboratory for 30 minutes while the parent and child interacted. Samples were transcribed using the format from the Systematic Analysis of Language Transcripts (SALT). A 100-utterance sample was collected for all the children except four. The measures used to analyze the spontaneous speech included mean

length of utterance in morphemes (MLU), the Index of Productive Syntax (IPSyn), and number of different word roots (NDWR).

Results: Overall results indicated language function to fall below age related comparisons on both types of assessments. Each child's performance was measured on an individual level. When looking at the correlations between spontaneous speech measures and standardized measures the authors saw that NWDR and MLU both correlated with measures found on standardized tests. It was seen that IPSyn had no significant correlations with any standardized measures.

Discussion: Each measure, whether it was a language sample analysis or a standardized test, showed a significant deficit in the language of children with Autism. After comparing these measures the authors concluded that language sample analysis and standardized tests both gave equal representations language abilities in children with Autism. There are aspects of pragmatic language that standardized tests are not sensitive to. The authors further stated that while both of these measures are used for different reasons, clinicians and researchers may make use of both with confidence.

DeRose, S. J. (1988). Grammatical category disambiguation by statistical optimization. *Computational Linguistics, 14*, 31-39.

Introduction: In order to understand language input a computer must be able to appropriately assign grammatical categories. The assigning of grammatical categories has been looked at using many different algorithms. This article was focused around a new algorithm called VOLSUNGA. VOLSUNGA has been shown to do well with ungrammatical utterances, gives about 96% accuracy, and runs in linear time. The author was particularly interested in VOLSUNGA's interaction to the Brown Corpus. The Brown Corpus consists of one million English words that have been used in many studies.

Previous Disambiguation Algorithms: The author reviews issues behind "lexical category ambiguity." Words that are categorically ambiguous make up 11.5% of vocabulary words or what the author calls "types." There are an even greater percentage of 40% categorically ambiguous words that make up running words or what the author calls "tokens." The authors firsts review algorithms created by Klein and Simmons that rely on suffixes and word set contexts to understand the tags. The article then reviews a program called TAGGIT, which received 77% accuracy on the Brown Corpus. In this program after all tags have been assigned the program eliminates tags based on context. The next program the author reviews is called CLAWS. This program is similar to TAGGIT. The main differences included using a larger portion of the tagged Brown Corpus and using a larger set of tag options. It is in this program that the ideas of relative tag probabilities are introduced.

The Linear-Time Algorithm VOLSUNGA: The program demonstrated here is linear based. VOLSUNGA uses what is referred to by previous programs as the "optimal path." The optimal path is what is seen when multiple probabilities multiple together and the highest number is chosen. VOLSUNGA also creates relative tag probabilities based on numbers from the Brown Corpus. This program does not make use of special exceptions lists, idioms, or tag triples. The

author then discusses how he uses what is called Dynamic Programming to find the optimal path. Here the program of VOLSUNGA focuses around reducing the number of possible paths to find the optimal path. Even with novel utterances VOLSUNGA uses contextual information to extract probabilities for a particular tag.

Accuracy Analysis: Accuracy of analysis was measured by running the program through the Brown Corpus. Tables in the study provided the number of types and tokens, as well as accuracy percentages for each category.

Conclusion: VOLSUNGA has improved upon areas of weakness in other program such as CLAWS. One example being, it makes use of 100% of the Brown Corpus. The program showed on overall accuracy of 96%. While other modifications can still be made to improve the software, the author concludes that it is a useful program for language analysis.

Ebert, D. K., & Scott, C. M. (2014). Relationships between narrative language samples and norm-referenced tests scores in language assessments of school aged children. *Language Speech and Hearing Services in the Schools, 45,* 337-350. Retrieved November 3, 2014 from http://web.a.ebscohost.com.

Introduction: Language assessment is a difficult task for the school clinician. Standardized testing and language sample analysis are what the practicing clinician will commonly use to tackle this task. Each has the potential to examine a variety of skills in a number of ways. Clinicians need to understand the pros and cons of each. The purpose of this study was to compare the features of both. Norm referenced tests dominate the language assessment process. They are quick, can test multiple modalities, and provide cut off scores to be used by districts. They do have limitations; they do not provide real life information, they do not work well in all populations, and there is often a loss of depth with each area tested. Language sample analyses make a good alternative form of testing. The information given provides knowledge about real world performance and allows the clinician to set goals. You also can glean information about global language structures and finite structures. Language sample analysis is sensitive for a variety of disorders. Unfortunately language sample analyses are time consuming and must be adjusted specifically for the age and needs of the child. Lots of research has examined the relationship between norm referenced tests and language samples. The evidence has been conflicting because of all the variables involved. Also little has been done to explore individual performance on these two measures. The authors here wanted to specifically look at how age related to the measures taken. They also wanted to explore a wide range of norm-referenced assessment and narrative measures. They created boundary for classification of a language problem to move past the typical correlation assessment.

Method: The study used explored past records of speech-language pathologists. Children had to have been referred with in the last 10 years. Language samples were accounted for from a wide variety of children. Children were only used if they had a language sample as well as standardized testing done. The sample size resulted in 73 children. There were 11 subtests included in the results of the study from 4 norm-referenced tests. Each measure of language was categorized with the area of language it assessed. The language samples were analyzed in 8 different ways using SALT software.

Results: There were only four correlations found in the older group of children. They were as follows Gray Oral Reading Test (GORT), Comprehension with Total Number of Words (TNW), Subordination Index (SI) with Understanding of Paragraphs, and Number of Different Words (NDW) with Formulated Sentences. Looking at the younger kids (below age 9) the authors found that there were far more correlations between norm-referenced tests and language analyses.

Discussion: There were far more correlations between the standardized measures and language sample analysis in the younger group. The authors suspect that this is due to the complex nature of language as children age. They suggest a variety of elicitation measures for older children. In younger children the correlations were found to be particularly strong at the word and sentence level. There was limited information on written language score. The authors suggest that future research needs to explore this area. The authors concluded that overall there was a moderate association between narrative analyses and norm-referenced measures.

Gutierrez-Clellen, V. F., Restrepo, M. A., Bedore, L. M., Peña, E. D., & Anderson, R. T. (2000). Language sample analysis in Spanish-speaking children: Methodological considerations. *Language, Speech, and Hearing Services in Schools, 31*, 88-98. Retrieved November 3, 2014 from http://web.b.ebscohost.com.

Introduction: There is no clear diagnostic process for Spanish Speaking (SS) children. This article was interested in developing a discussion on the selection and development of diagnostic procedures for Spanish Speaking children. The authors were also interested in how current measures can be applied in research to better understand their effectiveness.

Sociolinguistic Influences in the Language Performance of SS children: In the United States Spanish speakers are considered a minority group. There is an observed shift to be seen when children must move from their first language (L1) to a second language (L2). For each individual child it is difficult to see the patterns of the shift because each language sample may provide different information. Furthermore proficiency in L1 and L2 is always changing.

Selecting Measures of Spanish Grammar: Making measures of grammar in Spanish is a difficult challenge first because Spanish is a morphologically complex language. The variety of language abilities of SS children in the United States also makes it difficult. Assessment of Spanish relies on development of forms as well as measuring the most relevant aspects. One measure developed to assess Spanish morphology was the Developmental Assessment of Spanish Grammar (DASG; Toronto, 1976). This measure is similar to Developmental sentence soring in English. The DASG fell below clinical acceptance levels to identify children with language impairment. One alternative measure for assessing grammar is counting the number of grammatical errors per T-unit. This method has been shown to discriminate well between disordered and typical language form.

Selecting Measures of Sentence Length: The authors explored measures of sentence length. One measure that appears to be accurate when assessing SS is mean length of response in words (MLR-w). Understanding mean length of utterance in morphemes (MLU-m) is more difficult to

calculate because of the inflected nature of Spanish. Developing criteria for counting has been difficult for researchers, thus this process is not standardized for all speakers. Codeswitching in SS children as well as dialect can further make it difficult to make measures of MLU. Overall MLR-w has seen to be more reliable than MLU-m.

Clinical and Research Implications: There are a variety of options to consider when assessing a SS child. Because of the lack of understanding on certain measures, it is important to consider a variety of diagnostic procedures in order to get a full picture. The authors recommend that further research be done in the area of understanding diagnostic procedures for morphological counts and language complexity. Research further need to be done exploring the development of Spanish-speakers in their native language.

Gutierrez-Clellen, V. F., & Simon-Cereijido, G. (2009). Using language sampling in clinical assessments with bilingual children: Challenges and future directions. *Seminars in Speech and Language, 30,* 234-245. doi:10.1055/s-0029-1241722

Introduction: A reoccurring concern among researchers today is the validity of testing for language impairment in children who are bilingual. The authors of this article were particularly concerned for those children who are Latino. Many standardized tests include Latinos in their norms, but they do not have norms that focus specifically on the bilingual children. One study was noted for testing that showed moderate effects on scoring for bilingual children. Other tests that look at the Spanish language specifically have been shown to have poor sensitivity and specificity. Using a language sample on children who are bilingual may be the best option for clinical decision making.

Which Language to Assess: It is difficult to say which language a child is most proficient in because of the learning that occurs in different contexts. A child may be more familiar with home word in language one and less familiar with academic words at school. The opposite may also be true for language two. Development of each language varies over time. The authors believed that testing should be done in both languages.
Is Language Sample Analysis a Good Diagnostic Indicator: Research has shown language sample analysis to be a highly effective tool in the diagnostic process for both English-speaking monolingual children and Spanish-speaking children. In Spanish-speaking children analyses that were effective included looking at a child use of clitic pronouns and use of ditransitive verbs.

Language Sample Analysis Procedures: The authors use narrative retell to elicit language samples. Systematic Language Analysis of Language Transcripts (SALT) software has both and English and a Spanish version. In SALT clinicians can access a database that contains over 2000 language samples of Spanish-speaking children who are English language learners. The sample imputed by the clinician is analyzed by looking at the child's MLU for English and Spanish, overall grammaticality, morphosyntactic accuracy, and verb argument structures.

Spontaneous Language Markers in English: In English accuracy of scoring is achieved by looking at a set of finite verb morphology score. Verb forms were listed by the authors. The same verb morphology scoring system was used. Past research done by the authors indicated that this same scoring system worked well for Latino children. Because of children being tested

with limited English proficiency, it remains important for further testing to be conducted in Spanish.

Spontaneous Language Markers in Spanish: The article listed two ways to assess the grammar of Spanish-speaking children. The first way was to calculate mean length of utterance in words (MLU-w). This can be calculated using the SALT program. The SALT program can then give you a percentage of ungrammatical utterances used. A Spanish-speaking individual may also be assessed by looking at their ability to mark grammatical structures. If the narratives from both the Spanish and the English language samples show deficiencies then it is probable that the child has language impairment.

A Rubric for Assessing Verb Argument Structure: A point system is used by the authors to achieve a total argument structure (TAS) score. One point is given for the verb and each of its arguments, making the highest score a four. The rubric for each language is different because of the requirement of each language to include certain arguments. According to the research conducted by the authors the rubric was successful at identifying bilingual children with language impairment. The scoring rubric and norms were included in the article.

Conclusion: The article provides specific instructions and help when assessing a child who is a Spanish-speaker. Interviewing others and understanding performance in both languages is essential.

Hassanali K., Liu Y., Iglesias A., Solorio T., & Dollaghan C. (2014). Automatic generation of the index of productive syntax for child language transcripts. *Behavior Research Methods, 46,* 254-262. Retrieved November 3, 2014 from http://search.proquest.com.

Introduction: Research has explored a number of methods for quantifying expressive language abilities in people. Numeric scores provide a way for both the clinicians and researchers to measure performance. Manual analysis can make understanding and quantifying a language sample time intensive. There is great potential in automated analysis of language samples. With these faster analyses researchers and clinicians can more fully access information in language samples. The index of productive syntax (IPSyn) is one analysis that is widely known. This analysis categorizes using the following labels: noun phrase, verb phrase, questions and negations, and sentences. Under each of these areas there are 60 subcategories. A score is calculated for the number of times a certain structure appears in a sample. The authors review a number of software systems. The Computerized Profiling (CP) software can calculate IPSyn, but requires some manual input. The Sagae system computes IPSyn, but it is completely automated. Research has shown the Sagae system to be more accurate than CP, but there are still some limitations. The Sagae system is also not commercially available for use. The authors of this article wanted to explore a completely automated system that would be available to those involved in research. The authors not only wanted to study the IPSyn scoring, but they wanted to look at language development as well.

The Development of AC-IPSyn System: The authors developed a program called Automatic Computation of IPSyn system (AC-IPSyn) that would allow for transcribing in SALT and CLAN software as well. The program was designed to show a list of all the contracts along with the

score and a line connecting to where it first appeared. The program further gave a summary score for each category and subcategory. The program followed four steps. The first step was preprocessing, where the sample is appropriately formatted. CLAN and SALT can both be used here. The next step was parsing. The authors use what they call the Charniak parser, which labels each word with parts-of-speech (POS) tags. These POS tags are the same as what's used in the CP software. It is the statistical research done on the accuracy of the POS tagging that gives it higher accuracy. The Charniak parsing trees were fairly accurate with the child samples. Most parsing errors occurred when the software came across a word that was not in the corpus. The third step in the process involved identifying IPSyn structures. In this step the syntactic constructs were found using the POS tags and the parsing results. In step four the scores were computed. The software totals up the categories used. It also considered all the rules and exceptions in the scoring process.

Evaluation of the AC-IPSyn system: Two data sets were used to evaluate the system. Data Set A and Data Set B each contained 20 utterances. The system was graded using a point-by-point accuracy and a point difference method. The point difference compares the manual coding and the automatic coding by looking at the absolute difference between the scores. The point-by-point agreement looks at the number of agreements in each category and divides that number by the number of decisions made. Scores were also given for CP software and the Sagae software for Data Set A. Results showed that the AC-IPSyn system outperformed both the CP software and the Sagae software. The authors speculated that the Sagae performed lower because it used less robust rules. They also speculated that the reason the CP software did worse was because the POS tagging and the morphological analysis were used alone. Errors in the AC-IPSyn occurred most due to parsing and POS tagging errors.

Conclusions: The coding done for calculating IPSyn can be done much fast in an automated system. Software like CP has a longer processing time and involves some manual coding. The authors seek to improve their software by making the rules more robust. They want the system to be more compatible for older children.

Heilmann, J., & Malone, T. O. (2014). The rules of the game: Properties of a database of expository language samples. *Language Speech and Hearing Services in the Schools, 45,* 277-290. Retrieved November 3, 2014 from http://web.a.ebscohost.com/

Introduction: Specific language impairment (SLI) is a disorder that follows children through adolescence into adulthood. Speech and language pathologists working with older children in junior high and high school have a need to use language sample analysis (LSA) just like their colleges working with younger children. Most speech and language pathologists do not use LSA because of the lack of standardized procedures for older children. As with any language assessment the context has an effect on the type of information you can collect. Research has seen that expository discourse tasks will elicit more complex language. It is this expository discourse from older children that the authors suggested to be incorporated into a database such as the Systematic Analysis of Language Transcripts (SALT) software. The authors reviewed ways to standardize the testing done on students so it was unbiased for their study. They also review different types of analyses to understand which ones would be best to test on an older population.

Method: There were two groups of children from which language samples were taken. The groups were used in previous studies. The children were not pulled from those in special education services or English language services. There were 70 SLPs involved in collecting language samples from children in 55 different schools. The task involved the children explaining their favorite sport or game. The samples were all recorded digitally and transcribed by use of SALT software. The authors outlined language measures that appeared to be sensitive to difference in older children such as mean length of C-unit (MLCU), clausal density (CD), lexical diversity, productivity, expository scoring scheme, verbal fluency, and errors and omissions.

Results: Significant correlations with age and performance happened on MLCU, relative clauses, NDW, WPM, ESS, and mazes. An ANOVA was also conducted to understand difference in topic selection such as group sport, individual sport, and game. The NDW, NTW, and TCU all had statistically significant differences. When comparing the results to other studies the authors found that interview format did not differ with expository discourse in MLCU and CD values.

Discussion: The authors have created a database for older children undergoing LSA. This database will hopefully influence other research in the future. As consistent with other studies, age did influence language development, even in these older children. While topic did not influence most of the analyses there were some analyses that were influenced. The authors did caution against using games as a topic of discussion. They did believe that benefits of having the student choose a topic of interest out weighted the risks of topic choice. The authors concluded that context does influence performance. A factor analysis proved that children use a variety of language areas in expository discourse. The purpose here was to provide SLPs with a database for comparison. One direction future research can take is to study the effectiveness of the database in identifying children with LI.

Heilmann, J., Miller, J. F., & Nockerts, A. (2010). Using language sample databases. *Language, Speech & Hearing Services In Schools*, *41*, 84-95. doi:10.1044/0161-1461(2009/08-0075)

Introduction: The process of language sample analysis has evolved over the years. This evolution of language sample analysis has helped researchers better understand child language development. Language sample analysis further is thought to best represent a child's language abilities. Creating a reference in which to compare a child's language skills can be a difficult process. The creation of the Systematic Analysis of Language Transcripts (SALT) was put together by researchers to address this concern of reference databases. A protocol was created, after which databases were created to represent different groups of children. Throughout the years the databases created have been expanded. As of this time the authors noted that SALT had 6 databases that included 6,500 language samples from 4,000 children. Many researchers have back up the use of the SALT. The authors from this studied noted that identifying children with language impairment (LI) is a difficult task. Standardized tests are often used, but do not identify well. There is supported evidence that suggests that naturalistic testing better identifies children with LI. The study undertook aimed at comparing the clinical diagnosis of children to SALT's classification of these children.

Method: Language samples were collected from 244 children who were considered to have LI. Children with disabilities beyond LI were excluded from the study. At least 100 utterances were collected by SLPs who had undergone training in collecting language samples. There are more than 50 different types of analysis that can be done in SALT. The authors selected what were considered the most researched and sensitive measures.

Results: The study was interested in sensitivity, or the ability to identify children with language impairment, and specificity, or the ability to identify children without language impairment. Overall sensitivity was given at 69% and overall specificity was given at 84%. Each of these increased then the authors controlled for age. Sensitivity and specificity were both seen to be higher for the younger children. Older children were still observed to have a sensitivity of 77%.

Discussion: SALT software has potential to be of great use to clinicians. The number of protocol represented in the databases make it more flexible to get results. SALT software aims at making language sample analysis easier. SALT gives a good representation of language performance, but further research needs to explore SALT's ability to analyze other language areas such as a child's ability to ask questions.

Heilmann, J., Nockerts, A., & Miller, J. F. (2010). Language sampling: Does the length of the transcript matter? *Language, Speech & Hearing Services In Schools*, *41*, 393-404. doi:10.1044/0161-1461(2009/09-0023)

Introduction: Language sample analysis has strong evidence to support its clinical use. Like every form of analysis it has advantages and disadvantages. Clinical practice has required a standard of 50 utterances or less in a language sample. The authors noted conflicts in research over sample length. Some studies recommended lengthy transcriptions, while other supported shorter analysis to get needed information. The validity of analysis length may further be affected by the interaction of the context (e.g. narrative versus conversation). A child's age may further affect the validity of the shorter sample length. Here the authors used the SALT database to analyze their samples.

Method: There were 231 monolingual typically developing children in the study. A school SLP was trained on the collection of the samples for the SALT database. Each child gave two language samples. One sample was taken in a conversation context and another sample was taken in a narrative format. Instead of an utterance length cut the authors used a timing cut of 11 minutes for analysis. They then split each sample into a 1 minute analysis, a 3 minutes analysis and a 7 minute analysis as a reference for the smaller samples. The analysis they used included number of total utterances (NTU), words per minutes (WPM), number of different words (NDW), mean length of utterance in morphemes (MLUm), percentage of maze words (maze%), and composite scores for errors and omissions.

Results: The authors documented that there were modest differences between the 1 minutes samples, the three, minutes samples, and the seven minute samples, but there was no statistical significant difference between them. The authors also saw that there were no statistically significant differences with regards to sample context and sample length interactions or interactions between age and sample length. The last analysis done was a three way interaction

looking at age, context, and length. There were also no statistically significant differences seen here. Coefficients of variation were calculated and the authors did note greater variability for the smaller samples.

Discussion: While language sample analysis is seen as the most valid measure, clinicians do not use it as often due to time constraints. Taking a shorter language sample may be more functional, but are seen to be just as stable. The authors noted at the end the many limitations of this study, one being that the language samples collected were not designed specifically for short language sample analysis. It is true with a shorter language sample that children may not be able to demonstrate their range of performance. The authors recommend using shorter language samples in a comprehensive kind of assessment. They can also be used to monitor progress. When doing analyses on specific areas of language the authors recommend sticking to the 50 to 100 utterance format. Furthermore, information on each procedure during the study may be used to assess the reliability and validity of each measure. My study is also interested in making language sample analysis faster, while maintaining reliability and validity. Computer software will be tested using probabilities extracted from a 450 million word corpus.

Hewitt, L. E., Hammer, C. S., Yont, K. M., & Tomblin, J. B. (2005). Language sampling for kindergarten children with and without SLI: Mean length of utterance, IPSyn, and NDW. *Journal of Communication Disorders,* 38, 197-213. Retrieved November 3, 2014 from http://www.sciencedirect.com/

Introduction: Language sample analysis has been asserted to be the most valid form of assessment. There are many documented advantages of language sample analysis. Many clinicians use language sample analysis, but often not with standardized protocol. The authors presented evidence suggesting clinicians do not perform standardized analyses because there is a lack of reference data. With reference data clinicians can compare if a child is out of normal limits and see where the child is functioning. The most common standardized measurement used is mean length of utterance in morphemes MLU-m. Unfortunately this measure is not useful past 48 months. Although the authors believe that MLU-m may be helpful in assessing child older than 48 months who have specific language impairment (SLI). Research has shown that children with SLI have reduced MLU-m through even the school aged years. The authors mention the advantages and validity of using measures such as the index of productive syntax (IPSyn) and number of different word roots (NDW). Clinicians can take great advantage of these standardized measures while getting a great feel for natural communication.

Method: Language samples were taken from 27 typically developing children (mean age 5.99) and 27 children with language impairment (mean age 6.01). Each sample was required to have 50 utterances or more. Data from each child was audio recorded. Narrative retell and conversation were intermixed throughout the sample. All samples were transcribed using the SALT software. After being transcribed by an undergraduate or graduate student the samples were rechecked by the graduate student in charge of the project. The SALT software was then used to analyze the MLU-m and the NDW. Manual coding was used for IPSyn analysis.

Results: Using an analysis of variance (ANOVA) the authors noted that there was a significant difference between the children with LI and the typically developing children on MLU-m, NDW,

total score on the IPSyn, and the IPSyn sentence structure subscale. The IPSyn noun subscale and verb subscale was not found to be statistically significant.

Discussion: In this article the authors support language sample analysis as a valid and reliable tool for detecting language impairment. The authors speculated correctly with regards to the children's performance. They did note that there was still variability among the typically developing children and the children with SLI. If these measures of performance and cut off scores were to be used beyond the scope of this research project the authors showed that sensitivity would be poor, but specificity would be good. There were some limitations to the study. Some children had to be excluded because their utterances amounted to less than 50. There was also great variability on the level of the individual child. Further research needs to explore other types of analyses to increase sensitivity in the language sample. My study will focus on the reliability of language sample analysis with aid of computer software.

Johnson, B. (1992). *Automated grammatical tagging of spoken and written English* (Unpublished Master's Thesis). Brigham Young University, Provo UT.

Introduction: If high accuracy on language sampling analysis could be done automatically this would prove to be a great asset to speech and language pathologists. The present study was interested in using and algorithm used by DeRose in 1988 called VOLSUNGA. The VOLSUNGA algorithm has been shown to have high accuracy on the Brown University Corpus. The corpus contains all adult text and so it misses out on naturalistic utterances and child speech. The present study was interested in examining the aspects of the text (e.g. date of publication) and how they affected the results. The study also was interested in in how the GramCats software (Channell, 1991) made use of the VOLSUNGA algorithm.

Method: A variety of written texts and spoken utterances were used in the present study along with the Brown University Corpus. The probability data was extracted by GramCats through the use of relative tag probabilities and transitional tag probabilities. The tag set used in the study was the same one as used in the original tagging of the Brown University Corpus.

Results: Accuracy of the tagged texts fell between 89.47% and 94.49%. The samples taken were ranked in comparison to the Brown Corpus. The sample with the highest similarity was the edited written discourse from 1991. The sample with the lowest similarity was the unedited child spoken discourse

Discussion: Tagging accuracy was high for the unedited samples, but it was highest for the edited text. The VOLSUNG algorithm had lower accuracy overall in this study than in the original research presented by DeRose. Some of this may be due to manual tagging errors. A number of people have contributed to the tagging process of the Brown Corpus over the years and this may have an effect on the accuracy. Future research needs to explore the necessary size of the corpus for accuracy. Improvement of the automated tagging could one day prove to be helpful in the clinical world.

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. Retrieved November 3, 2014 from http://clt.sagepub.com/

Introduction: Clinical language samples have the potential to provide clinicians with a wealth of information not found in other measures. This study was interested in the actual clinical use of language samples by speech pathologists in the United States.

Method: A survey was mailed to clinicians. The survey consisted of 22 items that included 7 open response questions and 15 closed response questions. Information was collect on caseload, standardized tests, and language sample analysis.

Participants: Five-hundred Participants were randomly selected from nearly 4,000 speech-pathologists currently employed in a preschool setting. Two-hundred and fifty-three of the participants responded and were included in the study.

Results: Most clinicians used both language sample analysis and standardized testing to perform an assessment. Eighty-five percent reported using language sample analysis. Of those eighty-five there were eight percent that reported to be required to elicit a language sample. Ninety five participants total reported using standardized tests. The most common form of transcription method was real-time transcription. Language analysis procedures that were non-standardized were favored. Developmental Sentence Scoring (DSS) was the most favored standardized procedure. Only 8% of clinicians reported using computer programs in language sample analysis. SALT was used by 6% of the clinicians and other computer programs were used by the remaining 2%.

Discussion: The authors concluded that language sample analysis is a relevant part of the assessment process. Computer analysis was documented to be limited at the time this article was written. SALT was speculated to be preferred because there is no additional coding involved. The authors mentioned that the low percentage of clinicians using computer may be due to a lack of awareness; the authors believed that clinicians are unaware of what the computer can do.

Future directions: The interest in future studies is to make language analysis easier and more accurate. Clinicians could benefit from assistance. Computerized analysis will have to be used in the future. My study will examine the accuracy of computer tagged coding in language sample analysis as a means to make language analysis easier and more accurate.

Klee, T., Membrino, I., & May, S. (1991). Feasibility of real-time transcription in the clinical setting. *Child Language Teaching & Therapy*, *7*, 27-40. Retrieved November 3, 2014 from http://clt.sagepub.com/

Introduction: Language sample analysis is a long standing desirable measure in the diagnostic process. Computers have been seen to have great promise over the past decade to increase efficiency of language sample analysis. There have not been many advances in increase of efficiency of transcription time. This study compared real time transcription (RTT) to audio

taped transcription (ATT). Used with computer analysis software RTT could cut down substantially on time constraints.

Method: The participants included 22 children of which 21 had language impairment. Two of the children ended up being eliminated from the study because their MLU fell beneath requirements for the study. The children were also required to be at least 50% intelligible. Each language sample was 20 minutes long. A child was recorded using RTT and ATT while interacting alone with their mother. Two speech language pathologists complete the transcription. Each did 10 RTT and 10 ATT. The real time transcription was done with a transcriber in a room connected to the room with the child by a one-way mirror. The transcriber typed in utterances using the SALT format. The recorded analyses were transcribed 1-5 days after the initial recording to account to normal delay in transcription time. The two were compared using the following five measures: total number of utterances, total number of words (TNW), number of different words (NDW), percentage intelligibility, and mean length of utterance in morphemes (MLU).

Results: Looking at all the transcripts the results showed that the real time utterances compose about 90% of the audio recorded utterances. When complete and intelligible utterances were isolated the findings were similar. All the test done showed a high correlation between the two transcriptions. A t-test was done and it showed that mean difference in complete utterances and number of utterances was significant for each measure. For this set of speech samples intelligibility was rated to be about 80% for both types of transcription. There was no significant difference in MLU between the two transcriptions.

Discussion: Given the results from this study, RTT was seen to be successful. MLU and overall intelligibility were particularly successful. There are many other analyses that can be done in SALT. While there are correlations here, caution must be taken when comparing RTT scores to ATT norms. The authors still recommend using audio recorded analyses for clinical decisions with regards to production. They call RTT a starting point. More studies need to be done to know when RTT is inappropriate.

Long, S. H. (1996). Why Johnny (or Joanne) can't parse. *American Journal of Speech-Language Pathology, 5,* 35-40.

Introduction: Students in the field of speech language pathology can face difficulties when acquiring the necessary skill sets of metalinguistic awareness. Some are able to fill in the gaps as they work and others are not prepared to take on the challenge.

What do Students Need to Know About Language: Long states that the skills clinicians need to understand and apply include first the domains of language (e.g. morphology and syntax). They also need to understand the components under each of these different areas. Students further need to understand the way in which these components relate to each other and how to identify them. Being able to define these areas helps them communicate to a wide variety of audiences. They need to understand how to manipulate these areas as well.

How Well Are We Teaching Students: According to Long students do not understand the necessary principles behind language. This affects the students when they get into their first jobs and have to tackle clinical writing tasks. He then explains they are surprised when they realize how much they need metalinguistic awareness, lectures that focus around skills of metalinguistic awareness are difficult to follow, they have difficult interpreting standardized tests, and they cannot label grammar needed for computer programs. These gaps usually manifest themselves at the graduate level.

What's Gone Wrong? Long suspects a number of things contribute to this. It is a problem that is found not only in the United States, but in Australia and Britain as well. He thinks the field may not draw in as many students who are academically skilled. The field may bring in those interested in human interactions and less interested in analytical science. A change in the way literacy is taught may effect this change as well. There may also be less of an opportunity to practice principles of academic components in a clinical setting. Students may not have the opportunity to be exposed to computer analyzing techniques as well.

What Must We Do? Some aspects of this problem require much more drastic changes than is possible. Long suggests working to attract highly skilled people to the field. He recommends counseling to students and providing classes to those who need to fill in the gaps. He believes linguistics should be a bigger part of teaching and instructors need to focus on teaching computer skills with relation to language course work. Increasing students' opportunities to do analysis is also important. The goals and teaching of concepts needs to be consistent. Professors need to remain up to date on textbooks that would best suit their student's needs.

Conclusion: The study of grammar is an important aspect of speech and language pathology for students to become knowledgeable of. Metalinguistic awareness builds a strong clinical and academic foundation.

Long, S. H. (2001). About time: A comparison of computerized and manual procedures for grammatical and phonological analysis. *Clinical Linguistics & Phonetics, 15*, 399-426. Retrieved November 3, 2014 from http://web.b.ebscohost.com.

Introduction: The ability of a clinician to manage time is a huge assess in the workplace. Long makes a point of saying that language sampling is not just one task, but a series of tasks that take a considerable amount of time. While many articles are noted for saying this, few actually have recorded the amount of time it takes a clinician to do a language analysis. One study concluded that, with practice, to do a phonological assessment of a 200-word sample would take about 50 minutes to transcribe and another 50 minutes to analyze and interpret. Other studies were noted with comparable conclusions. Other studies were able to show that using a standardized test for phonological assessment shortened the assessment considerably. While there was great variation in time expected for a phonological analysis, it is suspected that there would be even greater variation for grammatical analysis. A simpler procedure such as MLU would take far less time than a detailed analysis such as LARSP; in one study MLU was predicated to take 25 minutes, and LARSP was estimated to take three hours. While these where all estimates, Long was interested in the actual time it took to do specific analyses.

Method: There were 256 participants in this study who were either students or practicing clinicians. They selected only the measures they were comfortable using in an analysis. Grammatical and phonological analyses were done on three language samples. The participants were provided with information as to sentence boundaries, proper nouns, and so on. Two of the participants had been diagnosed with some kind of language disorder and one was a typically-developing child. Each of the participants was to analyze the language sample by hand and record time spent on each sample in a log. Forms were also provided to record results. Computerized Profiling (CP) software was used by each participant in the study. They were educated on use of the software beforehand. They were required to import the text into the software, but they did not have to transcribe it. Ten phonological analyses were done and outlined by Long in the article. Five different grammatical analyses were done and included: MLU and descriptive statistics, Number of Syntactic Types (NST), a LARSP profile, Developmental Sentence Score (DSS), and Index of Productive Syntax (IPSyn). The study was biased for time against the computer. Each participant performed the analysis on the computer first and then did it by hand.

Results: While the interest of the study was in the time it took to complete each analysis, the authors were interested in the accuracy of each procedure. They created a point scoring system that compared manual versus computerized testing. The results scored computers as having higher or equal accuracy each time. Looking at the efficiency of analysis the authors noted that computer analysis was much more efficient. The computer ranged from 17 to 35 times more efficient than manual analysis. The author provided the reader with tables comparing the measures.

Discussion: There is strong evidence to suggest that language analysis done by hand is not a procedure that can be regularly done. The author also saw a wide variety in the amount of time it took to do a measure manually. The author was able to draw a number of conclusions. He saw that complexity of the language sample effected time spent. He also saw the grammatical measures were faster for the clinician to do than phonological measures. Lastly the article concluded that the type of analysis done affects the manual coding time. The author reviewed suggestions that may help in overcoming the time barrier. He mentioned shortcutting procedures by being familiar with patterns and protocol. Another solution to the problem may be found in using software speed up the analysis process. The data presented here repeatedly demonstrated time saving capabilities of computer-aided software. The article supports computerized language analysis procedures. My study focuses around computer software's accuracy at coding child language utterances.

Long, S. H., & Channell R. W. (2001). Accuracy of four language analysis procedures performed automatically. *American Journal of Speech-Language Pathology, 10,* 180-188. Retrieved November 3, 2014 from http://web.b.ebscohost.com/

Introduction: For many years past research has favored diagnostic procedures that use language sample analysis. While many standardized procedures have been developed for analyzing language samples research has shown that most clinicians use self-designed standards. Time constraints and lack of understanding are what hold most clinicians back from this more in depth analysis. Recent developments in computer software center on particular types of analysis.

There is a different level of functionality for each program because each software program demands something different from the user. Software has thus far only aspired to doing all the work and leaving human involvement out of the analysis process. The authors believe that the only way for language sample analysis to become completely free of human interaction is to improve the algorithms the computer uses. The interest of current research is in the decoding of child language. There has been some success seen with probabilistic computing. The authors involved in this study were interested in particular how automated language analysis works in a clinical setting.

Method: A total of 69 language samples were collected from a range of children. There was a mix of typical children and Language Impaired children. The ages ranged from 39 to 94 months. All of the language samples were analyzed using Computerized Profiling (CP) which uses two processes for analysis. The first involves the probabilities and corpus found in the GramCats program and the second involves the algorithms found in CP for a particular type of decoding. The four analysis procedures this article explored were mean length of utterance (MLU), LARSP, Developmental Sentence Scoring (DSS), and Index of Productive Syntax (IPSyn).

Results: Overall results indicated a higher score for each automated analysis of each category. The DSS scoring for the group of kids in the study who stuttered was the only measure found to be statistically significantly different. MLU was seen to be the most accurate of all the measures. IPSyn was seen to be highly accurate, but it was seen the scoring became more difficult as the utterances became more complex. LARSP was seen to have accurate coding at the word, phrase, and clause levels, but not the subordinate clause level. When comparing DSS and LARSP the authors found that DSS was more accurate in coding. The authors also found that correct coding was found for about 52% of all the utterances in the study.

Discussion: Using the results here the authors calculated the reliability of the automated software by using interrater reliability. It is generally accepted that percentages that fall above 85% are acceptable, 90% are good, and 95% excellent. Looking at each analysis and the data collect the authors concluded that LARSP was considered acceptable, DSS and IPsyn were considered good, and MLU was considered excellent. Further, the results of the automated analysis in this study were found to be comparable with the results of human analysis procedures in other studies. The authors concluded that CP software has the potential to produce information that is clinically useful and relevant. My study will focus on the accuracy of computer software to accurately decode language samples from probabilities extracted from a 450 million word corpus.

Overton, S., & Wren, Y. (2014). Outcome measures using naturalistic language samples: A feasibility pilot study using language transcript software and speech and language therapy assistants. *Child Language Teaching and Therapy, 30*, 221-229. Retrieved November 3, 2014 from http://web.b.ebscohost.com/

Introduction: An evaluation of a child's language skills is one of the most important parts of being a speech language pathologist. Many speech pathologists opt to give standardized tests over language samples. Going against what has been suggested by research, many will also select goals based on the gaps in the norm-referenced tests. Language sample analysis provides

a look at language use in real settings. It is believed to create a much stronger foundation for goal setting purposes. With language sample analysis you can see how all the areas of language work together for a particular client. The authors supported the use of computerized language sample analysis saying that it provide higher accuracy and more detail. Clinicians have complained in the past the language sample analysis done by the computer is confusing, unreliable, and inaccurate. Technology continues to move forward to bypass these constraints. The authors wanted to know if language transcription could be reliability carried out by Speech and Language Therapy Assistants (SLTAs) in computer software.

Method: While there are multiple options for language software the authors favored using Systematic Analysis of Language Transcripts (SALT) for their study. The SLTA selected for the study had some experience in working with children with communication needs, but did not have the training of a Speech and Language Therapist (SLT). The training was completed over two phases. In the first phase the SLTA completed a self-training over a three day period. Then a trial run was done in conjunction with the SLT. After this the SLTA did a transcription alone and errors were checked by the SLT before beginning the pilot study. A second part of training was completed half way through to discuss agreement and errors. Language samples were collected from 15 children between ages 5 to 12 years. Two samples were taken from each child. One sample was taken before a 10-week intervention and one sample taken after the 10-week intervention. The transcription procedures were taken from the SALT transcription manual. Agreements and disagreements were tabulated between the SLT and SLTA. Percentage of acceptable inter-rater agreement was considered to be 85%.

Results: Looking at the transcriptions done at the beginning of therapy (T1) there was an over 85% agreement between the SLTA and SLT for all the transcripts that fell above threshold. The transcripts that were taken after therapy were all found to be reliable. All of the measurements done were all reliable at the T1 measurements except for ‘words and morphemes in mazes’ and ‘maze placement.’ When measurements were done at T2 the results were all the same except for ‘maze placement’ had fallen with in reliability.

Discussion: The authors found that reliability for an SLTA to transcribe in SALT to be mostly accurate. The authors found that the intelligibility of the child greatly affected the outcomes of reliability. The low reliability found in mazes was not found to greatly affect the standardized measures. The noted improvement by the SLTA was attributed to the learning that occurred between T1 and T2. The authors concluded that the reliability found supported SLTAs undertaking transcription in software such as SALT.

Sawyer, J., & Yairi, E. (2006). The effect of sample size on the assessment of stuttering severity. *American Journal of Speech-Language Pathology*, *15*, 36-44. doi:10.1044/1058-0360(2006/005)

Introduction: There is has been a lot of debate among researchers regarding acceptable and reliable measures of stuttering. Much of the data collected on an individual’s stuttering is done in one speech sample. It is still unclear how long a speech sample needs to be in order to adequately represent a person’s disfluencies. Disfluencies are particularly difficult to measure

because they do not occur in a pattern. This study compared the differences seen in sample sizes in preschool children who stuttered.
Methods: There were 20 preschool children used in the study. The children were recorded in interactions with an experimenter and a parent. There was a minimum of 1200 syllables in each sample. Unintelligible utterances were discarded and disfluencies were marked in the Systematic Analysis of Language Transcripts (SALT) software. Well trained and experienced judges also rated the samples according to the criteria for stuttering-like disfluencies (SLD). There was a .98 interjudge agreement.

Results: The authors measured the differences between the first 300 syllables, then the first 600, then 900, and then 1200. They indicated whether there was an upward shift (U), downward shift (D), or no shift (N). Looking at the samples and the number of disfluencies identified the authors saw that length played a critical role in the identification of children with stuttering. Most children did not reach their maximum number of disfluencies in the first 300 syllables. The authors further took data on the RU, or the length of the disfluent events.

Discussion: Looking at SLD the researchers saw that longer samples can effect group outcomes where research is concerned and effect individual representation at the clinical level. The authors speculated that this may be due to the changes in the variables as the sample progressed (e.g. children felt more comfortable toward the end). The RU did not seem to be effected by sample size, but for individual children there were some differences seen. The authors suggested that it still be used by clinicians for individual diagnosis. The authors conclude by stating the variability of language sampling. Generally a longer sample is better, but there is a longer time commitment associated with that. Because there is such little research in this area more articles are needed to support the finding here.

Solorio, T., Sherman, M., Liu, Y., Bedore, L. M., Pena, E. D., & Iglesias, A. (2011). Analyzing language samples of Spanish–English bilingual children for the automated prediction of language dominance. *Natural Language Engineering*, *17*, 367-395. doi:10.1017/S1351324910000252.

Introduction: The diagnostic process relies heavily on the findings in language samples. A wide variety of language features can be explored through this process. Emerging research has focused on automated techniques to analyze language samples. Research is lacking in bilingual language samples. An interesting component of therapy with bilingual children is determining language dominance. A major question: Can automated analysis predict language dominance?

Method: Language transcripts of children about age 6 were used. The children retold four stories, 2 in English and 2 in Spanish. The authors developed a method where the three language groups balance bilingual (BB), English-dominant (ED), and Spanish-dominant (SD) were used to categorize language samples. The transcripts here were compared to monolingual Spanish-speaking children as well as monolingual English-speaking children. The children retold a story in each language and the features of their language were extracted. These features were combined and then the program assigned a category for each group. The part of speech (POS) tagging developed by the Child Language Data Analysis System (CHILDES) database was used. Language exposure was measured in this study as well.

Results and Discussion: The accuracy of predicting language dominance in different groups was displayed. No measurement fell above 75% accuracy or below 56% accuracy. The authors concluded that classification was not as accurate when using a feature selection approach. A measure of consistency between the assignments of the two language samples in each language showed an overall 83% accuracy. Research needs to further address making accuracy of language dominance higher.

Tilstra, J., & McMaster, K. (2007). Productivity, fluency, and grammaticality measures from narratives. *Communication Disorders Quarterly*, *29*, 43-53. Retrieved November 3, 2014 from http://web.a.ebscohost.com/

Introduction: Language impairment is a persisting disorder throughout the life of an individual. Speech language pathologists track the progress of children with language impairment using a variety of methods. The authors in the article explored general outcome indicators (GOI), which is an alternative method to documenting progress. The authors outline many limitations in using norm-referenced tests, criterion referenced measures, and language sample analysis. A GIO method is often referred in a more academic sense. The teach-test-retest method is used in the GIO process. Using this idea the authors chose to elicit language using a single picture for a narrative. The authors chose to measure language productivity, grammaticality skill, and verbal fluency. The authors were interested in which measure would be the best GIO of language proficiency.

Method: There were 45 participants in this study between 5 and 9. Three separate standardized black and white pictures were used to elicit narration. The child was required to construct a narrative about each. Some prompting was allowed from the examiner. The authors conducted five measures of productivity, four measures of fluency, and four measures of grammaticality. A narratives were recorded. The child all completed a standardized assessment on the *Oral and Written Language Scales* (OWLS). On 10% of the transcripts interrater reliability was found to be 95%. All the samples were transcribed using Systematic Analysis of Language Transcripts (SALT).

Results: The results showed reliability across samples for 10 different language measures. Examining the reliability at grade level for fluency there was a moderate to strong correlation for all the grades. Productivity was seen to be only reliable at the third grade level. Grammaticality was seen to only be reliable at the kindergarten level. There was a moderate correlation between the OWLS and the measurements of verbal fluency. Looking at grade differentiation the authors saw that total words per minute was significantly different for third graders when compared to kindergarteners and first graders. In summary the reliability of the 10 measures taken was strong.

Discussion: It is difficult to find a measure the will reliable and quickly measure a child's growth and performance. The study here indicated the in a narrative context measures of fluency are reliable for kid in third grade, first grade and kindergarten. Productivity only proved to be reliable for those kids in third grade. Criterion validity needs to further be explored beyond this study.

Van Rooy, B., & Schafer, L. (2002). The effect of learner errors on POS tag errors during automatic POS tagging. *Southern African Linguistics & Applied Language Studies*, *20*, 325. Retrieved January 9, 2015 from http://web.a.ebscohost.com/

Introduction: A corpus has the potential to be much more useful when the parts of speech (POS) have been tagged. Hand-tagging is a time consuming option for a large corpus. There is no existing perfecting automating software. There are complex factors to factor in when using automated tagging to aid the process. The corpus used in this study was entitled the Tswana Learner English Corpus (TLEC) that was taken from a larger 200,000 word corpus called the International Corpus of Learner English (ICLE). This corpus is composed of essays from advanced English language learners. The correlation between errors the English language learner and the tagging is not completely understood. The authors focused their research on exploring this relationship. Preparing the data for this type of an analysis is more time consuming than expected. The method they used for tag selection involves assigning all possible tags and then going through a "disambiguation" process.

Evaluation: Three taggers were selected for the study, which the main one being TOSCA-ICLE. The other taggers were the Brill-tagger and CLAWS7. As outlined by the authors a Brill-tagger is mostly a rule-based tagger whereas CLAWS7 and TOSCA-ICLE are more probabilistic taggers. Parts of the corpus were extracted and run through all three taggers. The authors manually checked the errors made. After data was taken on spelling mistakes and the mistakes were corrected the data was run through the software again.

Analysis of errors: In all the 2, 159 words in the data there were 76 spelling errors noted. The data showed that errors involving morphemes and syntax exceeded spelling errors. Types of errors on part of the writer that seemed to have little effect on the accuracy of that tagging were article, prepositions, and numerical features with in a noun. Predominately learner errors affected the tagging accuracy.

Conclusion: The authors concluded that learner errors greatly affected tagging accuracy in all three programs. CLAWS was seen to be the most accurate in tagging overall.

Vine, E. W. (2011). High frequency multifunctional word: Accuracy of word-class tagging. *Te Reo*, *54,* 71-82. Retrieved January 12, 2015 from http://web.b.ebscohost.com/

Introduction: Often automated word tagging is accomplished by use of a corpus and the probabilities extracted. The ability of a program to make distinctions between tags is especially difficult in a case of frequently occurring words with multiple uses such as *like*, *is*, and *so*. The author of this article wanted to explore the tagging of these multifunctional words. A comparison between the manual tagging and the automated tagging is made here. The author looked at two corpora: The Wellington Corpora of Spoken and Written New Zealand English (WCSNZE and WCWNZE). She also looks at the tagging of those corpora through two programs. The first is called the Constitute Likelihood Automatic Word-tagging System

(CLAWS) and the second is an unpublished program from Northern Arizona University developed by Douglas Biber.

Tagging the Written Corpus: The Biber and CLAWS both fell below acceptable limits for appropriate tagging for *like*, *as* and *so*. CLAWS did have a low error rate for the word *like*. The word *like* was tagged the most accurately on both programs, it was most easily identified as a preposition.

Tagging the Spoken Corpus: Only the Biber program was used to tag the spoken corpus and its performance dropped from what it was on the Written Corpus. Again the word *like* was the best tagged word. The author explored the different possible tagging mistakes made.

Conclusion: The author cautioned that the data here should not be completely reflective of the programs themselves because the words analyzed or problematic in general.