2015-06-01

Automated Grammatical Tagging of Clinical Language Samples with and Without SALT Coding

Andrea Nielson Hughes
Brigham Young University - Provo

Automated Grammatical Tagging of Clinical Language Samples with and Without SALT Coding

Andrea Nielson Hughes

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

June 2015

ABSTRACT

Automated Grammatical Tagging of Clinical Language
Samples with and Without SALT Coding

Andrea Nielson Hughes
Department of Communication Disorders, BYU
Master of Science

Language samples are naturalistic sources of information that supersede many of the limitations found in standardized test administration. Although language samples have clinical utility, they are often time intensive. Despite the usefulness of language samples in evaluation and treatment, clinicians may not perform language sample analyses due to the necessary time commitment. Researchers have developed language sample analysis software that automates this process. Coding schemes such as that used by the Systematic Analysis of Language Transcripts (SALT) software were developed to provide more information regarding appropriate grammatical tag selection. The usefulness of SALT precoding in aiding automated grammatical tagging accuracy was evaluated in this study. Results indicate consistent, overall improvement over an earlier version of the software at the tag level. The software was adept at coding samples from both developmentally normal and language impaired children. No significant differences between tagging accuracy of SALT coded versus non-SALT coded samples were found. As the accuracy of automated tagging software advances, the clinical usefulness of automated grammatical analyses improves, and thus the benefits of time savings may be realized.

Keywords: language sample, language impairment, automated tagging, software, SALT

ACKNOWLEDGEMENTS

I would like to thank the children in the Reno, Nevada area and their families for the language samples that provided a foundation for my thesis. I would also like to thank Dr. Nissen and Dr. Tanner for their cooperation and assistance throughout this process and Dr. Fujiki and Dr. Brinton for access to the language samples used. I especially would like to thank Dr. Channell for his constant encouragement and dedicated efforts. It has been a pleasure and delight to work with you. Finally, I would like to thank my family. I recognize that your faith, love, and support have guided me along this journey. Jordan, thank you for believing in me.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

DESCRIPTION OF THESIS STRUCTURE

The body of this thesis is written as a manuscript suitable for submission to a peer-reviewed journal in speech-language pathology. This thesis is part of a larger research project, dealing with the automated grammatical analysis of children's language samples for use in clinical language assessment. Portions of this thesis may also be published as a part of other articles on automated grammatical analysis, listing the thesis author as a co-author. An annotated bibliography is presented in the Appendix.

Introduction

Samples of client language can be of great benefit to the process of clinical language assessment and treatment but are resource-intensive for clinicians to collect and analyze. Naturalistic samples of the client's spontaneous, non-elicited language avoid the contrived nature of many standardized language tests, but clinicians may be hesitant to utilize this assessment measure due to impressions of a significant time commitment (Heilmann, 2010). The analysis of these samples is time intensive and requires skilled expertise for effective evaluation (Hassanali, Liu, Iglesias, Solorio, & Dollaghan, 2014). Long (2001) outlines the time commitment necessary for completion of various types of language sample analyses. Dependent on the efficiency of the clinician and the complexity of the language sample, a detailed analysis such as the Language Assessment and Screening Profile (LARSP; Crystal, Garman, & Fletcher, 1989) carried out by hand can take between 34 and 334 minutes to complete (Long, 2001), thus making such an analysis unlikely to be completed by a busy clinician.

While manual efforts at language sample analysis are time consuming, automated computer software programs are available to make the process more efficient (Channell & Johnson, 1999; Hassanali et al., 2014; Sagae, Lavie, & MacWhinney, 2005). Heilmann (2010) lists several of these computer programs, such as the Child Language Analysis software (CLAN; MacWhinney, 2015), Computerized Profiling (CP; Long, Fey, & Channell, 2006), and the Systematic Analysis of Language Transcripts (SALT; Miller, Andriacchi, & Nockerts, 2011). With an automated system to reduce time demands for language sample analysis, these programs open the possibility of greater language sample utilization in research and clinical practice (Hassanali et al., 2014). When compared to the time commitment for manual completion of LARSP, this same analysis using computerized software takes 15 to 98 minutes, again

contingent on the efficiency of the clinician and complexity of the language sample (Long, 2001).

The useful nature of automatic language sample analysis is clear; however, studies have compared the results of these automated systems to manual efforts and have found variable results in the accuracy of automated analysis. Long and Channell (2001) reported good but variable accuracy for automatic results compared to manual results using CP and its related systems. For example, accuracy at calculating the Phrase level of LARSP averaged 90.9% but accuracy on the Subclause level averaged only 13.5%.

Channell (2003) calculated the accuracy of automated Developmental Sentence Scoring (DSS; Lee, 1974) computed by CP. Comparing the automated system to manual efforts, Channell lists performance for each DSS category, displaying variable results. The overall accuracy of the automated system was 78.2%. Samples included both older and younger children as well as developmentally normal and language impaired participants. Point-by-point agreement for each sample yielded results no higher than 85%; this figure only occurred in two out of 51 samples.

Sagae et al. identified the grammatical relations (GR) in a language sample, which aid in producing a score for the Index of Productive Syntax (IPSyn; Scarborough, 1990). The precision of GR identification was variable, depending on the specific category (e.g., Subject, Predicate, Object, etc.). Using point-by-point agreement to compare the IPSyn scores of the GR system to manual efforts, the overall reliability was 92.8%. This system was compared to IPSyn scoring using CP, which amounted to a point-by-point agreement of 85.4%. While the results of Sagae et al. are superior to CP, the usefulness of this automated system analysis is limited as the software is not offered publicly (Hassanali et al., 2014).

Hassanali et al. (2014) utilize the Automatic Computation of the Index of Productive Syntax (AC-IPSyn) to evaluate language samples. The AC-IPSyn system acquired IPSyn scores that improved upon both the CP and Sagae et al. (2005) systems. Two data sets were studied, Set A and Set B. Set A included children between ages two and three. Set B included both developmentally normal and language impaired children who were six years old. Point-by-point agreement for Set A was 96.9% for the AC-IPSyn system, 92.5% for the Sagae system, and 86.2% for the CP system. Point-by-point agreement for Set B was 96.4% for the AC-IPSyn system and 87.39% for the CP system. The Sagae system was not tested using Set B.

These automated analysis systems provide a faster way of looking at language samples. Automated analyses bypass the significant time commitment needed previously in manual attempts. Although there are several automated programs for language sample analysis, there are variable results regarding the accuracy of these systems, as discussed above. One factor affecting the performance of these programs is the quality of the initial grammatical coding of the words in the utterances of the samples. All of these programs first grammatically code the words of the samples using a grammatical tagging program, and the accuracy of these programs relies heavily on the foundation of the accuracy of this initial tagging. For example, Hassanali et al. (2014) noted that their analysis system suffered largely from tagging and parsing errors. Few studies have specifically explored factors affecting the accuracy of automated tagging systems in child language.

One exception was Channell and Johnson (1999), who evaluated the accuracy of their automated tagging system, GramCats, using child language samples. Automated computer tagging was compared to manual tagging at the word and utterance level. At the word level, the range of accuracy was 92.9% to 97.4%. At the utterance level, the range of accuracy was 60.5%

to 90.3%. An average of 95.1% and 77.7% accuracy at the word and utterance level respectively indicated that many words are still tagged incorrectly. Sagae et al. (2005) describe the accuracy of the Charniak parser (Charniak, 2000) when applied to child language samples. The Charniak parser was originally trained on a data set of language from the *Wall Street Journal*, and its accuracy when applied to child language was 90.1% (Sagae, 2005). Thus many errors still occur in automated grammatical tagging. By improving the precision of automated grammatical tagging, the accuracy of clinically useful language sample analyses should increase.

One factor that might affect the accuracy of automated grammatical tagging is the format into which the samples are transcribed. Some of the grammatical aspects that are difficult for the computer can be precoded as the clinician or researcher transcribes the language sample. One commonly used precoding scheme is that of SALT (Miller et. al., 2011), which codes several inflectional morphemes such as possessive and plural markers on nouns or the progressive and regular past tense on verbs. This scheme was initially designed to allow software tabulation of a child's use of these markers. In contrast, the CLAN software used to require a similar precoding, but the use of precoding was dropped, as it was claimed that the computer was more accurate than the human transcriber in noting these grammatical details (MacWhinney, 2008), though no data were presented to back this claim.

Although SALT formatting may require extra time and effort, especially for the unfamiliar user, it may provide extra information regarding the assignment of tags and thus increase the accuracy of automated grammatical tagging. This increased accuracy might then require less correction and yield a net time savings. The purpose of this study is to investigate the effect of SALT formatting on the accuracy of automated grammatical tagging over a range of language samples, including samples from children with and without language impairment.

Method

Participants

Language samples had been collected from 30 children living in the Reno, Nevada area by Brinton, Fujiki, and Sonnenberg (1988). These samples were collected from children with language impairment and from typically developing children who attended the same school. The 30 children were subdivided into three groups containing 10 participants each. The three groups were (a) children with language impairment whose ages ranged from 7;6 (years; months) to 11;1, (b) typically developing children matched on language test scores (LTS) whose ages ranged from 5;6 to 8;4, and (c) typically developing children matched on chronological age (CA).

Manual Coding and SALT Formatting

Manual grammatical coding of the Reno Samples was previously performed in the Willmarth (1999) study using the grammatical tag scheme of Channell and Johnson (1999). Willmarth reported reliability levels for this tagging of 97%. The present author converted the language samples into SALT formatting preparatory to automated tagging.

Automated Tagging Software

This study used the gc5 software version 1.0 (Channell, 2015) to perform automated grammatical tagging. The gc5 software is not limited to samples in SALT format. However, several formatting features are required for the software to run properly: utterances must occupy separate lines, only proper nouns and the pronoun *I* are allowed in upper case lettering, mazes must be confined within parenthesis so as to prevent their influence on the tagging software, and a punctuation mark must be placed at the end of each utterance.

The gc5 software uses two forms of probability, extracted from a training corpus of tagged samples, to compute grammatical category assignment: the relative tag likelihood and the

tag transition likelihood. The relative tag likelihood is the probability that the word has one tag rather than another. For example, the word *break* could be either a noun or a verb. The number of times this word was used as a noun versus a verb was counted in the training corpus and stored in gc5's dictionary. The tag transition likelihood is the probability that a given tag was sequenced after two other tags. For example, the likelihood that a noun follows a verb and a preposition compared to other configurations creates a tag transition probability. The tag transition probabilities had also been calculated from the training corpus and stored in gc5's dictionary.

The dictionary used by gc5 included about 20,000 child utterances, which were drawn from the CHILDES database (MacWhinney, 2000) and hand tagged. The samples were taken from several children including Abe (aged 3;0 up to 5;0; Kuczaj, 1976), Sarah (aged 3;2 up to 4;11; Brown, 1973), 48 children conversing with other children, who ranged in age from 2;10 to 5;7 (Garvey, 1979; Garvey & Hogan, 1973), and children attending first, third, and fifth grade (Carterette & Jones, 1974). In addition to these samples obtained from the CHILDES database, the dictionary also contains verbs and adjectives gathered from the internet, which do not carry empirical probability information. Proper nouns are not included in the dictionary because all words beginning in an upper case letter are coded as such. Finally, words that are not found in the dictionary and therefore receive no information regarding their tagging scheme are tagged as nouns. The rationale for this decision is that words not included in the dictionary were mainly newer nouns such as *iPad* or *transformer*.

Procedure

The gc5 software automatically coded versions of the language samples both with and without SALT formatting. Using a utility program, the manual coding and automatic coding

were then compared for both the SALT and non-SALT formatted samples. The accuracy of automatic coding was inferred from its level of agreement with the manual coding and was calculated at the word and utterance level. A one-way analysis of variance (ANOVA) was used to look for differences among accuracy levels for coding samples from the children with language impairment, the LTS-matched but younger children, and the CA-matched children.

Results

The accuracy of gc5 in coding grammatical categories without SALT coding averaged 97.35% at the tag level. Comparatively, gc5 performance averaged 97.37% with SALT coding. At the utterance level, accuracy averaged 85.26% and 85.46% for non-SALT and SALT coded samples respectively.

Table 1, comprised of data from non-SALT coded samples, and Table 2, comprised of data from SALT coded samples, outline accuracy levels for each grammatical category used in the present study. Table 1 and Table 2 list each tag, the frequency with which each tag occurred in the samples, the accuracy of each tag usage, and the category confusions. Of the 79 different grammatical categories listed, 61 categories achieved accuracy levels at or above 90% for both SALT coded and non-SALT coded samples. For the samples that were not SALT coded, the computer software demonstrated marked (below 60%) difficulty accurately tagging auxiliary *get* (53%), auxiliary *had* (0%), and auxiliary *has* (49%). For SALT coded samples, marked difficulty was evident on the categories of auxiliary *get* (53%), auxiliary *gets* (0%), auxiliary *had* (0%), and auxiliary *has* (49%). SALT processed samples improved in four categories compared to the corresponding non-SALT coded categories (e.g., *possessives formed with 's*, *existential there, plural noun,* and *3rd present singular*). Similarly, non-SALT coded samples displayed increased accuracy in three grammatical categories (e.g., *copula is, proper noun,* and *auxiliary gets*).

Table 1

Grammatical Tagging Accuracy Without SALT Coding

Tag	Description	N	%	Confusions (%)
<*	negatives *not, n't*	(464)	100	
<$	possessive *'s*	(135)	92	BCZ (7) XBZ (1)
BC	copula *be*	(71)	100	
BCD	copula *were*	(16)	94	XBD (6)
BCDZ	copula *was*	(226)	98	XBDZ (2)
BCM	copula *am*	(17)	94	XBM (6)
BCN	copula *been*	(8)	100	
BCR	copula *are*	(226)	99	XBR (1)
BCZ	copula *is*	(1646)	98	XBZ (1)
CC	clausal conjunct.	(3622)	100	
CS	sub. conjunct.	(35)	60	IN (1) PD (20) PRL (9)
D$	possessives	(1130)	100	
DA	articles	(4734)	100	
DCN	cardinal numbers	(252)	95	PN(5)
DD	demonst. singular	(452)	100	
DDS	demonst. plural	(153)	100	
DN	indefinite det.	(454)	95	DPA (1) JJ (2) PN (1)
DON	ordinal number	(32)	75	NN (9) RB (16)
DPA	predeterminer	(131)	100	
DWN	noun clause det.	(17)	65	PWN (24) RBN (12)
DWQ	interrogative det.	(12)	75	PWN (8) PWQ (17)
DWX	exclamative det.	(1)	100	
EX	existential *there*	(356)	95	RB (5)
IN	preposition	(4272)	99	

JJ	adjectives	(2214)	95	NN (2) RB (1)
JJR	comparative	(20)	100	
JJT	superlative	(13)	100	
NN	singular noun	(6826)	99	
NNS	plural noun	(1854)	98	NN (1) PN (1)
NP	proper noun	(1243)	100	
P$	possessive pro.	(23)	87	D$(9) VB (4)
PD	demonst. singular	(1016)	98	DD (2)
PDS	demonst. plural	(113)	96	DDS (4)
PI	indefinite pro.	(218)	100	
PL	reflexive singular	(8)	100	
PLS	reflexive plural	(1)	100	
PN	quantifiers	(841)	93	NN (2) RB (1) DN (1) DCN (1)
PO	object case pro.	(1716)	99	
PRL	relative pro.	(299)	96	PD (3)
PS	subject pro.	(3372)	100	
PWN	nominal clause pro.	(143)	98	DWN (1)
PWQ	interrogative pro.	(78)	87	DWQ (1) PN (3) PWN (9)
PZ	3rd person pro.	(3237)	100	
RB	adverb	(3160)	95	IN (1) NN (2)
RBN	noun clause adv.	(149)	86	RBQ (1) RBS (13)
RBQ	questions wh-adv.	(64)	91	RBS (9)
RBR	comparative adv.	(5)	100	
RBS	subordinating adv.	(723)	93	IN (2) RB (4) RBN (1)
RP	verb particle	(1329)	88	IN (12)
RQL	qualifier	(425)	65	DD (2) DN (1) IN (2) JJ (2) NN (13) PN (3) RB (6) UH (4)
RQLP	post qualifier	(29)	93	PN (7)
TO	infinitive marker	(682)	100	
VB	verb	(4385)	97	NN (1) VBD (2)

VBD	past	(1557)	94	VB (4) VBN (1)
VBG	present participle	(1384)	99	NN (1)
VBN	past participle	(440)	80	JJ (5) NN (2) VB (2) VBD (11)
VBZ	3rd pres. singular	(1447)	99	NNS (1)
VPO	verb + pronoun *let's*	(51)	100	
VTO	catenative	(196)	100	
XB	aux *be*	(10)	100	
XBD	aux *were*	(26)	96	BCD (4)
XBDZ	aux *was*	(143)	99	BCDZ (1)
XBG	aux *being*	(3)	100	
XBM	aux *am*	(84)	95	BCM (5)
XBN	aux *been*	(4)	100	
XBR	aux *are*	(120)	89	BCR (11)
XBZ	aux *is*	(841)	98	BCZ (2)
XD	aux *do*	(286)	95	VB (5)
XDD	aux *did*	(84)	86	VBD (14)
XDZ	aux *does*	(76)	99	VBZ (1)
XG	aux *get*	(17)	53	VB (47)
XGD	aux *got*	(50)	80	VB (8) VBD (12)
XGG	aux *getting*	(1)	100	
XGZ	aux *gets*	(6)	83	VBZ(17)
XH	aux *have*	(58)	76	VB (24)
XHD	aux *had*	(2)	0	VBD (100)
XHZ	aux *has*	(81)	49	BCZ (2) VBZ (1) XBZ (47)
XM	modal	(480)	100	
XM*	modal + neg.	(85)	100	

Table 2

Grammatical Tagging Accuracy With SALT Coding

Tag	Description	N	%	Confusions (%)
<*	negatives *not, n't*	(464)	100	
<$	possessive *'s*	(135)	100	
BC	copula *be*	(71)	100	
BCD	copula *were*	(16)	94	XBM (6)
BCDZ	copula *was*	(226)	98	XBDZ (2)
BCM	copula *am*	(17)	94	XBM (6)
BCN	copula *been*	(8)	100	
BCR	copula *are*	(226)	98	XBR(2)
BCZ	copula *is*	(1647)	98	XBZ (1)
CC	clausal conjunct.	(3622)	100	
CS	sub. conjunct.	(35)	60	IN (11) PD (20) PRL (9)
D$	possessives	(1130)	100	
DA	articles	(4735)	100	
DCN	cardinal numbers	(252)	95	PN (5)
DD	demonst. singular	(452)	100	
DDS	demonst. plural	(153)	100	
DN	indefinite det.	(453)	95	DPA (1) JJ (2) PN (1)
DON	ordinal number	(32)	75	NN (9) RB (16)
DPA	predeterminer	(131)	100	
DWN	noun clause det.	(17)	65	PWN (24) RBN (12)
DWQ	interrogative det.	(12)	75	PWN (8) PWQ (17)
DWX	exclamative det.	(1)	100	
EX	existential *there*	(356)	96	RB (4)
IN	preposition	(4272)	99	
JJ	adjectives	(2214)	95	NN (2) RB (1)

JJR	comparative	(20)	100	
JJT	superlative	(13)	100	
NN	singular noun	(6826)	99	
NNS	plural noun	(1851)	99	NN (1)
NP	proper nouns	(1241)	99	NNS (1)
P$	possessive pro.	(23)	87	D$ (9) VB (4)
PD	demonst. singular	(1016)	98	DD (2)
PDS	demonst. plural	(113)	96	DDS (4)
PI	indefinite pro.	(218)	100	
PL	reflexive singular	(8)	100	
PLS	reflexive plural	(1)	100	
PN	quantifiers	(841)	93	DCN (1) DN (1) NN (2) RB (1)
PO	object case pro.	(1716)	99	
PRL	relative pro.	(299)	96	PD (3)
PS	subject pro.	(3371)	100	
PWN	nominal clause pro.	(143)	98	DWN (1) PWQ (1)
PWQ	interrogative pro.	(78)	87	DWQ (1) PN (3) PWN (9)
PZ	3rd person pro.	(3239)	100	
RB	adverb	(3161)	95	IN (1) NN (2) RQL (1)
RBN	noun clause adv.	(149)	86	RBQ (1) RBS (13)
RBQ	question wh-adv.	(64)	91	RBS (9)
RBR	comparative adv.	(5)	100	
RBS	subordinating adv.	(723)	93	IN (2) RB (4) RBN (1)
RP	verb particle	(1329)	88	IN (12)
RQL	qualifier	(425)	65	DD (2) DN (1) IN (2) JJ (2) NN (13) PN (3) RB (6) UH (4)
RQLP	post qualifier	(29)	93	PN (7)
TO	infinitive marker	(682)	100	
VB	verb	(4385)	97	NN (1) VBD (2)
VBD	past	(1555)	94	NN (1) VB (4) VBN (1)

VBG	present participle	(1384)	99	NN (1)
VBN	past participle	(441)	80	JJ (5) NN (3) VB (2) VBD (11)
VBZ	3rd pres. singular	(1451)	100	
VPO	verb + pronoun *let's*	(51)	100	
VTO	catenative	(196)	100	
XB	aux *be*	(10)	100	
XBD	aux *were*	(26)	96	BCD (4)
XBDZ	aux *was*	(143)	99	BCDZ (1)
XBG	aux *being*	(3)	100	
XBM	aux *am*	(84)	95	BCM (5)
XBN	aux *been*	(4)	100	
XBR	aux *are*	(120)	89	BCR (11)
XBZ	aux *is*	(842)	98	BCZ (2)
XD	aux *do*	(286)	95	VB (5)
XDD	aux *did*	(85)	86	VBD (14)
XDZ	aux *does*	(76)	99	VBZ (1)
XG	aux *get*	(17)	53	VB (47)
XGD	aux *got*	(50)	80	VB (8) VBD (12)
XGG	aux *getting*	(1)	100	
XGZ	aux *gets*	(6)	0	VBZ (100)
XH	aux *have*	(58)	76	VB (24)
XHD	aux *had*	(2)	0	VBD (100)
XHZ	aux *has*	(81)	49	BCZ (2) VBZ (1) XBZ (47)
XM	modal	(479)	100	
XM*	modal + neg.	(85)	100	

Table 3 displays the results of an ANOVA comparing the three groups of children as to tagging accuracy scores either with or without SALT coding. Differences between groups at either the per-tag level or per-utterance level were not significant. Paired *t*-tests comparing scores with or without SALT coding at either the per-tag or the per-utterance level were not significant.

Table 3

Summary Statistics for the Reno Samples: Means and Standard Deviations for Each Group and Level of Coding as well as F-Ratios and Probability Levels

		Tag-Level		Utterance-Level	
Group		GC	SALT	GC	SALT
LI					
	M	96.90	96.92	85.52	85.70
	SD	1.11	1.04	3.91	3.80
LTS					
	M	97.63	97.62	86.89	86.89
	SD	0.70	0.75	4.76	4.85
CA					
	M	97.52	97.58	83.37	83.78
	SD	0.25	0.29	2.26	2.68
F Ratio		2.58	2.64	2.20	1.63
p Level		.09	.09	.13	.21

Table 4 lists the accuracy levels of each child sample at the word and utterance levels for both SALT and non-SALT coded samples. Each child sample is listed according to group (e.g., LI, LTS, or CA), age in months, and number of utterances in the transcript. SALT coded samples achieved accuracy levels at or above GC-tagged samples in all but eight instances at the word

level. This increased accuracy ranged from 0.06% to 0.26%. At the utterance level, four non-SALT coded samples had increased accuracy ranging from 0.34% to 1.16%.

Table 4

Accuracy Measures for Each Participant Sample

Samples	group	age	n_utts	GC tag%	SALT tag%	GC utt%	SALT utt%
Angel	LI	111	175	98.00	98.00	89.14	89.14
Charlene	LI	90	407	97.32	97.36	87.47	87.96
John	LI	111	137	97.00	97.00	88.32	88.32
Marla	LI	104	269	97.05	96.99	85.50	85.87
Melissa	LI	104	381	97.78	97.86	88.98	89.24
Rachel	LI	113	326	97.42	97.42	87.42	87.42
Randy	LI	119	583	97.03	96.95	85.59	85.25
Russel	LI	133	301	96.83	96.72	83.06	83.06
Travis	LI	104	225	96.61	96.69	83.56	84.00
Wilbert	LI	109	231	93.99	94.24	76.19	76.72
Amanda	LTS	91	212	96.40	96.14	78.77	77.83
Desiree	LTS	88	192	98.69	98.69	93.23	93.23
Doug	LTS	95	256	97.10	97.15	81.25	82.03
Erin	LTS	66	272	97.08	97.08	86.03	86.40
Jason	LTS	82	207	98.27	98.20	88.89	88.41
Kimmie	LTS	100	384	97.77	97.87	87.24	87.24
Mario	LTS	69	273	97.40	97.40	87.18	87.18
Nicholas	LTS	77	265	98.14	98.28	91.32	92.08
Pete	LTS	83	424	98.15	98.06	91.98	91.49
Sarah	LTS	84	247	97.25	97.30	83.00	83.00
Alison	CA	90	325	97.52	97.72	84.92	85.85
Jacob	CA	108	279	98.02	98.02	86.02	86.02

Jennifer	CA	106	337	97.52	97.61	85.76	86.05
Kevin	CA	100	374	97.58	97.75	85.56	86.36
Luis	CA	122	241	97.13	97.13	81.74	81.74
Michelle C	CA	110	382	97.82	97.94	83.51	84.82
Michelle K	CA	106	260	97.48	97.35	79.62	78.46
Richard	CA	104	338	97.43	97.50	84.02	84.91
Ryan	CA	132	194	97.42	97.54	81.44	82.47
Shona	CA	110	265	97.32	97.23	81.13	81.13

Discussion

The results of this study indicate that no significant difference exists between the automated grammatical tagging accuracy of SALT coded and non-SALT coded samples. Although the overall accuracy of SALT coded samples was slightly higher than those without SALT coding, there were some individual samples where the opposite was true. This finding is particularly puzzling because SALT coding should only make improvements upon tagging accuracy; that is, SALT coded samples should produce accuracy levels equal or better than their GC counterparts because SALT coding only adds information relevant to grammatical tag assignment.

A potential factor contributing to this predicament is found in the assumptions made by the gc5 software. The gc5 software makes certain assumptions: words beginning with an uppercase are proper nouns (NP) and words ending in */s* are coded as plural. When these two instances co-occur, the software tags the word as a plural noun because there is no specific tag for a plural proper noun (i.e., the category NP includes both plural and singular forms). The topics of *Christmas vacation* and *favorite movies/television programs* were introduced when collecting the language samples from all the subjects in this study (Brinton et al., 1988).

Additionally, subjects had the opportunity to play with different stimulus toys and games (Brinton et al., 1988). Terms such as *Transformers* were used frequently across samples considering this was one of the stimulus toys and a conversational topic related to Christmas presents and/or favorite TV shows. Given that these samples facilitated the use of the term *Transformers* and other such plural proper nouns, it is possible that the assumptions made by the gc5 software contributed to decreased accuracy for SALT coded samples. The results support this finding as tagging accuracy of proper nouns decreased in the SALT coded samples compared to the GC samples.

Considering that difficulties tagging proper nouns may have contributed to decreased accuracy in SALT coded samples, it is also reasonable to suggest that those difficulties affected the tagging of the remainder of the utterance. Instances where a NP was coded incorrectly could have affected subsequent tags due to transitional probabilities.

Human error is clearly another possible factor that contributed to decreased accuracy in the eight SALT coded samples. However, even though human error is a reality in every human endeavor, the results of this study do not support the claim of MacWhinney (2008) that computers are more accurate than human transcribers in noting grammatical details such as those in SALT precoding. SALT precoding, performed by a human transcriber, provided increased accuracy in the majority of cases.

Wilmarth (1997) conducted a similar study using GramCats, an older version of gc5. This older software was only capable of utilizing tag pairs for transitional probabilities instead of the current software, which operates on tag triples. A SALT coded comparison was not completed as part of her study. Using 26 of the 30 samples used in the current study, Wilmarth reported accuracy results at the word and utterance levels. Comparing her results to those of the current

study, children with LI achieved an average of 93.5% accuracy at the word level compared to 96.9% accuracy achieved in the current study. The LTS children achieved 94.5% accuracy at the word level compared to the current study's 97.6%. The CA children averaged 94.0% accuracy at the word level compared to the averaged 97.5% accuracy of this study. Although word level accuracy increased for all three groups in the current study compared to Wilmarth, utterance level accuracy apparently decreased for all three groups, whether with or without SALT coding. This difference ranged from 3.37% to 5.11% for non-SALT coded samples and 3.32% to 5.11% for SALT coded samples. This suggests that while gc5 performed better at the tag level than the software used by Wilmarth, it performed worse at the utterance level. Unfortunately, Wilmarth's analyses are not available to allow this contradiction to be investigated.

Wilmarth (1997) reported struggles in tagging certain grammatical categories. This study demonstrated increased performance in tagging those difficult categories with and without SALT coding. Specifically, gc5 improved tagging accuracy for auxiliary *be*, *get*, and *have* when taking the average of the three groups used in the Wilmarth study.

When comparing those grammatical categories in the current study that were markedly difficult to tag with the findings of Wilmarth (1997), gc5 performed equal or better in every category. In the case of auxiliary *gets*, the difference in performance went from 0% in the Wilmarth study to 83% in the current study.

Channell and Johnson (1999) performed a similar study using an older version of GramCats. In their study, Channell and Johnson used 30 language samples from typically developing children between the ages of 2;6 and 7;11. Similar to the Wilmarth (1997) study, the version of GramCats used by Channell and Johnson utilized tag pairs for transitional probabilities. Automated tagging accuracy at the word level averaged 95.1% compared to the

non-SALT coded accuracy of 97.35% in the current study. Tagging accuracy at the utterance level averaged 77.7% in the Channell and Johnson study and 85.26% in the present study. These findings suggest improvements in the gc5 software compared to this older version of GramCats at both the word and utterance levels.

Grammatical categories specifically listed in the Channell and Johnson (1999) study which were notably difficult to accurately tag included subordinating conjunctives (0%), auxiliary *be* (86%), auxiliary *have* (71%), and auxiliary *get* (0%). The current study improved in all these categories. Similarly, gc5 performed better in the areas of marked difficulty reported in the current study compared to the corresponding performance of Channell and Johnson.

Channell and Johnson (1999) analyzed language samples without SALT coding from typically developing children. When matching the non-SALT coded samples of typically developing children from the LTS or CA categories with the age groups reported in the Channell and Johnson study, accuracy results from the current study were higher at both the word and utterance levels. For example, in the 7;0 to 7;11 category, samples from the current study (n = 5) averaged 97.4% at the word level and 84.2% at the utterance level. Channell and Johnson reported accuracy levels of 94.4% and 67.8% at the word and utterance levels respectively. In the 6;6 to 6;11 age group, gc5 coded samples from this study (n = 3) with 97.8% at the word level and 89.0% at the utterance level. GramCats coded the relevant samples from the Channell and Johnson study with 93.4% word level accuracy and 74.0% utterance level accuracy. For the age groups of 6;0 to 6;5 and 5;6 to 5;11 only one sample corresponded to each in the current study. The sample matching the age range of 6;0 to 6;5 was 98.1% accurate at the word level and 91.3% accurate at the utterance level. Channell and Johnson averaged 95.0% and 72.1% at the word and utterance levels for this age range. Finally, the one sample matching the range of 5;6 to

5;11 using gc5 achieved 97.4% word level accuracy and 87.2% utterance level accuracy compared to 94.3% and 74.9%.

Winiecke (2015) investigated the usefulness of SALT coding using the same samples from the Channell and Johnson (1999) study but used the newer gc5 software. Unlike the current study, Winiecke reported a small but significant improvement in the accuracy of automated SALT coded samples compared to samples without SALT coding. For samples without SALT coding, accuracy averaged 96.5% at the word level and 84.2% at the utterance level. SALT coded samples averaged 96.8% and 85.4% at the word and utterance levels respectively. Although a statistically significant difference was not found in the current study, average performance at both the word and utterance levels were better for both classes of samples compared to Winiecke. Winiecke did not experience the same predicament discovered in this study, in that all SALT coded samples from her study achieved accuracy results equal or better than the corresponding non-SALT-coded samples. As discussed previously, this may be because the samples utilized by Winiecke were not comprised of multiple plural proper nouns; however, this is just speculation as the content of Winiecke's samples have not been reviewed as part of the current study.

Similar to this study, Winiecke (2015) outlined performance at the grammatical category level. The current study obtained accuracy levels at least 10% greater in 11 of the categories outlined by Winiecke for non-SALT coded samples; only two grammatical categories in the Winiecke study achieved better accuracy than the current study when using the same standard. When comparing the SALT coded samples, the current study obtained accuracy levels at least 10% greater in nine of the categories outlined by Winiecke; four grammatical categories in the Winiecki study were better by 10% or more.

Except for Wilmarth (1997), the studies discussed above evaluated the performance of typically developing children. Of particular interest in the practicing world of speech-language pathologists is the resilience of the automated system in tagging individuals with language impairment. There were no significant differences between the accuracy levels obtained between the LI, LTS, and CA groups for both SALT and non-SALT coded samples. This suggests that the gc5 software is just as capable of performing automated tagging for the LI children as the typically developing children used in this study. This finding differed in the Wilmarth study, which found significant differences at the utterance levels between her LI and CA groups and LA and CA groups. Future research is needed to test automated tagging software on larger groups of both language impaired and typically developing children to further understand this relationship.

The results of this study must be interpreted within the purview of its limitations. The sample size consisted of 30 children, which is a relatively small group of individuals. Additionally, there were only 10 children per language group creating an even smaller subset in which to make conclusions based on language functioning. The participants used herein were not necessarily a representative sample of the nation's cultural and linguistic diversity. Participants reflected the population of those who met certain qualifications for study inclusion in the Reno, Nevada area.

Future research is needed to bypass the limitations of the present study. Obviously, studies with larger samples sizes are needed in order to provide more concrete conclusions. Additionally, future research should target culturally and linguistically diverse populations; as clinical case loads continue to incorporate children of various backgrounds, research on grammatical coding of samples from culturally and linguistically diverse children will become

even more valuable. Future studies investigating the accuracy of gc5 in coding different types of samples (e.g., conversational vs. narrative) would provide insight as to which sample type would yield superior coding accuracy. Finally, studies investigating the accuracy of analysis procedures using gc5 are needed in order to provide useful information for clinicians. An excellent tagging scheme is needed for accurate analysis, but tagging alone will not provide clinically useful information for diagnosis and treatment. Studies need to expand the current research to the level of respected clinical analyses such as LARSP, DSS, or IPSyn.

However, in spite of these limitations, the current study offers insight in three areas. First, the findings indicate that the performance of automated grammatical tagging software in accurately tagging child language samples at the word level has greatly improved over the last 15 years, and performance at the at the utterance level has intermittently improved. Second, the effect of SALT coding on tagging accuracy was found to be insignificant, although it generally improved accuracy. Given this finding, it is recommended that clinicians and researchers could better use the time it would take to SALT code previously transcribed samples if they would instead fix automated tagging errors. Third, this study supports the accuracy of gc5 across samples of both typically developing and language-impaired children. As automated grammatical coding software continues to improve, the clinical utility of such measures will increase and ultimately provide an efficient and trusted means for language sample analysis. Such naturalistic sources of information will support enhanced assessment and treatment, the ultimate goal of this inquiry.

References

Brinton, B., Fujiki, M., & Sonnenberg, E. (1988). Responses to requests for clarification by linguistically normal and language-impaired children in conversation. *Journal of Speech and Hearing Disorders, 53*, 383-391.

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, 12,* 369-375. doi: 10.1044/1058-0360(2003/082)

Channell, R. W. (2015). gc5 version 1.0 [Computer software]. Provo, UT: Brigham Young University.

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

Charniak, E. (2000). A maximum-entropy-inspired parser. *Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics.* Seattle, WA.

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

Garvey, C. (1979). An approach to the study of children's role play. *The Quarterly Newsletter of the Laboratory of Comparative Human Cognition, 12*.

Garvey, C., & Hogan, R. (1973). Social speech and social interaction: Egocentrism revisited. *Child Development, 44*, 562–568.

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. doi: 10.3758/s13428-013-0354-x

Heilmann, J. J. (2010). Myths and realities of language sample analysis. *Perspectives on Language Learning and Education, 2,* 4-8.

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

Long, S. H. (2001). About time: A comparison of computerized and manual procedures for grammatical and phonological analysis. *Clinical Linguistics & Phonetics, 15*, 399-426. doi: 10.1080/02699200010027778

Long, S. H., Fey, M. E., & Channell, R. W. (2008). Computerized Profiling (CP; Version 9.7.0) [Computer software]. Milwaukee, WI: Department of Speech Pathology and Audiology, Marquette University. Retrieved from www.computerizedprofiling.org.

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

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.), Vol. 2: The Database. Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (2008). Enriching CHILDES for morphosyntactic analysis. In H. Behrens (Ed.), *Corpora in language acquisition research: History, methods, perspectives* (pp. 165-196), Amsterdam, The Netherlands: Benjamins.

MacWhinney, B. (2015). The CHILDES project: Tools for analyzing talk (electronic edition). Retrieved from http://childes.psy.cmu.edu/manuals/clan.pdf

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.

Wilmarth, J. W. (1997). *Automated grammatical tagging of language samples from children with language impairment.* (Unpublished master's thesis). Brigham Young University, Provo, UT.

Winiecke, R. C. (2015). *Precoding and the accuracy of automated analysis of child language samples.* (Unpublished master's thesis). Brigham Young University, Provo, UT.

Appendix: Annotated Bibliography

Ambridge, B., & Lieven, E. V. M. (2011). *Child language acquisition: Contrasting theoretical approaches* (pp. 1-3, 191-209). New York, NY: Cambridge U Press.

This chapter provides an introduction to the different theoretical approaches related to language acquisition. The authors also consider the necessary steps to grammar acquisition, including understanding of syntactic categories (e.g., noun, verb). The authors discuss the work of Pinker who suggests that children have innate syntactic and semantic categories and begin assigning words to these categories through a linking approach. The authors discuss various limitations of this theory. Another theory, based on the idea that children can identify different syntactical phrases through prosodic patterns and function words, is discussed with weak support for its practicality. The distributional approach to grammar acquisition in which children categorize words based on their order in a sentence is discussed in detail. A phonological theory is presented which postulates that children may use cues like the stress patterns of words to differentiate syntactical categories and identify words that fit into multiple categories. The theories discussed in this chapter are presented with variable support, which displays the need for better grammar acquisition models.

Brinton, B., Fujiki, M., & Sonnenberg, E. (1988). Responses to requests for clarification by linguistically normal and language-impaired children in conversation. *Journal of Speech and Hearing Disorders, 53*, 383-391.

The manner in which language impaired (LI), language age-matched (LA), and chronological age-matched (CA) children responded to neutral requests for clarification was evaluated in this study. Having met specific requirements for inclusion, each group (i.e., LI, LA, and CA) consisted of eight children from the Reno, Nevada area. Each child participated in a 30-minute conversational sample with an adult examiner. The adult examiner included neutral requests for clarification in each sample. One sequence of a neutral request included the questions “Huh?” “What?” and “I didn’t understand that” in subsequent order. The manner in which each child responded was classified into one of five categories: repetition, revision, addition, cue, or inappropriate. Each response was also evaluated according the suprasegmental

and/or gestural cues employed. Among the results described, it was discovered that the LI children provided more inappropriate responses than the other groups. CA children were found to use the cue and addition categories more when compared to the other groups who used increased revisions. However, the authors note that when comparing the responses of the request sequence, LA children did utilize more cues on the third request. CA children used the cue response progressively over the three requests. Cue responses were limited when focusing on the LI group. The three groups did not seem to differ in their utilization of suprasegmental and gestural cues. The authors conclude that a child's ability to respond to a conversational request for repair is sensitive to language impairment. The linguistic immaturity of the LI group was reflected in comparison to the CA group; however, differences between the linguistically similar LI and LA groups were also manifest in this experiment indicating that linguistic immaturity alone is not able to explain all differences in the LI children.

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

This study evaluated the accuracy of automated Developmental Sentence Scoring using Computerized Profiling. Language samples were taken from both typical and language impaired children from the Reno, Nevada area and Jordan school district in Salt Lake County, Utah. The automated DSS accuracy was 78.2% (SD = 4.4). Overall, the lower point values of the grammatical categories were more accurate. Findings included that there was not a distinct pattern related to inclusions or misses for the computerized software. The Reno-LI group was statistically different from the Reno-LA and Reno-CA groups. Finally, although the accuracy of the automated DSS was calculated at 78.2%, the correlation between the automated and manual DSS was high (r = .97, p < .0001). The accuracy levels of the different grammatical categories and point values are compared to a standard of acceptable (greater than 85%), good (greater than 90%), and excellent (greater than 95%). Channell reports that only 11 out of 36 point categories obtained an acceptable level or above. Different relationships were explored including the relationship between accuracy scores and sample size. Further research is suggested to help improve the accuracy levels of automated DSS, which will make the process faster and more convenient for clinicians.

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

This study applied probabilistic automated grammatical tagging to child language samples. The GramCats software was used to tag 30 child language samples. Tagging was completed at the word level using 75 different tags. Of the probability matrices used, six were created based on 25 of the 30 language samples used in this study, and one was created based on a sample not used in the study. Accuracy at the word-by-word and utterance levels ranged from 92.9% to 97.4% and 60.5% to 90.3% respectively. There was a significant negative correlation between age and accuracy of automated coding. Reliability was fairly similar for both the automated system and manual coding with 95.1% and 95.4% to 97.6% respectively. The authors make note that although reliability was fairly similar the error types were not. A limitation of the present study was the inability of the tagging system to more accurately code a word due to what the authors term *limited look-ahead capabilities*. Previous research has shown elevated accuracy levels when tag triples (as opposed to tag pairs) are used to code adult language samples. Increased coding accuracy in automated tagging software is needed for language sample analyses to provide more useful and accurate information for diagnostic and treatment purposes.

Chemla, E., Mintz, T. H., Bernal, S., & Christophe, A. (2009). Categorizing words using 'frequent frames': What cross-linguistic analyses reveal about distributional acquisition strategies. *Developmental Science, 12*, 396-406. doi: 10.1111/j.1467-7687.2009.00825.x

The distributional approach to grammatical category acquisition using frequent frames has had some success in explaining how children gain linguistic knowledge. The authors aim to test the validity of frequent frames in the French language while exploring the resilience of other frequent frame-like patterns. The effect of a recursive application on the classification ability of frequent frames is also investigated. Using French samples, accuracy and completeness scores were above chance levels. Accuracy and completeness scores were subsequently calculated for frames that met a certain frequency criterion. For this calculation, accuracy scores were perfect and completeness scores were still above the level of chance. Results indicate that frequent

frames can be applied to French as well as English despite differences in the French vernacular. Accuracy levels of frequent back and front contexts in French and English were found to be above the level of chance but much lower than the accuracy levels of the frequent frames. Completeness scores were much lower in both French and English. Reportedly, English frequent front contexts were the only completeness score above chance level. These results indicated that the discontinuity inherent in a frequent frames approach results in superior accuracy and completeness scores when compared to frequent front or back contexts; this may be a result of the greater restriction imposed by a discontinuous frame to the syntactic patterns possible. The accuracy of the frequent recursive frames was above the level of chance for both French and English, but comparatively lower than the initial frequent-frames analysis. This revealed that it is necessary to consider the specific words that make up the frequent frame rather than the syntactic categories of the words for more accurate grammatical categorization.

Cluff, S. Z. (2014). *A model of grammatical category acquisition using adaption and selection.* (Unpublished master's thesis). Brigham Young University, Provo, UT.

This study uses an evolutionary approach to understand grammatical category acquisition. The adaptation and selection model overcomes previous limitations of assigning words to only one grammatical category. The five corpora used in the study were taken from the CHILDES database. The program cycled through 4,000 times. Four different mutation rates were selected (i.e., 1/400, 1/800, 1/1200, 1/1800), and every corpus was run through each. The number of grammatical category tags was specified in accordance with the known average for a given corpus. The mutation rate of 1/400 proved to be the least accurate across all corpora. Although the mutation rate of 1/400 showed the greatest gains in accuracy levels initially, it was later surpassed by all other mutation rates. Overall, as the mutation rate decreased the accuracy of coding grammatical tags improved. Accuracy levels decreased as the children aged. Consequently, the authors speculate that the language presented to children at different ages may have an affect on their ability to acquire grammatical categories. The corpora with the highest accuracy were also noted to have the greatest number of grammatical tags per word. Further research using this computational model of grammatical category acquisition is warranted.

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

The author of this paper outlines how the algorithm termed VOLSUNGA provides a solution to the problem of grammatical category ambiguity evident in computer tagging software. Previous attempts by researchers to address the issue of grammatical category ambiguity are presented. Contrasting VOLSUNGA to the previously established CLAWS, the author describes that VOLSUNGA operates linearly while CLAWS operates exponentially. The issue of exponential operation in CLAWS is that it requires significant time and space. The author also notes that VOLSUNGA does not include accessory processing, which requires special-case lists such as those needed to perform idiom tagging under CLAWS. VOLSUNGA also utilizes Relative Tag Probabilities (RTPs) in their raw form compared to the scaled nature of CLAWS. VOLSUNGA was tested on the Brown University Corpus. The Brown Corpus comprised the reference dictionary used by VOLSUNGA and provided the basis for RTPs. VOLSUNGA achieved a total accuracy of 96.04%. Comparatively, CLAWS achieved 96-97% accuracy on the Lancaster-Oslo/Bergen Corpus. The author also discusses parallels between VOLSUNGA and human language processing. Decreasing the time and space needed while still maintaining a high level of accuracy, VOLSUNGA is a useful algorithm for future implementation.

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. doi: 10.1044/2014_LSHSS-14-0034

This study investigated the relationship between norm-referenced language tests and narrative language samples in the context of language assessment. The authors evaluated four domains, including the factor of age on the association between norm-referenced tests and narrative language samples, the factor of language level on the association between the two measures, the relationship of written language tests and narrative language samples, and the ability of norm-referenced tests and narrative samples to similarly identify individuals with and

without language impairment. This study was a retrospective analysis; participants included individuals who previously participated in a language assessment comprised of both a narrative language sample and a norm-referenced test or subtest. The participants were separated into two groups: an older group ranging between 9;1 and 12;8 (years;months) and a younger group ranging between 6;0 and 8;11. The narrative sample was collected by way of a storybook stimulus comprised solely of pictures. A variety of norm-referenced tests and subtests were included in this study. Different measures such as MLU (mean length of utterance) and NDW (number of different words) were calculated from the language sample. Partial correlations between the norm-referenced tests and narrative samples revealed many more correlations of statistical significance in the younger group than that of the older group. The authors speculated that this could reflect the restrictive nature of the narrative task for the older children. Using the younger age group, the authors evaluated the factor of language level. They discovered no significant correlations between word levels, six correlations between sentence levels, five correlations between the word and sentence levels, and 1 correlation between the sentence and discourse levels. There were multiple correlations between the CELF subtests and the narrative measure that included grammatical errors. Due to this pattern, the authors suggested that evaluating grammatical errors in language samples could be useful in assessment. When evaluating the relationship of written language tests and narrative language samples, the older and younger groups both had a significant correlation based on a total of 40; the older group had a correlation of $r(7) = .99$, $p < .001$ at the discourse level. Norm-referenced tests and narrative samples displayed variable agreement in their assignment of language impairment. The variability in agreement, however, supports the position that norm-referenced tests and language samples should be used conjointly in language assessment. The authors noted that only composite scores were used to determine the presence or absence of language impairment. They also discussed that different cut-off levels of impairment improved or worsened agreement for a given measure. Future research may continue to explore the relationship between these two important aspects of assessment.

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.

This article outlined the various limitations regarding the assessment of Spanish speaking children with varying levels of second language (L2) proficiency. The authors pointed out that there is a lack of reference data and developmental data in which to compare results of assessment for this heterogeneous group. Language assessment is particularly challenging for Spanish speaking children because level of L2 proficiency and time/type of exposure affect the child's language picture. The authors discussed the potential impact of *language shift* on this picture and the importance of comparing the child to a relevant population. For example, the authors noted that comparing a Spanish-speaking child with a significant amount of L2 exposure to monolingual English speaking children or to Spanish speaking children with limited to no L2 exposure is inappropriate. Additionally, the authors pointed out the affect of timing on language acquisition, noting that bilingual learners, second language learners, and single language learners may each acquire language differently. Measures of language sample analysis for Spanish speaking children were explored including the Developmental Assessment of Spanish Grammar (DASG; Toronto, 1976), grammatical errors per T-unit, mean length of response in words (MLR-w), mean length of terminable unit (MLTU), and mean length of utterance in morphemes (MLU-m). Relevant limitations were noted for these measures. The lack of methodological consistency in calculating MLU-m was explored as different researchers computed this measure differently for Spanish language samples. The authors also emphasized the natural occurrence of codeswitching in the language of bilingual individuals; they stressed that assessment measures must take this occurrence into account as a typical, not atypical, process. MLU-w, as opposed to MLU-m, was cited as a measure that could generally take codeswitching into account without skewing the results of a child's language performance. Issues inherent in using MLU-m with different dialects were explored as well. Appropriate assessment is needed for Spanish speaking children who exhibit a range of L2 proficiency. Certain assessment protocols were outlined for children with different levels of L2 proficiency. Areas of future research were listed.

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. doi: 10.3758/s13428-013-0354-x

According to the authors, the purpose of this study is to create an automated system for calculating the Index of Productive Syntax (IPSyn) that can be accessed by researchers without cost which has the capability of performing hundreds of analyses simultaneously. The authors intend to make the automated system capable of utilizing both CHAT and SALT format. The system should also provide more information regarding the language sample including the number of occurrences of each IPSyn form. The Automatic Computation of IPSyn system (AC-IPSyn) was evaluated on a point-by-point and point difference basis. Two data sets were evaluated: set A was comprised of samples from children ages two to three and set B from a combination of developmentally normal and language impaired children of age six. Compared to manual coding, the AC-IPSyn system had a point difference of 3.05 (SD = 2.14) and a point-by-point accuracy of 96.9% for set A. Set B had the same point difference but an accuracy of 96.4%. The AC-IPSyn system displayed improved results compared to the Sagae and CP systems. Discrepancies in IPSyn scoring calculations were largely a result of part-of-speech tag and parse tree errors.

Heilmann, J. J. (2010). Myths and realities of language sample analysis. *Perspectives on Language Learning and Education, 2,* 4-8.

Heilmann discusses myths and facts regarding language sampling and analysis in efforts to encourage its clinical use in the field of speech-language pathology. The first misconception is that learning to take a language sample is overly complex. Heilmann includes references to language sampling software that can be of aid to clinicians who wish to complete language samples. The second misconception is that there is not enough time to complete a language sample analysis. Heilmann discusses the reality that language samples can be collected in a relatively short amount of time. The third myth is that language samples are not reliable and do not provide an accurate reflection of a child's language abilities. As with standardized measures, clinicians have to provide good judgment regarding whether or not the performance accurately

reflected the child's abilities. Data is presented supporting the reliable nature of language sample compilation. The fourth misconception is that one must have extensive knowledge in linguistics to draw conclusions from a language sample. The author presents simple analysis tasks that can be completed and notes that computer software provides valuable resources related to language sampling. The next myth is that taking a language sample online is acceptable. The limitations of this method are addressed including the limited research to support this approach to language sampling collection. The sixth misconception is that language samples can be collected whenever an opportunity is presented. The author stresses that language samples can only be compared to others collected in the same manner; performance is affected by context. The last myth presented is that language samples will fix everything for clinicians. Language samples are supported in the research and can be a useful tool in evaluating language. However, clinicians need to consider each child individually and select the best methods of assessment for a given child. Overall, language sampling can be a useful measure to evaluate language production.

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)

This study investigated the extent with which shorter language samples produced accurate and reliable results compared to longer language samples considering variables such as context and age. Conversational and narrative samples were collected from children ranging in age from 2;8 to 13;3 who were separated into a younger age group (2;8 to 5;11) and an older age group (6;0 to 13;3). Of these samples, segments of 1, 3, and 7 minutes were evaluated. Analysis measures included *number of total utterances* (NTU), *words per minute* (WPM), *number of different words* (NDW), *mean length of utterance in morphemes* (MLUm), *percentage of maze words* (% mazes), and *errors and omissions per minute* (errors & omissions). Results indicated no significant differences in analysis measures among all of the relationships explored including *Sample Length*, *Age Group*, and *Sampling Context*. To explore reliability, the authors cite their use of Cronbach's alpha tests (Cronbach, 1951) utilizing the 7-minute segment as the comparison for the other two short segments. Comparing the two short samples, alpha values were generally better for the 3-minute segment. The strength of the alpha values varied by analysis measure with

the strongest values reported overall for NTU, WPM, and NDW; generally weakest values were reported for % mazes and errors & omissions. For reference, Table 2 lists each measure's alpha value according to sample length, context, and age group. When evaluating the variability between sample lengths using coefficients of variation, it was discovered that an increase in variability was generally associated with a decrease in sample length. The authors elaborated on the limitations of their study and the implications of the findings for future clinical practice.

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

The purpose of this study is to evaluate the effectiveness of automated grammatical tagging while using the VOLSUNGA algorithm (DeRose, 1988) on samples that differ from the Brown University Standard Corpus of Present-day American English. The authors used the automated tagging system of GramCats (Channell, 1991), which implements the VOLSUNGA algorithm to perform this analysis. Additionally, this study evaluated the effectiveness of different transitional probability measures on automated grammatical tagging. A variety of language samples were obtained including edited written samples and unedited written and spoken samples from adults and children. Relative Tag Probabilities utilized by GramCats were compiled from the Brown University Corpus. Three different Transitional Tag Probabilities were used. These transitional probabilities were based on the Brown University Corpus, on samples comparable to the ones used in this study, and on the samples used in this study. The manual and automated tagging results were compared at the token level for each Transitional Tag Probability matrix. Accuracy of automated grammatical tagging for the variety of language samples ranged from 89.47%-94.49%, which was less than the accuracy results reported previously when only utilizing edited adult written samples. The effect of the transitional probability on the accuracy of grammatical coding was found to be statistically significant with transitional probabilities based on similar samples statistically improved compared to probabilities based on the same samples used. The transitional probabilities based on the same samples were also significantly better than the probabilities based off of the Brown University Corpus. The highest accuracy of automated tagging was achieved when the transitional probability was based on a group of similar samples

indicating that future tagging systems base the transitional scheme off of similar samples to achieve maximum results.

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*(2), 161-176. doi: 10.1177/026565909701300204

The purpose of this study was to investigate the current tendencies of practicing clinicians in the domain of language sample analysis. By way of a mailed questionnaire, 500 speech-language pathologists practicing with the preschool population were randomly selected to complete the survey. Of the 500 surveys sent to clinicians throughout the United States, 253 were returned for inclusion in the study. Results indicated that most clinicians completed language sample analyses on language-impaired children in their practice (85%). Various reasons were presented as to why some clinicians did not use language sample analyses, with time constraints identified most frequently (86%). For clinicians who did complete language sample analyses, online transcription (59%) was practiced more than tape or video recording and subsequent transcription. Collection of a 50-utterance sample was a common practice. With regard to the analysis portion, most clinicians performed nonstandardized assessments. Among the various standardized forms of analyses utilized, DSS was most commonly cited. Only 8% of clinicians capitalized on automated programs for analysis, with SALT most frequently cited. The authors provide various suggestions to help increase the usage and efficiency of language sample analysis, including the increased use of computers to expedite the process.

Klee, T., Membrino, I., & May, S. (1991). Feasibility of real-time transcription in the clinical setting. *Child Language Teaching & Therapy*, *7*(1), 27-40. doi: 10.1177/026565909100700102

The purpose of this study was to discover the degree in which real-time transcription (RTT) and subsequent analyses matched audiotaped transcription (ATT) and its analyses. Understanding this relationship was meant to affirm or negate the usefulness of RTT in providing immediate objective information regarding the language functioning of a child

following language sampling. Conversational language samples were collected from 22 children as part of an evaluation or re-evaluation service. Only 20 samples were used in this study based on qualification measures of MLU, complexity, and intelligibility. Samples were transcribed in real time and with audiotaped recordings by two speech-language pathologists experienced in this discipline. Even though all measures evaluated between RTT and ATT were highly correlated, only two measures were not significantly different. Results of this study indicate no significant difference between RTT and ATT in identifying overall intelligibility and MLU. Comparing the RTT and ATT measures of MLU to group data, all corresponding samples matched with regards to placement within or outside the normal range (i.e., 1 SD). No significant difference was discovered regarding the transcription abilities of the two speech-language pathologists when analyzing the mean MLU difference between RTT and ATT. The authors note that while RTT may provide some immediate feedback, including quantitative data, regarding the language functioning of a child, it should undergo later correction using audiotaped recordings.

Long, S. H. (2001). About time: A comparison of computerized and manual procedures for grammatical and phonological analysis. *Clinical Linguistics & Phonetics, 15*, 399-426. doi: 10.1080/02699200010027778

This study investigated the time commitment of phonological and grammatical analyses given different variables such as expertise in analysis and severity of the sample. This study also compared manual and computerized analyses for efficiency. Participants consisted of 256 individuals including students and clinical speech-language pathologists. All participants received training on the analyses they conducted. Computerized Profiling was used to complete the computerized analyses. Each participant first conducted the computer analysis and then the corresponding manual analysis. On a comparison of accuracy, computerized analyses were consistently deemed more accurate than manual attempts. The amount of time needed to complete manual phonological analyses was affected by variables such as sample size and complexity. For phonological analysis, sample type had a different impact on efficiency when comparing manual and computerized efforts. The gap in time difference between manual and computerized grammatical analyses was less than that of phonological analyses. However, the

time commitment for grammatical analyses was impacted more by the efficiency of the clinician. With limited exceptions, the least complex grammatical sample was the fastest to analyze. Comparing the time needed to complete manual versus computerized grammatical analyses, there were highly significant correlations for the complex procedures (i.e., LARSP, DSS, and IPSyn). Manual language sample analysis can take long periods of time depending on the complexity of the sample, the efficiency of the clinician, and the analysis procedure utilized. Computer software provides more efficient means for language sample analysis.

Long, S. H., & Channell, R. W. (2001). Accuracy of four language analysis procedures performed automatically. *American Journal of Speech-Language Pathology, 10*, 180-188. doi: 10.1044/1058-0360(2001/017)

The purpose of this study was to discover the automated accuracy of four different language analysis procedures in comparison to manual efforts. Computerized Profiling, with GramCats functioning as the automated tagging software, was used to calculate MLU, LARSP, IPSyn, and DSS scores for conversational child language samples. In this study 69 samples were used which included those of typically developing children, children with Specific Language Impairment (SLI), children with Specific Expressive Language Impairment (SELI), and children who met a certain stuttering criteria. Each analysis procedure was first performed automatically as Condition 1. The automated analysis was then corrected by hand, and results were classified as Condition 2. On a point-by-point basis, accuracy for LARSP, IPSyn, DSS, and MLU were respectively 85.2%, 91.4%, 90.0%, and 99.3%. The authors of this study compared these findings to acceptable, good, and excellent levels with LARSP attaining acceptable (above 0.85), IPSyn and DSS achieving good (above 0.90), and MLU reaching excellent (above 0.95). In comparing the automated findings from this study to previous studies in which accuracy levels were achieved manually, MLU accuracy exceeded previous findings, and IPSyn and DSS accuracy scores were similar to previous calculations. The authors of this study also compared Condition 1 and Condition 2 of the current study to reference information for MLU, IPSyn, and DSS. Using the reference information, they concluded that the results obtained from automated conditions used in this study to calculate MLU were comparable to manual efforts; automated IPSyn calculations were very similar to Condition 2, but manual speculation should take place when scores are close to the standard cutoff; and manual review and correction should follow

automated DSS. Automated LARSP calculations were believed to convey information similar to manual efforts regarding future treatment targets. Language analysis assisted by computer technology holds promising results for future clinical and research use.

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*(2), 221-229. doi: 10.1177/0265659013519251

This pilot study investigated whether speech and language therapy assistants (SLTAs) could reliably transcribe language samples for automated analysis using SALT software. Language samples were collected from 15 children with language impairment at the start of a treatment period (T1) and at the end of a treatment period (T2). An SLTA was selected to transcribe all samples using SALT formatting for subsequent automated analysis. The SLTA underwent two training periods, the second of which occurred following T1 transcription. Two of the T1 transcripts were used for training purposes. Inter-rater reliability was established at a mean of 89% for the 13 samples of T1 and 93% for 4 samples randomly selected from T2. Four separate areas of SALT formatting were additionally evaluated with mean reliability ranging from 62% to 95%. Specific areas of conflict, which contributed to decreased reliability levels, were discussed including the most prominent of maze boundary errors. Although further research is needed, findings suggest the potential for utilizing SLTAs for language transcription and automated analysis: a practice that would save time for clinicians while still providing a naturalistic picture of a child's language abilities during assessment and follow-up.

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.

The purpose of this paper was to display the benefit of automated tagging systems for language analyses. Specifically, the Grammatical Relations (GR) automated system is explored.

Precision results of this process were calculated at 86.9% overall. The GR tagging process is outlined which includes running samples through MOR and POST, which evaluates morphology and tags part of speech respectively; filtering the samples through constituent trees to derive unlabeled dependencies; and finally, assigning GR labels using a trained classifier. In this study, IPSyn analyses are computed through the GR automated system, manual coding, and CP and then compared using point difference and point-to-point accuracy. Samples included those in the age ranges of two to three years and eight to nine years. On average, GR differed from manual coding by 3.3 points while CP differed from manual coding by 8.3 points. CP displayed a greater difference between the two age ranges explored compared to GR. Mean point-to-point accuracy for GR was 92.8%, with CP accuracy at 85.4%. IPSyn scores more closely matched manual efforts when using the GR automated system compared to CP. Although positive results are evident using the GR system, improvements are still needed.

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*, 325-335.

The authors of this study investigated the performance of three automated tagging systems when applied to a portion of the Tswana Learner English Corpus (TLEC). The authors also evaluated the effect of what they term *learner errors* on the tagging accuracy. The three automated tagging systems differed in the number of tags they used and the way they operated. Results indicated superior performance of CLAWS (96%) compared to the TOSCA-ICLA (87%) and Brill taggers (89%). The authors noted that spelling errors were an issue, but were relatively easy to correct. Once spelling errors were corrected, the samples were automatically tagged again. CLAWS (98%) continued to perform better than the TOSCA-ICLA (90%) and Brill taggers (91%) in the spelling-corrected version. In addition to spelling errors, the authors discovered other learner errors affecting tag accuracy. These errors were separated into 10 categories. Learner errors, including spelling errors, accounted for 34%, 25%, and 67% of tagging errors for the TOSCA-ICLA, Brill, and CLAWS taggers respectively. When spelling errors were corrected, learner errors accounted for 19%, 14%, and 38% of tagging errors for the TOSCA-ICLA, Brill, and CLAWS taggers respectively. The authors noted that while learner

errors do affect tagging accuracy, there are still a number of tagging errors not associated with learner errors.

Vine, E. W. (2011). High frequency multifunctional word: Accuracy of word-class tagging. *Te Reo*, *54,* 71-82.

The accuracy of automated tagging programs in correctly classifying high frequency words with multiple tag possibilities was investigated in this study. The Constituent Likelihood Automatic Word-tagging System (CLAWS) and Douglas Biber's automated program were compared to manual tagging efforts. This study only evaluated the tagging accuracy of the words *like, so,* and *as*. Samples from the Wellington Corpora of Spoken New Zealand English (WCSNZE) and the Wellington Corpora of Written New Zealand English (WCWNZE) were used in this study. Within the WCWNZE samples, the overall error frequency of CLAWS in tagging *as*, *like*, and *so* was 54.0%, 6.5%, and 23.0% respectively. The Biber program exhibited error rates at 78.0% for *as,* 33.0% for *like*, and 53.0% for *so.* The multifunctional word *like* was explored in more detail. The confusions of each program in tagging the various classes of *like* were presented. CLAWS demonstrated particular success in coding *like* according to the grammatical categories of preposition and verb. Both CLAWS and the Biber program displayed difficulty accurately tagging *like* according to the classes of conjunction and noun. Automated tagging utilizing CLAWS was not performed on the WCSNZE; however, error rates were reported for the Biber program at 89%, 67%, and 88% for *as, like,* and *so* respectively. Additionally, the accuracy in tagging specific grammatical categories and associated confusions were reported. The higher error rates associated with tagging fiction prose in the WCWNZE samples and the overall poorer performance using the WCSNZE samples were attributed to the possibility that the Biber program lacked flexibility in tagging dialogue and spoken language compared to its CLAWS counterpart. Results displayed the difficult nature of tagging high frequency, multi-class words accurately using an automated system. Individuals should be careful to not assume high accuracy for these types of words when automated systems are utilized.

Wilmarth, J. W. (1997). *Automated grammatical tagging of language samples from children with language impairment.* (Unpublished master's thesis). Brigham Young University, Provo, UT.

This study aimed to discover the accuracy of automated grammatical tagging on spoken language samples of children with language impairment in order to increase clinical usefulness. Language samples from 26 children were used in this study. Comprised of language impaired children (LI), language age matched children (LA), and chronologically age matched children (CA), these language samples were collected in a previous study. GramCats was used to automatically tag each language sample. Accuracy results at the word level were averaged at 93.5% (*SD* = 1.9), 94.5% (*SD* = 1.8), and 94.0% (*SD* = 0.80) for the LI, LA, and CA groups respectively. Accuracy results at the utterance level were averaged at 90.4% (*SD* = 2.8), 92.0% (*SD* = 3.6), and 87.1% (*SD* = 2.4) for the LI, LA, and CA groups. A significant difference was found regarding the number of utterances automatically tagged correctly between the LI and CA samples as well as the LA and CA samples. A negative correlation was discovered related to the number of words per utterance and accuracy in automated grammatical tagging of utterances. Previous research indicates decreased tagging accuracy on older children compared to younger children, which matches the findings of this study as well. Specific difficulties experienced by the GramCats software are explored including challenges coding certain tag schemes. Future studies should consider increasing the transitional probability from tag pairs to tag triples.

Winiecke, R. C. (2015). *Precoding and the accuracy of automated analysis of child language samples.* (Unpublished master's thesis). Brigham Young University, Provo, UT.

The purpose of this study was to determine the effect of SALT coding on the accuracy of automated grammatical tagging and discover the effect of modern computer capacity on such tagging accuracy. Participants consisted of 30 developmentally normal children in the Provo, Utah area between the ages of 2;6 and 7;11. The language samples were run through the automated tagging software of gc5 with and without SALT coding. Accuracy results were obtained by comparison to the manually tagged versions at the word and utterance levels. SALT coded samples achieved 96.84% and 85.4% accuracy while samples without SALT coding

achieved 96.54% and 84.18% accuracy at the word and utterance levels respectively. These statistically significant results indicated increased accuracy with SALT coding. At the utterance level, the relationship between MLU and number of utterances indicated a positive correlation for samples with and without SALT coding. Additionally, the relationship between MLU and utterance level accuracy indicated a negative correlation for samples with and without SALT coding. Compared to the software utilized in the Channell and Johnson (1999) study, the updated gc5 software demonstrated overall improved accuracy on samples without SALT coding by 1.4% and 6.5% at the word and utterance level respectively. Despite improvements in accuracy with SALT coding, the author supports the notion of fixing computer errors rather than SALT coding previously automated samples as a more productive use of time. However, the author does recommend SALT coding samples during initial transcription, as this requires minimal additional effort. Areas for future research are outlined including the effect of SALT coding on the automated tagging accuracy of child samples with language impairment.