2003

Automated Grammatical Tagging of Language Samples from Children with and without Language Impairment

Deborah Millet
Brigham Young University - Provo

AUTOMATED GRAMMATICAL TAGGING OF LANGUAGE SAMPLES FROM CHILDREN WITH AND WITHOUT LANGUAGE IMPAIRMENT

by

Deborah Millet

A thesis submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Master of Science

Department of Audiology and Speech-Language Pathology

Brigham Young University

April 2001

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a thesis submitted by

Deborah Millet

This thesis has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

Date	Ron W. Channell, Chair
Date	Bonnie Brinton
Date	Martin Fujiki

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the thesis of Deborah Millet in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

Date

Ron W. Channell
Chair, Graduate Committee

Accepted for the Department

Ron W. Channell
Department Graduate Coordinator

Accepted for the College

Robert S. Patterson
Dean, David O. McKay School of Education

ABSTRACT

AUTOMATED GRAMMATICAL TAGGING OF LANGUAGE SAMPLES FROM CHILDREN WITH AND WITHOUT LANGUAGE IMPAIRMENT

Deborah Millet

Department of Audiology and Speech-Language Pathology

Master of Science

Grammatical classification ("tagging") of words in language samples is a component of syntactic analysis for both clinical and research purposes. Previous studies have shown that probability-based software can be used to tag samples from adults and typically-developing children with high (about 95%) accuracy. The present study found that similar accuracy can be obtained in tagging samples from school-aged children with and without language impairment if the software uses tri-gram rather than bi-gram probabilities and large corpora are used to obtain probability information to train the tagging software.

ACKNOWLEDGMENTS

I thank the faculty at BYU for providing me with valuable instruction as well as being examples of careful, committed research. I also thank my family and friends for their support of my thesis and education. Finally, I thank Dr. Ron Channell for his influence, example and guidance. I will forever value the counsel and friendship he generously gave me.

TABLE OF CONTENTS

Page

LIST OF TABLES

Introduction

Being able to describe and quantify the level of syntactic development is valuable for the diagnosis of language impairment and for the assessment of progress toward structural language goals in the clinic. However, studies have shown that clinicians often lack the time or the grammatical training necessary to perform reliable, accurate analyses (Hux, Morris-Friehe, & Sanger, 1993; Kemp & Klee, 1997; Long, 1996). Computer software has been used to increase the speed of grammatical analysis (Long, 1991; Long & Fey, 1995). Although automated grammatical analysis software may be useful, few data have been obtained regarding the accuracy of these programs.

One approach that offers accuracy data for automated grammatical analysis uses probabilistic methods to statistically analyze syntactic structure. Software based on this approach uses two types of probability data extracted from a training corpus. The first is the likelihood that a particular grammatical tag describes the use of a word. For example, the word *can* is far more likely to be an auxiliary verb than a noun. The second type of probability is the likelihood that one tag will follow another tag. For example, a noun is far more likely to follow an adjective than is a verb. A tag dictionary is created from the training corpus containing words, their tag options and the frequency of each, and the frequency of tag options. Probabilistic methods provide high levels of accuracy when applied to adult language (DeRose, 1988; Garside, Leech, & Sampson, 1987). Recently, probabilistic methods have been applied to the analysis of naturalistic language samples from typically developing children.

Channell and Johnson (1999) used software to computer-tag 30 naturalistic samples from children between the ages of 2;6 (years;months) and 8;0. Their software, GramCats, used probability information extracted from a training corpus of 5,000 manually-tagged utterances. These authors compared the computer-tagged and manually-tagged samples both by word and by utterance. Word-by-word accuracy averaged 95%. Tagging accuracy was negatively correlated with age; accuracy on tagging samples from school-aged children was about 2% lower. Accuracy on an utterance-by-utterance basis, where every tag in the utterance had to be correct, averaged 78%. Areas that had lower accuracy in the automated grammatical analysis were subordinate clauses, the ambiguity of auxiliary verb versus main verb, uses of *be, have, do,* the word *that*, and the form *'s*. When a word appeared in the test corpus that was not in the training corpus dictionary, the word was manually added to the program's dictionary.

Channell and Johnson's (1999) study showed probabilistic methods of automated grammatical analysis to yield high levels of tag-by-tag accuracy even for the analysis of language samples collected from children in non-standardized conditions. These authors also mentioned several ways that the accuracy of automated grammatical tagging might be improved and the generality of this approach expanded.

The present study expanded upon and advanced the work of Channell and Johnson in five ways. First, new software, entitled gc3 (Channell, 2000), was written to use tri-gram rather than bi-gram probability information. This new software was motivated by the belief that expanded tag sequence information would provide additional grammatical context for resolving ambiguous tags. When tagging unknown

words, tri-gram probability information determines the most likely tag option for the unknown word based on the tag of the previous two words.

Second, the gc3 software applies a “guessing rule” for unknown words was applied. Although manually adding new words may be acceptable to researchers, adding every unknown word to the program's dictionary would not be practical for clinicians. When a word appears that was not in the training corpus dictionary, a “guessing rule” is used to select a tag for that word. Data obtained on the frequency with which gc3 or similar programs would need to use a guessing rule are of clinical interest.

Third, the present study investigated the use of training corpora larger than the 5,000 utterance size used by Channell and Johnson (1999). In the present study, the training corpus was used to create dictionaries ranging in size from 5,000 to approximately 20,000 utterances. This was based on the notion that a larger training corpus dictionary would decrease the need to use the guessing rule and would lead to fewer tagging errors.

Fourth, the gc3 software uses a tagging scheme which is clinically well known. Current software, such as GramCats used by Channell and Johnson, uses tagging code systems which are unfamiliar to most clinicians. The scheme used in the gc3 software was adapted from the Language Assessment, Remediation and Screening Procedure (LARSP; Crystal, Garman, & Fletcher, 1989) approach. Presumably, probability methods should not be dependent on a specific tagging scheme; therefore, incorporating a more familiar scheme would make the automated analysis clinically more applicable.

Fifth, the automated grammatical analysis was applied to samples from children with and without language impairment. In addition to improvements in the grammatical tagging software, data on the accuracy of the use of probabilistic methods to perform automated grammatical analyses of children with language impairment are of clinical interest. These automated methods may offer clinical potential if high accuracy levels can be maintained when applied to children with language impairment.

Review of Literature

Clinical methods of measuring syntactic development in spontaneous language samples vary from quantitative measures such as mean length of utterance (MLU), to Developmental Sentence Scoring (DSS; Lee, 1974), to completely qualitative measures such as the LARSP (Crystal et al., 1989).

Language Sample Analysis

Hux, Morris-Friehe, and Sanger (1993) surveyed school-based speech-language pathologists about their language sampling practices. Hux et al. found language sampling was routinely used in assessment and therapy planning. Respondents used language samples to supplement standardized procedures (80%), assist program planning (77%), document treatment effectiveness (62%), and to evaluate program effectiveness (54%). Sixty percent of respondents reported collecting 51 to 100 utterances per sample. However, Hux et al. did not report how many utterances were used in analysis procedures. In addition, the number of utterances collected was not broken down from the 51 to 100 level leaving exact totals ambiguous. Respondents who reported their most preferred procedure used either a

self-designed analysis or DSS. Only 3% of respondents used a computer analysis. Hux et al. reported time, required expertise, and financial limitations as complaints, and concluded, “school speech-language pathologists recognize the importance of language sampling” (p. 90) but may rely on self-designed analyses which will “forfeit the benefits offered by standardized procedures with respect to reliability and validity” (p. 90).

Long (1996) reviewed the grammatical competency of English and Speech Language Pathology students. Because of poor syntactic skills, some students cannot interpret standardized language tests, do not understand nor can identify grammatical constituents, and “make minimal use of computer software for language sample analysis” (p. 36). Because of language constituent complexity and time constraints, students may be getting substandard instruction and not enough practice in analysis. Long stated that the field of speech-language pathology “has been slow to accept, much less require, the use of computers” in the classroom and clinic (p. 37).

Kemp and Klee (1997) surveyed 253 speech-language pathologists on their child language assessment practices, particularly on language sample analysis. This survey showed that 85% of respondents used “a language sample analysis to assess children with language impairment” (p. 164). Clinicians reported using a language sample analysis for diagnosis (92%), intervention (77%), post intervention (64%), and screening purposes (44%). Kemp and Klee found that the reasons reported for not performing a language sample analysis were lack of time (86%), lack of computer resources (40%), lack of training or expertise (16%), and financial constraints (15%). Almost half of the respondents reported using samples containing 50 utterances, 28%

used less, and 24% used more than 50 utterances. Thirty five percent of clinicians reported using DSS to analyze the language sample, 29% used Lahey's (1988) Content/Form/Use, and 10% used other systematic procedures. Kemp and Klee's survey also probed problem areas regarding language sample analysis. Respondents' concerns were lack of time and necessary technical skills or training regarding analysis. Kemp and Klee determined that the majority of clinicians find language sample analyses to be an important and useful clinical tool. Stating that the "clinical practice must change in some way to accommodate the majority of clinicians who find language sample analysis useful but can't find the time to do it" (p. 169), Kemp and Klee suggested the possibilities of using transcription machines, analysis labs, and computers.

Language Sample Size

Initially, Lee and Canter (1971) recommended using 50 utterances when performing a DSS analysis. Johnson and Tomblin (1975) estimated the reliability of DSS on sample sizes ranging from five to 250 utterances and found that "estimated reliability values increased for all scoring categories as the sample size increased" (p. 377). Estimated reliability coefficients found for DSS composite scores were as follows: 0.60 reliability for 25 utterances, 0.75 for 50 utterances, 0.82 for 75 utterances, 0.86 for 100 utterances, 0.90 for 150 utterances, 0.91 for 175 utterances, 0.93 for 200 utterances, and 0.94 for 250 utterances. Johnson and Tomblin looked directly at Lee's recommended 50 utterance sample size and determined a value of 0.75 fell between the values for temporal and split-half reliabilities reported by Lee (1971). Most of the component items "fell below the value obtained for the total

score" (p. 377). A 175-sentence sample "must be collected before the standard error of measurement drops below 2.43 score points" (p. 378). Johnson and Tomblin noted, "these results indicate that a very large sample (three and one-half times the size recommended by Lee) is necessary before even a limited reduction in error can be achieved" (p. 378). Johnson and Tomblin pointed out that although Lee (1974) recommended DSS be used to isolate target areas for therapy, the "component scores that could perform this function have no norms, and so the method outlined cannot be used" (p. 378).

Gavin and Giles (1997) studied sample size effects on temporal reliability. Gavin and Giles argued that although the importance of language sample analyses has been established, reliability of each measure must be established if it is to be used for diagnostic or descriptive purposes. Gavin and Giles stated, "no studies have empirically evaluated whether the temporal reliability improves with increases in the number of utterances" (p. 1259). Because "high levels of reliability are necessary for the accurate identification of an impairment" (p. 1259), Gavin and Giles considered a correlation coefficient of .71 to be the minimum standard. Samples of duration (12 to 20 minutes) and total number of utterances (25 to 175 in increments of 25) were examined. Findings showed temporal reliability increased in all areas when sample size increased. The minimum correlation of .71 was not reached until sample size was at 100 utterances in all areas examined. A coefficient of .90 was not reached in any area until number of utterances reached 175. "For time-based samples, temporal reliability was higher in the 20-minute than the 12-minute sample" (p. 1261). Based on these findings, Gavin and Giles stated that language sample analyses for diagnosis or

client classification must be obtained from samples approaching 175 utterances. Gavin and Giles also stated that smaller sample sizes with lower reliability may be acceptable when tracking treatment progress. Because language sample analyses are frequently used to plan therapy goals, Gavin and Giles called for further research regarding validity and reliability of language sample size.

Muma (1998) addressed language sample size in a study that examined syntax in general using 400 utterances. Muma determined that "with a sampling error rate of about 15%, language samples should use 200 to 300 utterances to estimate most grammatical repertoires" (p. 310). The current use of 50 utterances has an error rate of 55% and using 100 utterance samples has a 40% error rate. Muma warned that making intervention goals from samples with such large error rates can be "inherently flawed" (p. 310). Although larger samples take more time, a lower error rate may prevent lost time in therapy due to inadequate goal selection.

Although there is no simple answer to the adequacy of sample size, as a general rule reliability increases as sample size increases. In the 1993 survey done by Hux et al., clinicians reported using between 51 and 100 utterance language samples. Kemp and Klee's (1997) survey reported clinicians using 50 utterances. Rondal (1978) used DSS to examine differences between language delay and language deviance in children with Down syndrome. Rondal used samples containing 75 utterances to better meet Johnson and Tomblin's (1975) recommendation. Blaxley et al. (1983) used 50 utterance samples, but questioned their results based on research calling for larger samples. Gregg and Andrews (1995) reported that a language analysis is only valid if the language sample it uses is representative of the client's performance. Completing

multiple linguistic analyses of a non-representative sample does not improve reliability and validity.

Developmental Sentence Scoring (DSS)

In 1971, Lee and Canter presented DSS as a "clinical procedure for estimating the status and progress of children enrolled for language training in a clinic" (p. 315). DSS was based on "a developmental scale of syntax acquisition" (p. 315) for children from 3;0 to 6;11. DSS "gives weighted scores to a developmental order of pronouns, verbs, negatives, conjunctions, yes-no questions, and wh-questions" (p. 315). "The mean score per sentence estimates the child's ability to formulate sentences with a high grammatical load" (p. 315). Lee and Canter intended DSS to be used to estimate a child's generalization of grammatical rules to guide therapy plans.

Lee and Canter (1971) discussed the use of DSS to estimate a child's progress from "one quarter to another" (p. 334) comparing a child's score "against his own previous scores" (p. 334). Lee and Canter stated, "DSS should not be considered... a test of syntactic or morphological development, but rather as a clinical procedure for analyzing verbal performance and planning appropriate remedial measures" (p. 335). Lee cautioned that the percentile chart included in this version of DSS was not to be used as normative data.

In 1972, Leonard presented a detailed description of deviant language and used DSS in a follow-up study to compare deviant and normal speakers. Leonard found that deviant speakers differed from their normal peers in terms of frequency of "certain deviant forms and transformations--not on the number of different deviant forms and transformations used and not on the developmental level of the structure used" (p.

443). Leonard determined that DSS may be “the most effective means of distinguishing between deviant language users requiring treatment and slow language developers who may not require clinical attention” (p. 441). Leonard reported that “a syntactic or morphological structure requiring clinical attention” (p. 438) should be a dependable member of the child’s linguistic repertoire. A reasonable estimate of this dependability is the reliability of the structure over time. A high temporal reliability correlation, then, should be required.

Later Lee (1974) published a book detailing a Reweighted DSS procedure which added normative data on children ages 2;0 to 2;11. This extended the use of DSS to children ages 2;0 to 6;11. A DSS analysis consists of a corpus of 50 complete (noun and verb in subject-predicate relationship), consecutive, and different sentences. The corpus should exhibit the child’s best performance from a spontaneous speech sample. No unintelligible utterances, echoed utterances, or repetitions of sentences are to be included. The speech sample should be obtained from the child’s conversation with an adult, using stimulus materials, pictures, and toys in which the child is interested. Lee recommended DSS as a “detailed, readily quantified and scored evaluation of a child’s use of standard English grammatical rules” (p. xix). After further statistical analysis, Lee claimed “DSS may be used with any clinical population to assess a child’s incorporation of adult grammatical rules into his spontaneous speech and to the degree that he is behind schedule, he will be called language-delayed” (p. xxi).

Johnson and Tomblin (1975) noted that norms were available for the DSS total score, but still no component score norms had been published. Johnson and Tomblin

reported the split-half reliability, Spearman-Brown correction formula for reduced sample size, and total score coefficient of stability.

Blaxley, Clinker, and Warr-Leeper (1983) used DSS as the basis for determining the accuracy of two screening tests in identifying language impairment. Although their conclusions depended entirely upon DSS results, Blaxley et al. noted that using DSS alone for diagnosis has limitations. Clinicians were reminded that Lee "also encouraged the use of other diagnostic measures in addition to the DSS to determine whether a child should be enrolled for therapy" (p. 45).

Lively (1984) identified common scoring errors associated with DSS in order to assist clinicians in learning the correct procedure. Quantifying expressive language, determining intervention goals, and evaluating treatment progress were discussed as benefits of a DSS analysis. Lively noted that although DSS is "one of the most popular and frequently used" (p. 154) syntactic analyses, it has its drawbacks. In order to achieve valid results, learning DSS requires "study and practice" (p. 154).

Hughes, Fey, and Long (1992) discussed three uses of DSS. First, DSS is a "numeric variable that can then be compared readily with the scores of other children or with previous and later scores from the same child" (p. 3). Second, DSS provides developmental data that can be used in diagnosis. Third, DSS provides a means for "asking and answering clinical questions" (p. 3). Hughes et al. reported DSS to be an "extremely useful" (p. 6) analysis when making diagnostic judgment, stating that DSS may be used when selecting goals and planning treatment. Hughes et al. stated, "a DSS analysis of the grammatical forms present, absent, infrequently used, or produced

in error can lead to hypotheses about the nature of the child's impairment and provide a basis for goal selection and therapy planning" (p. 6).

Clinicians have long used DSS to determine a child's progress over time. Hughes et al. (1992) reported the DSS advantages of sensitivity to grammatical change over time and a potentially high interrater reliability. However, longitudinal reliability of DSS using same-child data has not yet been established.

Hughes et al. (1992) noted DSS data available for comparing their clients' performance with other children. They pointed out that these data do not meet standard criteria and "clinicians must be cautious not to base diagnostic judgments entirely on these comparisons" (p. 2). However, Hughes et al. advocated the use of DSS in determining whether the gains their clients have made are large enough to be a result of treatment.

LARSP

LARSP is a form of grammatical analysis developed by Crystal, Fletcher, and Garman (1976). LARSP divides the stages of grammatical acquisition into seven levels. Crystal (1982) describes the LARSP procedure in *Profiling Linguistic Disability*. The clinician transcribes a 30-minute language sample and transfers word, phrase, and clause level information onto the profile chart under the categories of statements, questions, and commands. Stage I of the LARSP profile includes information characteristic of ages 0;9 to 1;6. Recorded at this stage are minor sentences including responses, vocatives, and interjections. Stage II is characteristic of ages 1;6 to 2 years. Utterances at this level usually contain two elements of the following: subject, verb, object, complement, and adverbial. Stage III represents

utterances at the age 2 to 2;6 level. Sentences at this stage usually include 3 elements. Stage IV "runs from about 2;6 to 3 years of age" and includes 4 or more elements (p. 30). Stage V includes utterances characteristic of age 3 to 3;6 and includes complex sentences. Stage VI represents the ages of 3;6 to 4;6 years and includes information not seen at other stages as well as errors of construction. Crystal states that stage VII has "little real assessment value" because it has been studied little in acquisition research. This stage includes information typically acquired after age four, including discourse, syntactic comprehension, grammatical style. After the profile has been completed, the clinician can determine which syntactic levels the child is producing and what information was not seen.

Automated Syntactic Analyses

Computers can significantly reduce the time required for analyzing language samples as well as improve accuracy. Long (1991) looked at how language analysis software as a whole could be used for assessment in the clinic. He found that computers could be used in transcription by efficiently recording and editing responses. The benefits of computers include processing large quantities of data, and scanning for subtle patterns within the data. Three features are shared by all computer programs. First, the clinician must code the utterances before the program can identify grammatical structure. Second, the computer can "recognize, analyze, and tabulate the information contained in the transcript" (p. 6). Third, the computer provides analysis results for interpretation. Long described the clinician's job as an interaction "with the program in order to code all the utterances accurately" (p. 6). Syntactic analyses such as DSS, involve a great deal of grammatical decisions for each utterance. Long stated

that “most clinicians would benefit from a program that is designed to detect and code automatically some of these structures” (p. 12). Using a computer in language analysis can relieve “the clinician from many tasks that are both time consuming and, because of their repetitiveness, mentally exhausting” (p. 2).

LARSP has been a popular method of syntactic analysis. Bishop (1984) introduced an automated version of LARSP, in which the clinician enters a typed sentence into the computer, the computer then breaks the sentence into words and assigns each to a grammatical category. The assignment of grammatical categories is done automatically, however; if the computer encounters an ambiguous or unfamiliar word, it is displayed on the screen to be manually assigned. The computer divides the words into phrases automatically with confirmation by the clinician. For example, the clinician “might be asked to specify whether a conjunction is phrasal or clausal, or whether a string of nouns constitute a single phrase” (p. 80). Phrases are then automatically scored by level, phrases are grouped together into clauses, relationships between clauses are determined, and the clause level is scored. Following this analysis, a summary in LARSP format is printed.

Long and Fey (1993) developed *Computerized Profiling* (CP). This program was developed to aid clinicians in language sample analysis. Each utterance of a language sample is manually entered into the computer. The CP program is designed to analyze the sample phonologically, semantically, and syntactically using the LARSP profile, the Profile of Phonology (PROP), and DSS. CP “generates a file of LARSP+ codes that describes the grammatical structure of sentences” (p. 7). The clinician must “edit incorrect codes generated by the program”, but the program

calculates the information and provides a complete LARSP profile (p. 7). In addition, a DSS analysis may be extracted from a LARSP profile whether or not the profile has been corrected.

Baker-Van Den Goorbergh (1994) outlined the benefits and drawbacks of some of the current software programs available for language sample analysis. Five automated language analysis programs were discussed. The *Systematic Analysis of Language Transcripts* (SALT; Miller & Chapman, 1990) program requires the clinician to code morphemes, verb tenses, missing lexical items, pauses, and unintelligible utterances. The analysis is therefore subject to clinical error, but also provides a detailed data summary. The *Parrot Easy Language Sample Analysis* (Weiner, 1986) uses percentages to make across-sample comparisons, but the calculations were incorrect in nine of ten samples tried. The *Pye Analysis of Language* (PAL; Pye, 1987) lacked a detailed user manual and not all the data were included in the analysis. Benefits of using PAL included ease of use and multiple analyses such as phonological, lexical, and syntactic. *Computerized Profiling* (CP; Long & Fey, 1993) also performs multiple analyses using five different modules. These modules are based on DSS, LARSP, PROP, PRISM-L and PROHP methods. Baker-Van Den Goorbergh reported incorrect analysis using the LARSP module that created time-consuming corrections. The *Computerized Language Error Analysis Report* (CLEAR; Baker-Van Den Goorbergh and Baker, 1991) presented accessible information through graphs, tables and offered comparisons of data over time or across patients. CLEAR requires users to be familiar with the LARSP system and uses two different editing modes, possibly causing confusion. Baker-Van Den Goorbergh recognized the valuable role

of computers for language sample analysis stating, “manipulating and categorizing data is a tedious, time-consuming task for the therapist, but a matter of seconds for a computer” (p. 331).

Long and Fey (1995) argued that the programs in Baker-Van Den Goorbergh’s (1994) article were inaccurately described. Long and Fey pointed out that all language analysis programs require coding, offer the ability to create new analysis categories, and CP and SALT both offer all the advantages Baker-Van Den Goorbergh listed for CLEAR. Two modules contained in CP, APRON and Early Vocabularies, were overlooked by Baker-Van Den Goorbergh. Long and Fey discussed the different purposes and development of the modules as well as clarified other aspects of each analysis program. Long and Fey agreed that using computer programs for language sample analysis can “increase both the efficiency and quality” (p. 185) of clinical work; however, clinicians often select the appropriate program based on professional reviews. Long and Fey emphasized the necessity for these reviews to be fair and accurate. Because of the complexity of computer programs, Long and Fey encouraged users to research the programs that are designed for their purposes rather than assume these programs can be adequately described in a short journal article.

In an unpublished master’s thesis, Boyce (1995) studied the accuracy of Computerized Language Analysis (CLAN; MacWhinney, 1991) and CP software in performing DSS. She stated that due to the increased validity of language sample analysis over other highly structured methods, researchers have turned to computers for analysis assistance. Boyce compared these programs to manual analysis and found a wide range in accuracy for various DSS categories (0% to 94% agreement). In four

of the eight overall DSS categories, CP outperformed CLAN; however, CLAN was better at scoring interrogative reversals. For the DSS developmental levels, CP significantly outperformed CLAN in 14 category/level combinations, CLAN did better in eight, and no significant difference was found for the remaining 14. Boyce concluded that although computer software is a valuable tool in some areas of assessment, "further development... is necessary before clinicians can rely on automated DSS analysis" (p. 2).

Despite the reported benefits of using computers in language sample analysis, surveys by Hux et al. (1993) and Kemp and Klee (1997) showed computers were not being used by most clinicians. Only 8% of respondents in the Kemp and Klee survey and 3% of respondents surveyed by Hux et al. reported using computers for analysis purposes.

Automated Grammatical Tagging

Syntactic analysis can be made even faster by automating the grammatical tagging procedure. Grammatical tagging involves assigning each word to a grammatical category. A computer program uses these grammatical codes (tags) to determine the syntactic accuracy of the utterance. The accuracy of automated tagging is affected by grammatical ambiguity. For example, the word *can* may be used as a noun in the sentence *He opened the can* and as an auxiliary verb in the sentence *He can run*. Ambiguity of grammatical categorization by computers has been addressed using algorithms such as CLAWS (Garside, Leech, & Sampson, 1987).

Probabilistic grammars have been developed to account for grammatical category ambiguity. These programs have resulted in high levels of accuracy.

Probabilistic methods are based on the assumption that there are discoverable regularities about language in terms of grammatical category frequency and order. These methods are used to select the most likely grammatical tag option based on two probabilities. The first probability is the likelihood that a particular grammatical tag describes a particular use of a word. The second probability is the likelihood that a particular tag follows another specific tag.

DeRose (1988) developed a program for grammatical tag disambiguation. This program used probabilistic methods to determine the most likely syntactic interpretation of a sentence. The program was tested on the million words of the Brown University Standard Corpus of English and was shown to be 96% accurate.

Charniak (1993) discussed the use of probabilistic grammars in grammatical tagging programs. Because tagging programs deal with 40 or more parts of speech, many sentences can be read using different interpretations. This ambiguity is due to the fact that many words can be used in several grammatical categories. However for a given word, there will be a more likely grammatical category than others. The probability that a word will be used in a particular category because of grammatical rules can be estimated. Probabilistic grammars use statistical methods to determine the most likely syntactic interpretation of a sentence.

Channell and Wilmarth (1998) applied probabilistic methods of automated tagging procedures to samples from children with language disorders. The authors used the same probabilistic algorithm used by Garside et al. (1987), DeRose (1988), and Church, (1988). Thirty language samples were manually and computer tagged followed by a comparison on a tag-by-tag basis. Ten samples obtained from children

with language impairment generated a mean of 93.5% (*SD* = 1.9). Ten samples from children who were language age matched showed a mean of 94.5% (*SD* = 1.8). Ten samples obtained from the chronological age group yielded a mean of 94.0% (*SD* = 0.8). Channell and Wilmarth stated that with further improvement in the accuracy of automated grammatical tagging, grammatical analysis will consequently become more accurate.

Channell & Johnson (1999) applied probabilistic methods of automated grammatical tagging to child language samples. They used 5,000 utterances from conversational language samples of 30 typically developing children to perform a word-level analysis comparing manual and computer tagging. The results of their comparison showed automated grammatical tagging to be from 95.1% accurate at the word level and between 60.5% and 90.3% accurate at the utterance level.

Method

In this study, six collections of language samples were divided into two subsets. One set was designated as the training corpus, the other set as the test corpus. The training corpus included approximately 20,000 utterances from four collections of language samples. The training corpus served as the source of probability data used in tagging both samples in the test corpus. The test corpus consisted of approximately 12,700 utterances from two collections of language samples obtained from typically developing children and children with language impairment. Only the child utterances from these samples were grammatically tagged.

Samples Used Only For Training

Abe. Kucjaz (1973) collected and transcribed approximately 1,900 utterances from his son Abe. The sample was created as a diary study from 1973 to 1975. Abe was recorded twice a week in one half-hour sessions from age 2;4 to 4;1. Additional utterances were obtained once a week in one half-hour session from age 4;1 to 5;0. Abe was considered to be a typically developing child. The Abe corpus is available from the CHILDES archives (MacWhinney, 1991).

Sarah. Roger Brown (1973) and his students collected approximately 6,073 utterances from Sarah, a typically developing child, between the ages of age 2;3 to 5;1. The samples were recorded as Sarah interacted with family members and a graduate student informally in her home. This corpus is also available as part of the CHILDES archive (MacWhinney, 1991).

Wymount. Channell and Johnson (1999) used 30 samples previously collected from typically developing children. These children were between the ages of 2;6 and 7;11. The samples were recorded as each child interacted with a graduate student informally in the child's own home. Altogether, the corpus consists of approximately 6,000 child utterances each.

Garvey. Approximately 3,000 utterances were used from samples collected by Garvey (1973). These utterances were obtained from typically developing children between the ages of 2;10 and 5;7. The samples were created from dialogues between two members of a triad. Each child in a triad was represented as A, B, or C. The children were then paired for transcription in sets of AB, AC, and BC. These samples are from the CHILDES archives (MacWhinney, 1991).

Samples Used For Training and Testing

Reno. Thirty samples collected by Fujiki, Brinton, and Sonnenberg (1990) from children in the Reno, Nevada area were used in the test corpus. Approximately 8,700 utterances were obtained from these samples, including 10 children with language impairment (LI), 10 language similar (LS) matches, and 10 chronological age (CA) matches. The children with LI averaged 9;1 years in age, showing deficits in expressive and receptive language. The children with LI scored at least one standard deviation below the mean on two formal tests. Each child with LI was matched to a child of language age within six months according to performance on the Utah Test of Language Development (Mecham, Jex, & Jones, 1967). The LS matches averaged 6;9 years of age. The children in the CA group were paired to a child with LI from their same elementary school. The CA children averaged 9;0 years of age and were achieving academically at grade level. The samples from these children were collected as part of a study on conversation repair. Therefore, the samples reflect a disproportionate number of requests for repair made by the examiner.

Jordan. Collingridge (1998) collected 20 samples of children with language impairment from the Jordan School District in Salt Lake City, Utah. This set of samples consisted of approximately 3,700 utterances. At the time of collection, these children were receiving services in the school for communication or learning disorders. The children were judged by speech-language pathologists as having language impairment but also with adequate language skills to actively participate in conversation. The children were required to be at least 80% intelligible. The children

were either receiving pull-out services or attending a classroom for children with communication or learning disabilities.

Tagging the Corpora

Utterances were grammatically tagged using a tagging scheme adapted from the LARSP approach of Crystal et al. (1989). Each word was assigned a grammatical tag from the list in appendix A. Samples were either tagged by hand or were auto-tagged by GramCats. If the samples had been tagged by GramCats, they were manually converted to the current tag scheme and manually corrected.

Interrater reliability was established on the test corpus samples by examining or correcting 10% of the data and then comparing agreement on a tag-by-tag basis. The Reno samples were previously corrected by Wilmarth (1997) and reliability of this tagging was found to be at or above 95%. The reliability of tagging the Jordan samples was found to average 96%.

Tagging Software

Each word in each utterance of the test samples was tagged by the software. For known words, the tagging software assigned the appropriate grammatical tag. If there was more than one tag option for the word, the software determined the best tag option based on (a) the likelihood of that tag for that particular word, and (b) the likelihood of that tag given the one or two tags in front of it, depending on whether bi-gram or tri-gram probabilities were being used.

The assignment of tags to unknown words was handled in the following manner. Unknown words, or words missing from the tag dictionary, were assigned the possible categories of "<N 9 <V 3 <AJ 1", unless the word ended in *s*, in which case

the tag options applied were "<N.s 3 <V.z 1". This means that an unknown word was guessed to be most likely a (singular) noun, less likely a verb, and minimally likely to be an adjective. Words ending in *s* were guessed to be most likely a plural noun, or less likely to be a 3rd person singular verb form. The program then combined this guessed tag option data with the other form of probability data used, the likelihood of a grammatical tag given the tags of the one or two words in front of the unknown word, to select the most likely tag for the unknown word.

Procedure

The training corpus and the test corpus were created after all samples had been manually tagged or corrected. The Reno corpus was included with the training corpus when testing the Jordan samples. Conversely, the Jordan corpus was combined with the training corpus when testing the Reno samples.

A utility program was used to extract bi-gram and tri-gram probability information as well as word and tag options with their frequencies from each training corpus. The program then used these data to create new dictionaries to examine differences in training corpus size. These dictionaries thus consisted of data from randomly selected sets of utterances of the following sizes: 5,000, 10,000, and 15,000. When the large Reno corpus was combined with the training set, one additional size (20,000 randomly selected utterances) was possible for tagging the Jordan corpus.

To determine which probability sequence model was more accurate, the test corpora were tagged twice, once using a bi-gram probability sequence and once using tri-gram. The reliance on a guessing rule was determined by examining each dictionary for the number of missing words and tag options relative to the manually

corrected test corpus. A utility program was used to compare manual and auto-tagged utterances of each file on a word-by-word and utterance-by-utterance basis.

Results

Bi-gram and Tri-gram Probability Models

Both per-tag and whole-utterance accuracy were higher when using tri-gram rather than bi-gram probabilities. Table 1 presents accuracy data obtained using dictionaries derived from the full training corpora. The accuracy using tri-gram probabilities was higher for every sample than accuracy using bi-gram probabilities. Paired t-tests, comparing the accuracy of tagging each sample with tri-gram versus bi-gram probability data, were statistically significant for both per-tag t (49) = 14.53, $p < .0001$ and whole-utterance accuracy t (49) = 17.52; $p < .0001$.

Table 1

Tag and Utterance Accuracy Percentages Using Bi-gram and Tri-gram Probabilities

	Tag Accuracy				Utterance Accuracy			
	JOR	R-LI	R-LS	R-CA	JOR	R-LI	R-LS	R-CA
Bi-gram	93.4	92.0	93.2	91.7	73.7	64.6	68.3	56.4
Tri-gram	95.0	93.1	94.6	93.4	79.6	70.6	74.0	63.8

Size of Training Corpus

Larger training corpora resulted in better accuracy. As can be seen in Tables 2 and 3, the accuracy of tagging either corpus using a dictionary based on the full training set was more accurate than tagging which used randomly-drawn subsets of the training utterances. Because tri-gram probabilities always yielded higher accuracy than did bi-gram probabilities, the comparisons below were made using only tri-gram probabilities.

Table 2

Training Corpus Size and Jordan Corpus Tagging Accuracy

Size	Tag %	Utterance %
5,000	93.9	75.8
10,000	94.6	77.9
15,000	94.9	78.7
20,254	95.0	79.6

Table 3

Training Corpus Size and Reno Corpus Tagging Accuracy

	Tag %				Utterance %			
Size	R-LI	R-LS	R-CA	All	R-LI	R-LS	R-CA	All
5,000	92.3	93.5	92.4	92.7	67.6	71.4	60.6	66.4
10,000	92.7	94.2	93.0	93.3	69.2	73.0	62.0	67.9
14,601	93.1	94.6	93.4	93.7	70.6	74.0	63.8	69.3

It can be seen in these tables that higher accuracy was obtained using dictionaries based on larger numbers of utterances from the training corpora, but the amount of this improvement diminished or plateaued for both test corpora.

Amount of Guessing

A total of 3385 word tokens (5.6% of the total) were in the Reno sample but were not in its full training corpus set and thus not in the tagging program's dictionary. Another 413 words (0.7%) were in the dictionary but the correct tag options for those words were not. About 68.3% of the missing words or tag options were nouns, 21.7% were verbs, 5.8% adjectives, 2.0% adverbs, and 2.3% were from other categories.

In contrast, 579 words (6.9% of the total) were in the Jordan sample but not in its full training corpus and program dictionary. Another 92 words (1.1%) were in the dictionary but the correct tag option was not. About 79.1% of the missing words or tag options were nouns, 13.1% were verbs, 4.7% adjectives, 1.8% adverbs, and 1.3% were from other categories.

Group and Individual Differences

Groups differed in terms of the accuracy of tagging. The mean accuracy levels obtained for all groups are presented in Table 4. In the Reno corpus, accuracy for samples from the 10 children with language impairment was lower than that for samples from the language similar or chronological age matched groups on the per-tag basis but was between the other groups on a whole-utterance basis. The accuracy of tagging samples in the Jordan corpus (all 20 children having language impairment) was higher than the accuracy obtained for any group in the Reno corpus. One-way

ANOVAs revealed significant differences between groups for both per-tag accuracy $F(3,46) = 8.5$, $p = .0001$ and whole-utterance accuracy $F\ (3,46) = 21.6$, $p < .0001$.

Table 4

Tag and Utterance Accuracy Percentages By Groups

	Group			
	JOR	R-LI	R-LS	R-CA
Tag Mean	95.2	92.9	94.7	93.3
Tag SD	1.4	1.8	0.9	1.2
Utterance Mean	79.9	70.5	74.0	63.7
Utterance SD	5.7	4.6	5.0	5.8

Accuracy for tagging individual samples in the Reno corpus ranged from 88.8% to 96.3% on the per-tag basis and from 53.6% to 84.4% for whole utterances. Accuracy for tagging individual samples in the Jordan corpus ranged from 92.9% to 97.8% per-tag and from 71.5% to 92.2% for whole utterances. Samples with a lower ratio of the number of tags to the number of utterances generally had higher whole utterance accuracy ($r = -.71$, $p < .0001$) but not necessarily higher per-tag accuracy ($r = -.21$, $p = .14$).

Accuracy by Grammatical Category

Some grammatical categories were tagged with higher accuracy than others. Table 5 shows the accuracy of each tag in both samples when tagged using tri-gram probabilities and the largest training corpus size to form the dictionary.

Table 5

Accuracy by Grammatical Category

Tag	Description	Reno		Jordan	
		N	%	N	%
<$	possessive 's	137	79	33	70
<+	phrasal coord.	862	78	102	66
<AJ	adjective	2213	83	207	88
<AJ.r	comparative adjective	20	80	8	75
<AJ.t	superlative adjective	13	62	1	0
<AM	modal aux	478	99	108	100
<AM.ng	modal aux + negation	85	100	33	100
<AO	other aux	574	89	147	84
<AO.d	past aux	163	56	29	79
<AO.dz	past 3rd person aux	143	86	15	67
<AO.g	progressive aux	4	75	2	100
<AO.n	perfective aux	4	75	0	
<AO.z	3rd person aux	1000	82	109	94
<AV	adverb	3157	92	608	94
<AV.r	comparative adverb	5	80	4	100
<CC	clausal coord.	2761	93	261	98
<CP	copula	314	91	101	93
<CP.d	past copula	16	100	4	50
<CP.dz	past 3rd person copula	226	98	15	67
<CP.g	progressive copula	0		1	100
<CP.n	perfective copula	8	13	1	100

(*table continues*)

Tag	Description	Reno		Jordan	
		N	%	N	%
<CP.z	3rd person copula	1647	99	281	98
<D	determiner	7354	99	964	98
<IF	intensifier	459	72	122	83
<LT	let's	51	100	2	100
<N	noun singular	8098	98	1097	97
<N.s	noun plural	1821	97	246	95
<NG	negation	464	100	80	100
<PO	pronoun other	2244	92	382	91
<PP	pronoun	8323	100	1162	100
<PR	preposition	4273	97	461	97
<PT	particle	1331	84	125	89
<Q	wh-question	152	76	74	86
<SB	subordinator	1366	87	187	91
<TH	existential	356	26	39	92
<TO	infinitive to	682	98	96	96
<V	verb	4384	96	734	98
<V.d	past verb	1562	86	162	86
<V.g	progressive verb	1371	82	157	97
<V.n	perfective verb	429	57	49	55
<V.z	3rd person verb	1448	96	142	99
<VTO	verb + to	196	100	68	100

Frequently occurring categories with particularly low accuracy were possessive *'s*, phrasal coordination, past tense aux, intensifier, wh-question, existential *there*, and the perfective forms of verbs.

Discussion

Automated grammatical tagging works as well on samples from children with language impairment as it does on samples from typically-developing children. The greater amount of grammatical context information available when using a tri-gram probability model significantly improves accuracy over that of a bi-gram model. Larger training corpora lead to more accurate tagging, probably due to reducing the program's reliance on guessing rules.

The findings of the present study using the gc3 software are similar to those obtained by Channell and Johnson (1999) using the gramCats software, though the present data were obtained under more severe conditions. Channell and Johnson found gramCats to yield 95% accuracy at the word level and 78% accuracy at utterance level. In the present study, overall accuracy levels using the gc3 software were about 95% at the word level and 80% at the utterance level. However, the gc3 software guessed words not present in the training corpus instead of having unknown words manually added to the program's dictionary. Likewise, the children sampled in the present study were older than those of the Channell and Johnson study, which had obtained accuracy levels about 2% lower when tagging samples from older children.

Further work on automated grammatical tagging software and its application to language samples is warranted. Once a sample has been entered into the computer, this software is very fast, tagging about 100 utterances per second. Two clinicians tagging

the same transcript using the software will get the same results, unlike manual analysis. The more accurate the automated analysis, the less subsequent correction needed.

The present study found that tagging accuracy increased as the training corpus size was increased, but the amount of improvement diminished. Whether further increases in training corpus size or variety would keep improving accuracy is unknown, but it appears that using more than 20,000 utterances would be not cost effective. An alternate approach to this issue is used in the most recent version of Computerized Profiling (Long, Fey, & Channell, 2000) wherein several thousand words never encountered in a training corpus have been added along with the possible tag options for those words to the program's dictionary, albeit with flat (i.e., even) information relative to tag likelihood. The relative accuracy of this large but flat dictionary strategy has not received empirical evaluation.

No matter how large the dictionary, some procedure for handling unknown words will still be required. This topic has yet to receive research scrutiny, and better techniques for guessing the grammatical categories of unknown words than the one used by the gc3 software in the present study might well be possible.

Clinical Applications

The findings of this study suggest that clinicians might make immediate clinical use of the gc3 software. Using the software, clinicians are able to quickly and accurately assign a grammatical code to every word in a language sample. Where manually tagging a language sample once took hours, tagging of a formatted sample can be done in seconds. Once a language sample has been tagged, other clinical

analyses can be carried out on the sample to assist in the description and quantification of syntactic development. For example, the "Find" function of a typical word processing program could be used to examine a client's use of modal auxiliaries, personal pronouns, or subordinating conjunctions by searching for occurrences of the tags <AM, <PP, or <SB.

The gc3 software makes the tagging process much faster but may also offer greater accuracy. The accuracy of manual tagging is dependent on the syntactic proficiency, grammatical analysis training, and attentiveness of the clinician. However, auto-tagging achieves its high accuracy levels consistently, and saves time even when post-corrected by clinicians whose accuracy is higher than that of the software.

The gc3 software can be applied to language samples in order to facilitate the identification of specific structures thought to characterize children with specific language impairment (SLI). Recent studies have isolated grammatical markers that can be used to identify SLI in children (Rice & Wexler, 1996; Rice, Wexler, & Hershberger, 1998). For example, children with SLI use grammatical tense at a much lower level than their age peers. Using syntactic analyses from naturalistic language samples to identify impairment offers greater validity than methods that focus solely on low test scores.

Overall, the present study provides data showing that high accuracy levels are possible when grammatically tagging samples from children with language impairment, when using a clinically familiar, LARSP-based tagging scheme, and when guessing words not encountered in a training corpus. We may thus expect

automated language sample analysis to play an ever more useful role in language research and intervention.

References

Baker-Van Den Goorbergh, L. (1994). Computers and language analysis: Theory and practice. *Child Language Teaching and Therapy, 10,* 329-348.

Baker-Van Den Goorbergh, L., & Baker, K. (1991). *Computerised language error analysis report (CLEAR).* Kibworth, Leics: FAR Communications.

Bishop, D. V. M. (1984). Automated LARSP: Computer-assisted grammatical analysis. *British Journal of Communication Disorders, 19,* 78-87.

Blaxley, L., Clinker, M., & Warr-Leeper, G. (1983). Two language screening tests compared with Developmental Sentence Scoring. *Language, Speech, and Hearing Services in Schools, 14,* 38-46.

Boyce, L. L. (1995). *Accuracy of automated Developmental Sentence Scoring.* Unpublished master's thesis, Brigham Young University, Provo, UT.

Channell, R. W. (2000). gc3 [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.

Channell, R. W., & Wilmarth, J. W. (1998, November). Automated analysis of language from children with language impairment. Poster presented at the national meeting of the American Speech-Language Hearing Association, San Antonio.

Charniak, E. (1993). *Statistical language learning.* Cambridge, MA: MIT.

Church, K. W. (1988). A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing,* pp. 136-143. Somerset, NJ: Association for Computational Linguistics.

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

Crystal, D. (1982). *Profiling linguistic disability*. London: Edward Arnold.

Crystal, D., Garman, M., & Fletcher, P. (1976). *The grammatical analysis of language disability.* London: Edward Arnold.

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

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

Fujiki, M., Brinton, B., & Sonnenberg, E. A. (1990). Repair of overlapping speech in the conversations of specifically language-impaired and normally developing children. *Applied Psycholinguistics, 11,* 201-215.

Garside, R., Leech, G., & Sampson, G. (Eds.). (1987). *The computational analysis of English: A corpus based approach.* London: Longman.

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

Gavin, W. J., & Giles, L. (1997). Sample size effects on temporal reliability of language sample measures of preschool children. *Journal of Speech and Hearing Research 39,* 1258-1262.

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

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

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

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

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

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

Lahey, M. (1988). *Language disorders and language development.* Needham, MA: Macmillan.

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

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

Leonard, L. B. (1972). What is deviant language? *Journal of Speech and Hearing Disorders, 37*(4), 427-446.

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

Long, S. H. (1991). Integrating microcomputer applications into speech and language assessment. *Topics in Language Disorders, 11*(2), 1-17.

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

Long, S. H., & Fey, M. E. (1993). *Computerized Profiling.* The Psychological Corporation.

Long, S. H., & Fey, M. E. (1995). Clearing the air: A comment on Baker-Van Den Goorbergh (1994). *Child Language Teaching and Therapy, 11,* 185-192.

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

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

Mecham, M., Jex, J. L., & Jones, J. D. (1967). *Utah Test of Language Development*. Salt Lake City, UT: Communication Research Associates.

Miller, J. F., & Chapman, R. S. (1990). *Systematic analysis of language transcripts (SALT) Version 1.3 (MS-DOS).* [Computer program]. Madison, WI: Language Analysis Laboratory. Waisman Center on Mental Retardation and Human Development.

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

Pye, C. (1987). *Pye Analysis of Language (PAL).* [Computer Program]. 200 Arrowhead, Lawrence, KN.

Rice, M. L., & Wexler, K. (1996). Toward tense as a clinical marker of specific language impairment in English-speaking children. *Journal of Speech and Hearing Research, 39,* 1239-1257.

Rice, M. L., Wexler, K., & Hershberger, S. (1998). Tense over time: The longitudinal course of tense acquisition in children with specific language impairment. *Journal of Speech and Hearing Research, 41,* 1412-1431.

Rondal, J. A. (1978). Developmental sentence scoring procedure and the delay-difference question in language development of Downs Syndrome children. *Mental Retardation, 16,* 169-171.

Weiner, F. F. (1986). *Parrot Easy Language Sample Analysis (PELSA)* and *Parrot Language Sample Utility (PLSU).* [Computer Program]. Parrot Software, 190 Ridge Road, State College, PA.

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

Appendix

Tag Category Scheme (13 June 2000)

This tagging scheme pretty much uses LARSP tags, but it has been adapted so that every word gets a tag. The tags are thus mainly phrase-level, with a few clause-level ones added when no phrase-level tag would be assigned. In addition, subcodes are added for LARSP word-level coding. Only the classes <N, <V, <AJ, and <AV are open.

LARSP phrase-level tags

<D determiner (includes quantifiers & numbers)
<AJ adjectival (<AJ.r comparative, <AJ.t superlative)
<N noun, proper or common; <N.s plural
(nouns functioning adjectivally are coded <N)
<$ *'s* marking possession
<PP personal pronoun (I me you she her he him it we us they them)
<PO other pronoun
<V verb
<V.z 3ps, <V.d past, <V.g -ing, <V.n -en
<CP copula *be* (use same verb subcodes but <CP.dz *was*)
<NG negation (*not* or *n't*)
<AM modal (AM.ng for *can't, won't*)
<AO other auxiliary (*be have do get*)
(use verb subcodes but <AO.dz *was*)
<TO infinitive marker *to*
<VTO catenatives *gonna*
<PT verb particle (the 10 words *about, across, down, in, off, on, out, over, through, up* except when used as prepositions)
<AV adverb (<AV.r comparative, <AV.t superlative)
<IF intensifier/qualifier
<+ phrasal conjunction
<PR preposition

Tags originally on the LARSP clause-level

<CC clausal conjunction
<LT *let's*
<Q question word (pronoun or adverb)
<SB subordinator
<TH existential *there*