ABSTRACT

Title of Thesis:

FACTORS THAT INFLUENCE EFFICACY OF AUTOMATED LANGUAGE SAMPLE ANALYSIS FOR CHILDREN WITH DEVELOPMENTAL LANGUAGE DISORDER

Mackenzie Leigh Jones, Master of Arts, 2026

Thesis Directed By:

Professor Yi Ting Huang, Department of Hearing & Speech Science

Automated language sample analysis (LSA) offers clinicians the ability to reduce time and labor loads often associated with the manual task process. Children with DLD are one of the main populations LSA benefits, due to individualized language deficit profiles of this population. To provide clinicians with straightforward standards to increase the accuracy and clinical utility of automated LSA for children with DLD, we investigated what factors at the language sample retrieval, transcription, and analysis stages influence accuracy. Our study specifically explored what acoustic-related factors during sample retrieval may influence transcription, and what metrics during analysis may influence result interpretations. Automated LSA for children with DLD (n=44) performed with greater accuracy once children had received language treatment for 10 weeks. Speech intelligibility ratings of children influenced automated transcription consistently, while background noise and audio recording equipment quality were not found to significantly influence transcription accuracy. At the metric analysis stage, automated transcription metric analysis did not track morphological and lexical gains to the same significance as manual transcription analysis did. Automated metric analysis with metrics that can only feasibly be computed by voice activity detection did find statistically significant gains between pre- and post- treatment transcriptions. If opting to use automated LSA, clinicians should consider speech intelligibility of children and metrics used in analysis, as these factors seem to impact accuracy of LSA as a tool for designing and assessing treatment for children with DLD.

FACTORS THAT INFLUENCE EFFICACY OF AUTOMATED LANGUAGE SAMPLE ANALYSIS FOR CHILDREN WITH DEVELOPMETAL LANGUAGE DISORDER

by

Mackenzie Leigh Jones

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Master of Arts
2026

Advisory Committee:
Professor Yi Ting Huang, Chair
Dr. Amanda Owen Van Horne
Assistant Professor José Ortiz

Table of Contents

Table of Contents……...……..ii
List of Figures…….……iii
List of Tables……...…iv
Chapter 1: Introduction..... 1
Chapter 2: Methods 9
Section 1: Participants……..……9
Section 2: Materials and instruments……..……...10
Section 3: Automated metrics……11
Section 4: Manual Coding…….13
Subsection 1: Dimension #1 (speech intelligibility)……..……13
Subsection 2: Dimension #2 (background noise)……..……14
Subsection 3: Dimension #3 (audio quality)……..….15
Chapter 3: Results..... 17
Section 1: Research Question #1……..…….……17
Section 2: Research Question #2……..……..…19
Section 3: Research Question #3……...……22
Section 4: Research Question #4……..……25
Chapter 4: Discussion..... 29
Chapter 5: Limitations..... 33
Chapter 6: Conclusion 36
Appendices……..…….....38
Bibliography……..39

List of Figures

Figure i. Word error rate between pretreatment transcriptions. The histogram shows frequency of WER values. Dashed line indicates median split. 17
Figure ii. Word error rate between posttreatment transcriptions. The histogram shows frequency of WER values. Dashed line indicates median split 18
Figure iii. Correlation between speech intelligibility ratings and WER values at the pretreatment stage. The legend key indicates shape associated with age of each plotted participant. Red pentagon for age group 4;0-4;11, blue circle for age group 5;0-5;11, yellow triangle for age group 6;0-6;11, green square for age group 7;0-7;11, pink star for age group 8;0-8;11, and purple rhombus for age group 9;0-9;11. 21
Figure iv. Correlation between speech intelligibility ratings and WER values at the posttreatment stage. The legend key indicates shape associated with age of each plotted participant. Red pentagon for age group 4;0-4;11, blue circle for age group 5;0-5;11, yellow triangle for age group 6;0-6;11, green square for age group 7;0-7;11, pink star for age group 8;0-8;11, and purple rhombus for age group 9;0-9;11. 21
Figure v. Correlations between manual transcription LSA metric measurements and ASR transcription metric measurements at the pretreatment stage. Metrics included are: mean length of utterance (MLU), number of different words (NDW), and total number of words (TNW), from left to right. 22
Figure vi. Correlations between manual transcription LSA metric measurements and ASR transcription metric measurements at the posttreatment stage. Metrics included are: mean length of utterance (MLU), number of different words (NDW), and total number of words (TNW), from left to right. 23
Figure vii. Correlation(s) between participant DELV-NR standard score and mean length of utterance in morphemes (MLUm) metric measurement from both manual (blue line and dots) and ASR (red line and dots) pretreatment transcription. 24
Figure viii. Paired line plots demonstrating change in manual transcription metric measurements for MLUm, NDW, and TNW from pre- to post-treatment. Bolded line indicates mean change of group. 25
Figure ix. . Paired line plots demonstrating change in ASR transcription metric measurements for MLUm and NDW from pre- to post-treatment. Bolded line indicates mean change of group 26
Figure x. . Paired line plots demonstrating change in ASR transcription metric measurements for mean length of speech segment (MLSS) from pre- to post-treatment. Bolded line indicates mean change of group. 27

List of Tables

Table i. Multivariable regression analysis where speech intelligibility, background noise, and audio quality are predictors of WER outcome. 19

CHAPTER 1: INTRODUCTION

Language sample analysis (LSA) is a non-standardized assessment widely used for evaluating language abilities, especially in children with developmental language disorder (DLD). The LSA procedure involves retrieving a speech recording, transcribing the speech, analyzing the language through selected metrics, and then inferring language patterns from analysis. It complements standardized assessments by offering deeper insights into language deficits across environments and providing clinicians with flexible options for analysis. Despite its value, LSA is often viewed as time-consuming and resource-intensive, which has resulted in reduced clinical use (Pavelko et al., 2016). One promising solution is automated speech recognition (ASR), which has been explored for streamlining LSA transcription and analysis (Agmon et al., 2023; Attia et al., 2024; Fox et al., 2022; Liu et al., 2023). ASR-assisted protocols have the potential to address many clinicians' concerns by offering the ability to speed up the transcription and analysis portions, without sacrificing decision-making accuracy. It also offers novel metrics that humans often do not have time or ability to measure, expanding the analysis portion of LSA. While clinicians are eager to find a user-friendly ASR system to reduce labor involved in LSA, research has not yet explored the efficacy of many systems designed for pediatric DLD populations.

Ongoing research aims to optimize ASR for efficient and reliable use in this population. However, most ASR systems have been trained on adult populations and therefore there is limited research for ASR use in pediatric populations. The quality of children's language samples differs from adults due to challenges specific to the pediatric population (Lammert et al., 2025; Lüdtke et al., 2023). Children tend to have poorer speech intelligibility and their developing executive functioning skills can result in less structured, more variable language use

and/or language interference in their speech samples (e.g., getting distracted with the task at hand, side comments from adults redirecting children to task at hand). Within the pediatric population, children with DLD experience even greater variability in language use. LSA is an invaluable tool at identifying children with DLD from children with TL (Bååth et al., 2019; Wilder & Redmond, 2022; Winters et al., 2022). While children require longer language sample analysis compared to adults (Lüdtke et al., 2023), research has found that children with DLD require even longer language samples due to increased language variability (Wilder & Redmond, 2022). The expectation of increased language sample length for the pediatric DLD population adds onto this time-intensive issue. These factors complicate both manual and automated, transcription and analysis. The lack of research on how to use ASR effectively in pediatric populations highlights a critical gap in both technology development and clinical application. This gap underscores the need for focused investigation into ASR's performance in pediatric LSA contexts, especially for children with DLD.

The goal of our study is to identify ASR transcription performance for children with DLD as well as what potential confounding factors found in audio samples may interfere with transcription performance. By the end of our study, we hope our identifications can enhance clinical use of ASR by increasing efficiency, accuracy, and ease in understanding how to use ASR software. The proposed research will examine how accurately ASR can capture the linguistic variability seen in children with DLD and what models/evaluation metrics are best suited to ensure ASR efficacy of this population. With the growing component of telehealth in clinical settings, challenges are compounded by additional sources of environmental noise such as background sounds from other children, eating, movement, and device interference. Such acoustic factors degrade audio quality and further complicate both human and automated

analysis performance (Meera et al., 2025), and likely transcription performance. These challenges can be particularly pronounced in children with (DLD), which ASR has not yet been trained on due to the need for thousands of hours of data training for greater reliability of ASR in this population (Lüdtke et al., 2023). Children with DLD are also prone to experiencing morphosyntactic errors, as that is a hallmark of their diagnosis and an essential feature in their treatment planning through LSA. However, while humans may struggle with overcorrecting speech and language in LSA transcripts, ASR is known to perform similar overcorrections (Attia et al., 2023). After you obtain a language sample, LSA needs to be transcribed exactly how someone's speech is (e.g., including any misarticulations, incorrect grammar) to achieve the most authentic analysis. Research is not yet sure if/how ASR may "correct" the speech produced by children with DLD in LSA. Research is also unaware of how clinicians who perform pediatric LSA are using ASR, as well as their opinions about ASR use for this population (Lammert et al., 2025).

To determine primary efficacy of ASR in LSA, researchers commonly rely on Word Error Rate (WER) – a standard metric for transcription accuracy. In a WER calculation, two transcripts are needed from the same language sample: a human-coded transcription and an ASR-coded transcription. The substitutions, omissions, and additions between the two files are calculated to determine overall WER. A lower WER indicates higher alignment between the ASR transcription and human-coded transcription, and by extension, greater ASR accuracy. This initial calculation provides baseline information about the status of ASR transcription in comparison to human transcriptions and if ASR can reliably support transcription tasks currently performed by humans. Prior research has explored the impacts of ASR's lack of training on pediatric populations and found weaker cultural and linguistic sensitivity (Meera et al., 2025)

and higher WER for children from lower socioeconomic backgrounds (Attia et al., 2023). WER provides baseline insight into transcription performance, but without the consideration of cultural, linguistic, and population characteristic factors that humans can.

While WER offers a quantitative measure of ASR performance, considering the factors that contribute to high or low error rates is critical. Our study focuses on the pediatric DLD population. We aim to connect population-specific factors and potential sources of ASR challenge. If humans experience challenges in transcription accuracy by way of speech intelligibility, background noise, and quality of the sample, does ASR experience the same challenges and at the same level? There is currently limited research on how ASR experiences similar or different challenges in the LSA process as humans. This study focuses on four variables that may influence WER: (1) the true error rate between ASR and human transcriptions, (2) speech intelligibility, (3) background noise interference, and (4) audio quality.

The true error rate is the goal of WER calculation, but the understanding of how audio quality, speech intelligibility of samples, and background noise interference interact with WER are relevant considerations for understanding the efficacy of ASR for clinical applications. Some research has investigated the use of ASR in assessment classification of DLD versus typical language (TL) (Bååth et al., 2019; Ortiz & Huang, 2024), but there remains a gap in the literature exploring automated LSA in clinical settings (Lammert et al., 2025). Even when trained on how to use automated analysis software (e.g., CLAN, SALT), clinicians still report time and knowledge barriers of how to efficiently and effectively use these programs (Klatte et al., 2022). So, determining WER of ASR adds to research, but likely does not add to clinical practice, which can increase the research-to-practice gap (Olswang & Prelock, 2015) and results in

clinicians not using ASR tools. By identifying other metrics or sample conditions that could potentially influence transcription accuracy, we can not only begin to define standards for language samples that consistently yield low WER, and in turn accurate transcriptions, but also better train clinicians on how to responsibly use ASR.

Instead of focusing solely on minimizing WER, it is relevant to understand under what conditions WER increases, particularly in clinical contexts. The variables explored in this study (e.g., background noise interference, audio quality, and speech intelligibility) reflect sample conditions and may present different effects on ASR. Background noise interference has presented issues in ASR software; however, some pretraining of noise on the software has shown improvements in reducing error rates (Attia et al., 2023). Audio quality is another variable that has been explored in some adult settings. Researchers have found that output of ASR software is influenced by the quality of language sample input (Liu et al., 2023) and high-quality recording equipment (Vogel & Morgan, 2009). Speech intelligibility is the third variable that has been mentioned as influence in ASR as the training required for ASR to be successful in transcription is greater for pediatric populations, compared to adult populations (Lüdtke et al., 2023), due to the diversity in speech intelligibility. Current research only offers preliminary speculation about how these three variables impact ASR performance. Thus, there remain significant gaps in understanding how LSA-specific audio variables in the pediatric DLD population function as barriers to accurate transcriptions. By addressing these gaps, we can not only improve ASR systems but ensure proper use in clinical applications.

Beyond transcription accuracy, evaluating clinical utility of ASR requires examining how well it replicates language sample metrics relevant to assessment. These include standard LSA metrics

such as mean length of utterance (MLU), total number of words (TNW), and number of different words (NDW). Accurate metrics are one of the baselines for DLD classification. This potential diagnostic application of ASR draws from previous LSA research that investigated how certain LSA metrics best distinguish DLD-specific deficits from other language impairments (Guo & Scheider, 2016). LSA studies have highlighted that the ability of metrics to distinguish DLD from TL depends on the specific indices selected, particularly those capturing lexical diversity (Guo & Schneider, 2016; Bååth et al., 2019; Kapantzoglou et al., 2019; Treffers-Daller et al., 2018). Prior research has shown that novel multidimensional metrics specific to ASR can improve analyses of syntax in adults (Agmon et al., 2024). Since most research exploring ASR metrics has focused on adult samples, questions remain on how ASR can be applied to pediatric LSA. Further investigation is needed to determine whether ASR can support accurate, efficient, and developmentally appropriate language sample analyses in children with DLD.

Recently, Ortiz and Huang (2024) introduced LSA+, an ASR-based tool capable of distinguishing between children with and without DLD using metrics such as mean length of utterance (MLU), number of different words (NDW), and total number of words (TNW). Ortiz and Huang reported strong correlations ($r > 0.7$) between ASR-generated and manually derived metrics. Notably, they also introduced a novel metric—mean length of speech segment (MLSS)—which outperformed conventional measures in sensitivity and specificity, highlighting the potential for ASR-specific metrics. Their dataset included 108 children aged 4 to 5 years both with typical language (TL) development (n=91) and children with DLD (n=17). These findings suggest that ASR can not only replicate traditional LSA metrics but may also yield new diagnostic indicators that enhance classification accuracy. This opens promising avenues for the use of ASR in both diagnostic and treatment planning contexts, particularly through the

continued exploration of novel, ASR-generated metrics tailored to the unique speech characteristics of children with DLD.

In addition to its role in classification, LSA also offers strong clinical utility in monitoring children's language progression over the course of treatment (Gallagher & Hoover, 2020; Wilder & Redmond, 2022; Wilder & Redmond, 2024). Clinicians express interest in using LSA as an indicator for treatment efficacy for children with DLD on their caseload. However, the same time and effort-based barriers exist (Klatte et al., 2022; Wilder & Redmond, 2022). Although there have been recent attempts in automating the coding and scoring of language metrics in narrative language samples that monitor progress in language treatment (Fox et al., 2022), research has still not validated a tool that automates the transcription and analysis of LSA for assessment nor progress-monitoring. In the pursuit of validating an ASR tool that accomplishes all current uses of LSA, it is not only important to investigate transcription accuracy and metric validity but also examine how sensitive ASR-derived metrics are to morphosyntactic changes over the course of treatment.

Despite these recent advancements in ASR, a critical question remains: Can ASR systems adequately detect the wide variability of linguistic profiles seen in children with DLD? Additionally, we also aim to investigate how different sources of noise commonly present in telehealth language samples (e.g., background noise, device interference, speaker movement) interact with transcription accuracy and ASR performance. The current study leverages a larger and more age-diverse dataset centered on children with DLD. Research focused on children with DLD requires not only larger data sets due to population variability, but more diverse datasets to increase ASR sensitivity (Lammert et al., 2025). This dataset acknowledges how DLD language

profiles and transcription challenges vary substantially across developmental stages by utilizing an R01- funded dataset comprising a subset of 44 children with DLD, ages 4 to 9, who participated in a treatment efficacy study. Participants received either recast therapy or book reading therapy, conducted by a clinician or clinician-trained parent. Language samples were collected for both pre and post intervention, with a goal of tracking efficacy of treatment models. Each sample was transcribed manually and processed using ASR, enabling direct comparisons. Additionally, each participant also completed standardized assessments, including the Differential Ability Scales II (DAS-2) and the Diagnostic Evaluation of Language Variation – Norm Referenced (DELV-NR), allowing for further contextual analysis. The present study aims to evaluate ASR as a tool for increasing efficiency in language sample analysis for children with DLD, while also identifying the factors that contribute to transcription inaccuracies. Beyond transcription accuracy, this study also examines how ASR-derived metrics can aid in clinical decision making, particularly as a tool for progress monitoring during the course of treatment. Specifically, we address the following research questions:

1. How accurately does ASR perform in language sample transcription for children with DLD?
2. What audio-file variables are correlated to high WER?
3. What is the concurrent validity of LSA metrics derived from ASR, relative to their manual counterparts, for children with DLD?
4. How sensitive is ASR to morphosyntactic changes over the course of treatment for children with DLD, relative to manual LSA?

CHAPTER 2: METHODS

Participants

Forty-four children with DLD aged 4;0 to 9;8 (years; months) were a part of a larger data set comprised of 136 children who participated in a randomized clinical trial (Koppy et al., 2025). These children were recruited from friend/family referrals, their speech-language-pathologist, school, Facebook, BuildClinical (a social media recruitment campaign platform), previous study participations, Reddit, or other. Participants reported locations from 23 different states in America: Arkansas, California, Colorado, Delaware, Florida, Idaho, Illinois, Iowa, Indiana, Kansas, Kentucky, Louisiana, Massachusetts, Michigan, New York, North Carolina, Ohio, Oklahoma, Pennsylvania, South Carolina, Texas, Utah, and Virginia. Of the participants whose parents reported gender, 26 reported that their child was a boy, and 16 reported that their child was a girl. Parents also reported ethnicity of children; of the 44 children there were reports of 2 American Indian, 5 Asian, 10 black, 1 native Hawaiian, 27 white, 8 Hispanic, and 5 reported as other.

All participants were screened to determine if they presented with DLD through standardized assessments. Standardized information on language performance (syntax, semantic, and pragmatic subtests) was collected through the Diagnostic Evaluation of Language Variation – Norm Referenced (DELV-NR) (Seymour et al., 2005) assessment. Assessments were chosen for inclusionary criteria of dialectal differences in language performance. A standard score of below an 85 on the DELV qualified children as having DLD. Participants in this study averaged a DELV composite standard score of 77.5 (SD=6.36) with a minimum standard score of 53 and a maximum standard score of 84. The Differential Ability Score II (DAS-2) (Elliot, 2007) was the other standardized assessment used for inclusionary criteria; children's t-scores had to be at or

above 35 for qualification of the study. The DAS-2 (Elliot, 2007) assessment measured the nonverbal cognitive abilities unrelated to language-learning for children. Children who presented with other psychiatric disorders or concerns were not qualified for the study. Children whose speech was determined to be >50% unintelligible (i.e., mean length of utterance had to be >2.5 words in length) by trained clinicians were also excluded from the study.

Materials and instruments

Language sample analysis is an expansive assessment of someone's language skills. The language samples used were narrative language samples where children participated in a story-tell task. In eliciting narrative samples, the Edmonton Narrative Norm Instrument (ENNI) (Schneider, 2005) was used. This is a set of three to six unique stories centered around animal characters. The ENNI instrument was displayed over zoom, and children were instructed to first silently look through the story and then verbally tell their story. Narrative language samples were recorded over zoom.

The automated language sample analysis software used was LSA+ (Ortiz, 2024). This software has been used in previous studies that achieved reliable classification accuracy for children with DLD from children with typical language development (Ortiz & Huang, 2024). The model used in LSA+ was a Faster Whisper medium model (SYSTRAN, 2025), which has the ability to quickly and thoroughly go through large masses of audio files and create automated transcriptions as well as analysis. Control settings for the transcription were set to align with the protocol which included 1) number of clauses set to two clauses per utterance 2) number of speakers set to one 3) a supplemental raw transcript and pre-ASR audio segmentation transcript 4) analysis metrics MLUw, MLUm, NDW, NDW lemma, and type token ratio (TTR).

All participants participated in pre-test data screening that identified expressive and receptive morphosyntactic performance to determine qualification for the study. Participants were randomly assigned to receive treatment either from a clinician via Zoom or a trained caregiver. Treatment spanned over 10 weeks and targeted passive or object-relative constructions through recast or book reading therapy. Children who received recast therapy were expected to obtain 960 exposures or 45–60-minute sessions for a total of 16 hours. Children who received book reading therapy were expected to obtain 960 exposures or 15–20-minute sessions for a total of 16 hours. For the subset of children in this study (n=44), 24 received treatment from clinicians while 20 received treatment from their trained caregiver; 22 received recast therapy while the other 22 received book-reading therapy; 26 received treatment targeting object relative clauses while 18 received treatment targeting passive constructions.

The aim for 16 hours of treatment for all conditions spanned over 10-weeks. Post-intervention data collection occurred two weeks after treatment wrapped up, where another narrative language sample analysis was conducted with an alternative version of the ENNI materials. Both pre-intervention language samples and post-intervention language samples were transcribed for further analysis. They were first transcribed and coded manually via SALT (Miller & Iglesias, 2012) by trained research assistants. Each transcription was then checked by at least one other research assistant. Two children in this study withdrew from treatment; however, both of those children participated in the secondary language sample analysis.

Automated metrics

ASR transcripts were generated through a process created by lead author (M. Jones). For this step, original audio files of the narrative samples were chopped and edited on Adobe Audition to only include the kid-speech. The edited audio files were then fed through LSA+ transcription tab.

LSA+ uses voice activity detection and natural language processing to generate transcriptions with timestamps that allow for speaker diarization and segmentation of speech for metric analysis. After inputting audio files, .txt files are automatically generated and uploaded for analysis on LSA+. In the analysis tab of LSA+, users can click to automate standard analysis on the .txt files that were automatically uploaded during transcription. This standard analysis provides users with MLUm, MLUw, NDW, NDW (lemma), TNW, syllables, type token ratio, and utterance values. Per the selected LSA+ settings including an output of a supplemental raw transcript, further ASR-specific analysis was performed by uploading the supplemental raw transcript .wrd file into LSA+ analysis tab. This analysis provided measurements for mean length of speech segment (MLSS) detected in ASR transcription. MLSS metric represents syntactic length values by measuring the average number of words produced in an identified speech segment based on voice activity detection (Bredin et al., 2020; Ortiz & Huang, 2024). MLSS is an emerging metric used in ASR analysis with recent evidence of specificity of DLD identification through LSA (Ortiz & Huang, 2024). This metric is not expected to be used in manual analysis due to impracticality of measurement processes and the metric measurement relying on voice activity detection algorithms that have been pre-trained to detect true continuous speech segments (Bredin et al., 2020).

Word error rate (WER) computation was used to examine transcription accuracy between ASR transcriptions and manual transcriptions. It takes the substitutions, insertions, and deletions between transcriptions and divides that sum by the total number of words in the transcription. A lower WER signals higher accuracy while a higher WER signals lower accuracy. The words in files are strung together in a single line for computation. To prepare files for analysis, human-

transcribed files were cleaned up by removing certain SALT codes (Miller & Iglesias, 2012) for WER computation.

Manual coding

Each participant audio file was then blindly rated (i.e., rater was masked to diagnostic status of each participant) on child intelligibility, background noise interference, and audio quality. Intelligibility of child refers to the perceived understanding of the child's speech. Background noise refers to any noise present in environment that is not the speech of the child. Audio quality refers to unique characteristics of the recording file, atypical to noise that would be present in real-time. Ratings occurred on a Likert scale of 1-5. Each Likert scale rating also included qualitative information that aided in the decision of the scaled rating. Descriptions were created by a research assistant who listened through 75% of audio files while developing the scales. These scales were then tested for interrater reliability by a trained speech-language pathology graduate student. The reliability rater was randomly assigned 44% of the audio files and used the scale to assign raters for the three acoustic dimensions. Reliability scoring occurred on a designed agreement scale. If both coders agreed on the same value, that acoustic factor rating for the audio file was scored as a 1.0 agreement; if coders differed by one numerical value, that acoustic factor rating for the audio file was scored as a 0.5 agreement; if coders disagreed by more than one numerical value, then the acoustic factor rating for the audio file was scored as a 0 agreement. All agreement scores were then added up and divided by the sum of participants scored. A 91% agreement was achieved for speech intelligibility and background noise ratings while a 92% agreement was achieved for audio quality ratings.

Dimension #1: child intelligibility. Child intelligibility ratings ranged from very poor (rating 1) to very good (rating 5). At a score of 5 (very good), the child's speech intelligibility in

audio files can be described as “clear”. A score of a 5 is also determined to have required no additional effort in understanding the child during manual transcription. A score of a 4 (good) was determined to have occasional unintelligibility that required minimal effort to understand. Misarticulation is seldom present in audio files with this score. At times, children would also eat during their narrative sample, and this was considered into speech intelligibility scoring, since the food typically interfered with precise articulation. At a speech intelligibility score of a 4, some occasional eating could be possible, but nothing that interferes with speech intelligibility. Identifiable misarticulations (e.g., gliding) may occur occasionally, but it seldom prevents misunderstandings of what the child said. At a speech intelligibility score of a 3 (fair), there is a moderate amount of unintelligibility that takes a moderate amount of effort to understand. The child’s speech in these files could now have multiple misarticulations and/or phonological processes that occur within every utterance. Any eating of food in these audio files starts to somewhat interfere with articulation. At a speech intelligibility score of a 2 (poor), the transcriber will likely have to relisten to understand the child’s speech. There is now a maximum effort to understand the speech. Any eating that took place during the narrative sample interfered with articulation. At a speech intelligibility score of a 1 (very poor), a majority of the audio file is unable to be transcribed due to an inability to understand the child’s speech, even with relistening.

Dimension #2: background noise. Background noise interference ratings ranged from very noisy (rating 1) to no noise interference (rating 5). An audio file would receive a rating of 5 (no noise), if there was no background noise present. If an audio file receives a rating of 4 (low noise), then there may be some background noise present; however, it does not distract the speaker's signal. Common acoustic factors that resulted in a rating of 4 for background noise

interference would be occasional adult (e.g., parent, speech language pathologist) interruptions that overlapped with child speech. At a rating of 3 (moderate noise), an audio file was determined to have noise that is slightly distracting from the speaker. The background noise may be louder at times compared to the signal. Some observational patterns of this occurred when a child was using their tablet to be on Zoom, and would move their tablet around, creating some occasional noises. A background noise rating of a 2 (noisy), was given to audio files that had background noise that moderately distracted child speech. The child may be in an environment where other children and/or adults are constantly speaking in the background. At a rating of 1 (very noisy), the background noise in an audio file would be constant and very distracting from the child's speech.

Dimension #3: Audio Quality. Audio quality ratings ranged from very poor (rating 1) to very good (rating 5). To receive a score of 5 (very good), audio files were described to be "clear and crisp" and have no distortions. Distortions of audio quality were described to be "audio components that are recognizable to the human ear that the audio was recorded through a device rather than audio that is natural to the human ear". "Crisp and clear" audio quality acknowledges that Zoom recordings may have some grainy quality, but these recorded files sound as natural as possible. The grainy quality in these files goes almost unnoticeable when listening without intention of rating the file. A rating of 4 (good) described audio files as "crisp" with minimal distortion (e.g., faint buzzing of audio, graininess that overpowers clear quality). When audio files are rated at 4, the grainy quality is more noticeable throughout the recording. A rating of 3 (fair) described audio files as "dull" with occasional distortion as well as potential occasional audio fadeouts. Dullness quality was described as audio that sounds more robotic to the human ear. Some observable occurrences for an audio fadeout were when children would use a higher

pitch towards the end of their sentences. It is relevant to note that with context cues, both ASR and human-coders may be able to fill in a lost word. Audio files rated at a 3 may also have a louder sound source than typical, which can be described as “boomy” or “sharp” when the recorded speaker talks. A rating of 2 (poor) described audio files as “muddy” with more frequency distortions and audio fadeouts. Muddiness was described as audio quality that sounds like the speaker is talking through a large mass (e.g., mud). The audio quality might be "quieter" volume-wise, even with the volume turned to the maximum level. “Loud” audio files receive higher/better ratings than “quiet” volume levels since it is easier to recognize and transcribe files with loud speaking rather than quiet speaking. A rating of 1 (very poor) described audio files as “muffled”, “very distorted” and/or “very quiet”. To receive this rating, audio files would have to exhibit at least two qualities of distortion (e.g., quietness and muddiness, dullness and loudness).

CHAPTER 3: RESULTS

Question #1: How accurately does ASR perform in language sample transcription for children with DLD?

To understand how accurately LSA+ performed language sample analysis for children with DLD, ASR-derived transcripts were compared to manually derived transcripts by way of a WER computation. Computations for WER were done for both pre-treatment transcripts and post-treatment transcripts.

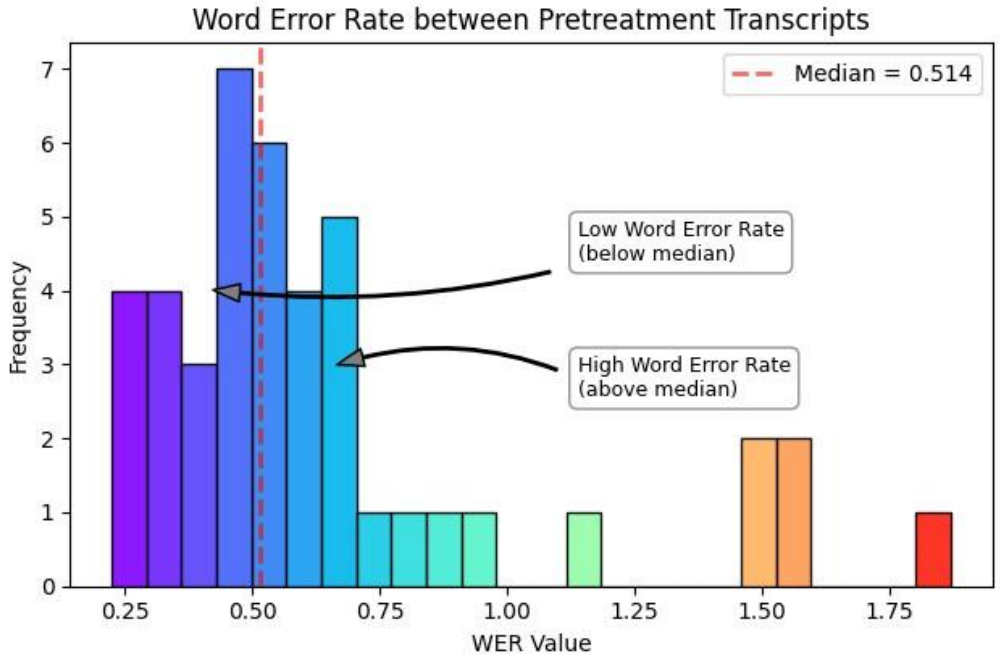

FIGURE I. WORD ERROR RATE BETWEEN PRETREATMENT TRANSCRIPTIONS. THE HISTOGRAM SHOWS FREQUENCY OF WER VALUES. DASHED LINE INDICATES MEDIAN SPLIT.

The average WER for the 44 pre-treatment transcriptions was 0.646, meaning that ASR transcriptions were about 65% inaccurate when compared to manual transcription counterparts. The minimum WER value of this group was 0.224 (~ 22% ASR transcription inaccuracy) while

the maximum WER value for this group was 1.87. WER values >1.0 mean that the ASR-derived transcript was 100% inaccurate and inserted additional inaccurate words.

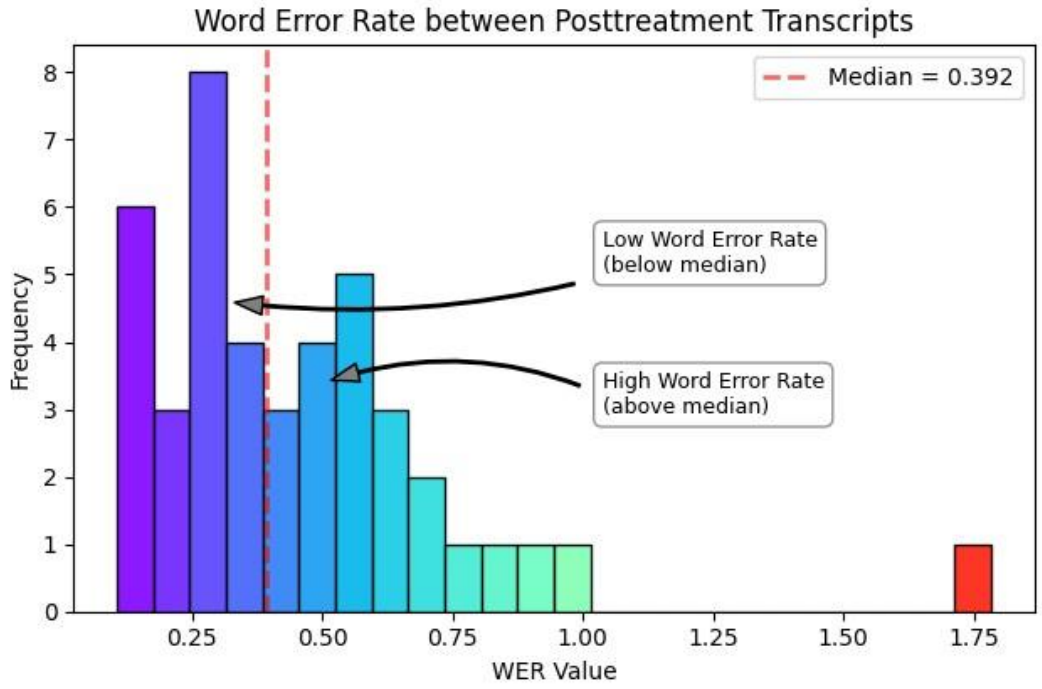

FIGURE II. WORD ERROR RATE BETWEEN POSTTREATMENT TRANSCRIPTIONS. THE HISTOGRAM SHOWS FREQUENCY OF WER VALUES. DASHED LINE INDICATES MEDIAN SPLIT

For the 44 post-treatment transcriptions, the WER average was 0.451, meaning that ASR transcriptions were about 45% inaccurate when compared to manual transcription counterparts. The minimum WER value of this group was 0.106 (~ 11% ASR transcription inaccuracy) while the maximum WER value for this group was 1.78.

Average WER values decreased about 20% from pre- to post-transcription, signaling higher rates of accuracy for children's transcriptions after treatment. WER values decreasing in post-transcription reflect ASR's improved performance for children with DLD after a course of language treatment, thus ASR performs better as children's language became more predictable.

Question #2: What audio-file variables are correlated to high WER?

To understand the relationship between our participant audio file dimensions (e.g., speech intelligibility, background noise interference, audio quality) and WER outcome, we conducted a multivariable regression analysis. In this analysis, WER is the outcome while the three audio file dimensions are predictors of that outcome. The analysis includes both the composite prediction of WER outcome and the three individual predictors of WER outcome.

TABLE I. MULTIVARIABLE REGRESSION ANALYSIS WHERE SPEECH INTELLIGIBILITY, BACKGROUND NOISE, AND AUDIO QUALITY ARE PREDICTORS OF WER OUTCOME.

	R squared	Adj. R squared	Coefficient	Std Error	P>|t|
Pretest Regression	0.652	0.625	1.6728	0.302	0.001
Speech Intelligibility			-0.3045	0.036	0.001
Background Noise			-0.0584	0.056	0.304
Audio Quality			-0.0457	0.048	0.351
Posttest Regression	0.666	0.64	1.4232	0.174	0.001
Speech Intelligibility			-0.235	0.174	0.001
Background Noise			-0.0392	0.032	0.226
Audio Quality			0.0152	0.28	0.596

Within the pre-treatment group, an R-squared value was calculated to determine the level of variance the three predictors (i.e., speech intelligibility, background noise, audio quality) have on the outcome value (i.e., WER). The R-squared value for the pre-treatment group was 0.652, meaning that about 65% of the WER variance can be explained by the three acoustic factors. The coefficient value in a multivariable regression analysis represents the predicted weight of change

in the outcome value for every 1-unit change in the predictor variable. The coefficient value for the pre-treatment regression analysis was 1.6728, meaning that together, the three acoustic factors have statistical significance with 1-unit changes in WER. The coefficient value for speech intelligibility as a predictor of WER in the pre-treatment group was -0.3035, meaning that for every 1-unit increase in speech intelligibility ratings, WER values decreased by 0.3035 percentage points. Both background noise and audio quality did not have significant coefficient values.

In the post-treatment group, the R-squared value was 0.666, suggesting that about 67% of WER variance can be explained by the three acoustic factors. The coefficient value for the post-treatment group was 1.4232; the coefficient value for speech intelligibility as a predictor of WER in the post-treatment group was -0.235, meaning that for every 1-unit increase in speech intelligibility ratings, WER values decreased by 0.235 percentage points. Both background noise and audio quality did not have a significant coefficient value. Results suggest that speech intelligibility of children with DLD is the main acoustic-related predictor for ASR transcription accuracy, with lower levels of intelligibility correlating with higher WER values.

To further understand the strong negative correlation between speech intelligibility ratings of a participant and their WER, graphs were plotted to visually demonstrate this relationship with the consideration of participant age. Participants were categorized by age groups and distinguished by a colored shape (e.g., children aged 6;0-6;11 can be identified by a yellow triangle).

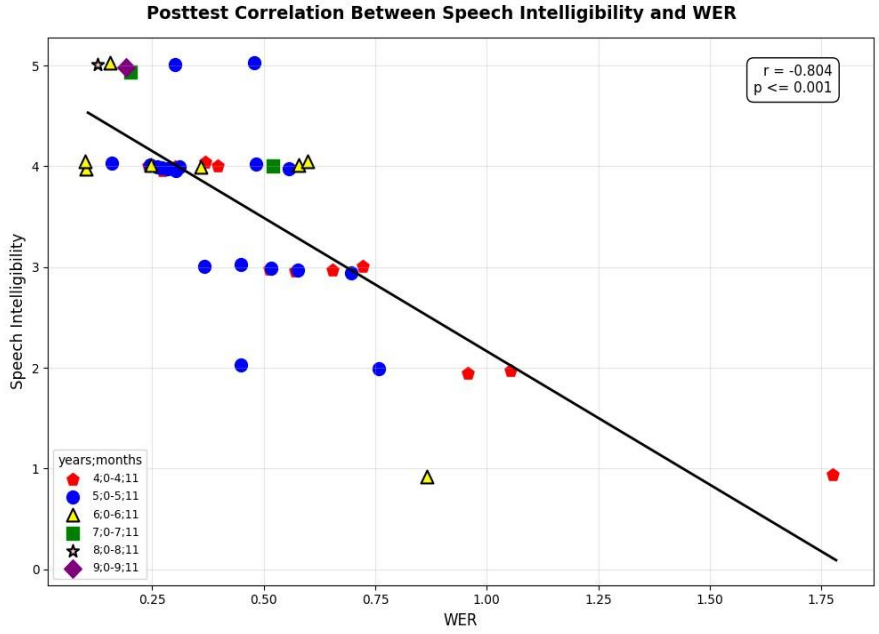

FIGURE III. CORRELATION BETWEEN SPEECH INTELLIGIBILITY RATINGS AND WER VALUES AT THE PRETREATMENT STAGE. THE LEGEND KEY INDICATES SHAPE ASSOCIATED WITH AGE OF EACH PLOTTED PARTICIPANT. RED PENTAGON FOR AGE GROUP 4;0-4;11, BLUE CIRCLE FOR AGE GROUP 5;0-5;11, YELLOW TRIANGLE FOR AGE GROUP 6;0-6;11, GREEN SQUARE FOR AGE GROUP 7;0-7;11, PINK STAR FOR AGE GROUP 8;0-8;11, AND PURPLE RHOMBUS FOR AGE GROUP 9;0-9;11.

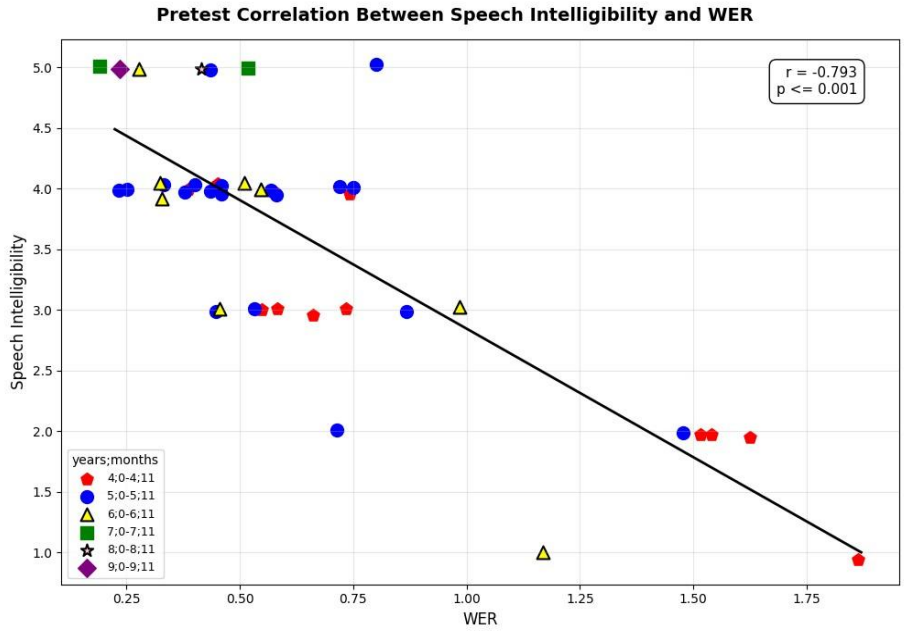

FIGURE IV. CORRELATION BETWEEN SPEECH INTELLIGIBILITY RATINGS AND WER VALUES AT THE POSTTREATMENT STAGE. THE LEGEND KEY INDICATES SHAPE ASSOCIATED WITH AGE OF EACH PLOTTED PARTICIPANT. RED PENTAGON FOR AGE GROUP 4;0-4;11, BLUE CIRCLE FOR AGE GROUP 5;0-5;11, YELLOW TRIANGLE FOR AGE GROUP 6;0-6;11, GREEN SQUARE FOR AGE GROUP 7;0-7;11, PINK STAR FOR AGE GROUP 8;0-8;11, AND PURPLE RHOMBUS FOR AGE GROUP 9;0-9;11.

Participants who were both rated for lower intelligibility levels and had higher WER were generally the youngest participants, while the oldest participants were consistently rated for higher intelligibility levels and lower WER. Younger participants in this study (children aged 4-5 years) were more likely to have lower levels of speech intelligibility ratings, which correlated with higher WER values. This suggests that younger children are more likely to have lower levels of speech intelligibility and thus less accurate ASR transcriptions.

Question #3: What is the concurrent validity of LSA metrics derived from ASR, relative to their manual counterparts, for children with DLD?

To understand the concurrent validity of ASR-derived LSA metrics, we analyzed correlations between metrics derived from ASR and manual transcripts through multiple bivariate correlation analyses. For each, we calculated Pearson's parametric to understand the magnitude of difference between ASR and manually retrieved metrics. We then created a bivariate regression plot to visualize the strength of the relationship between automated MLUm and DELV standard scores as well as manual MLUm and DELV standard scores.

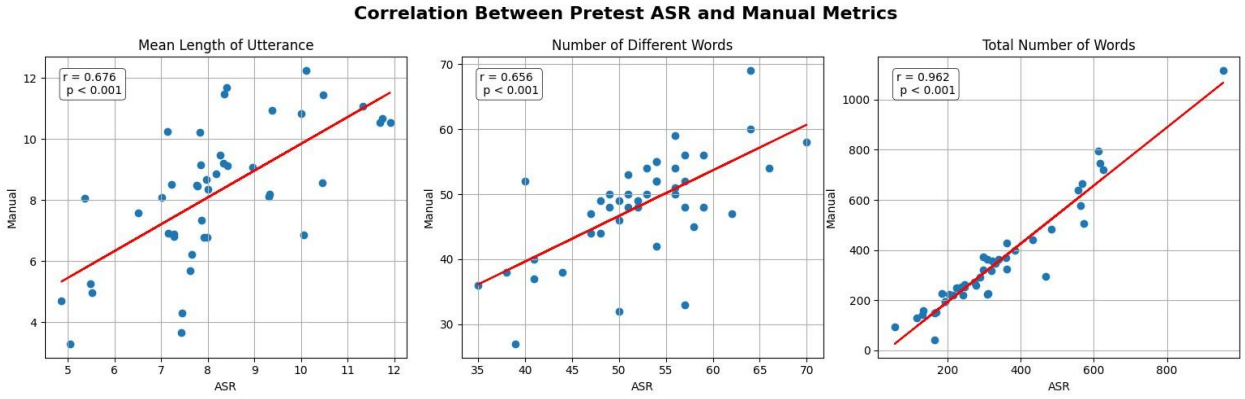

FIGURE V. CORRELATIONS BETWEEN MANUAL TRANSCRIPTION LSA METRIC MEASUREMENTS AND ASR TRANSCRIPTION METRIC MEASUREMENTS AT THE PRETREATMENT STAGE. METRICS INCLUDED ARE: MEAN LENGTH OF UTTERANCE (MLU), NUMBER OF DIFFERENT WORDS (NDW), AND TOTAL NUMBER OF WORDS (TNW), FROM LEFT TO RIGHT.

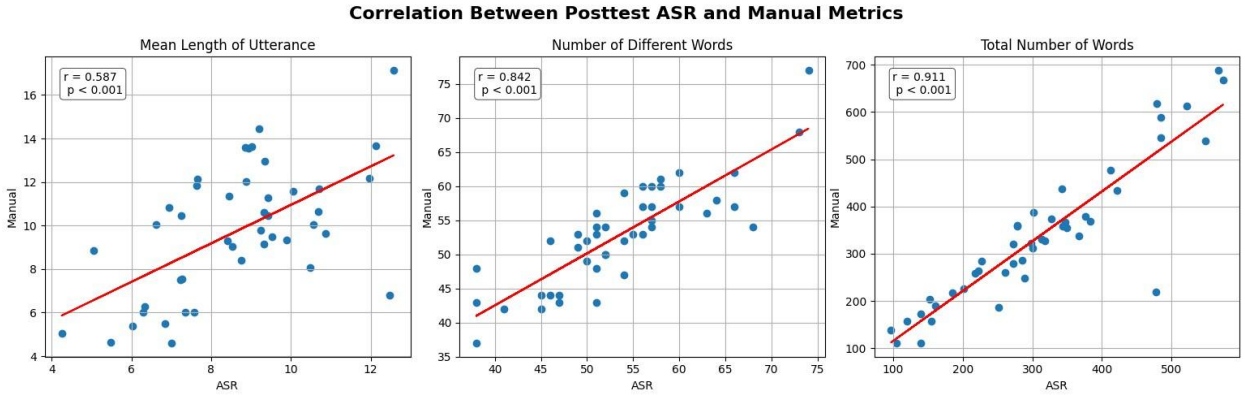

FIGURE VI. CORRELATIONS BETWEEN MANUAL TRANSCRIPTION LSA METRIC MEASUREMENTS AND ASR TRANSCRIPTION METRIC MEASUREMENTS AT THE POSTTREATMENT STAGE. METRICS INCLUDED ARE: MEAN LENGTH OF UTTERANCE (MLU), NUMBER OF DIFFERENT WORDS (NDW), AND TOTAL NUMBER OF WORDS (TNW), FROM LEFT TO RIGHT.

These calculations were performed for each of the 44 participants' pre-treatment and post-treatment transcriptions. For pre-treatment metric analysis, MLUm between ASR and manual analysis had a moderately strong positive correlation (r= 0.676). NDW between ASR and manual analysis also had a moderately strong positive correlation (r=0.656). TNW for pre-treatment ASR and manual analysis had a very strong positive correlation (r=0.962). For post-treatment metric analysis, MLUm between ASR and manual analysis had slightly decreased to a moderately positive correlation (r=0.587). NDW for post-treatment ASR and manual analysis increased to a strong positive correlation (r=0.842). TNW for post-treatment ASR and manual analysis maintained a very strong positive correlation (r=0.911). Results suggest ASR's strength with transcribing just as many words as manual transcribers do. They also suggest that ASR became more accurate with hearing just as many different words as humans did after participants received treatment. ASR, however, did not consistently transcribe the same type of words as humans did, which is reflected in lower MLUm correlations in both pre- and post-treatment transcriptions.

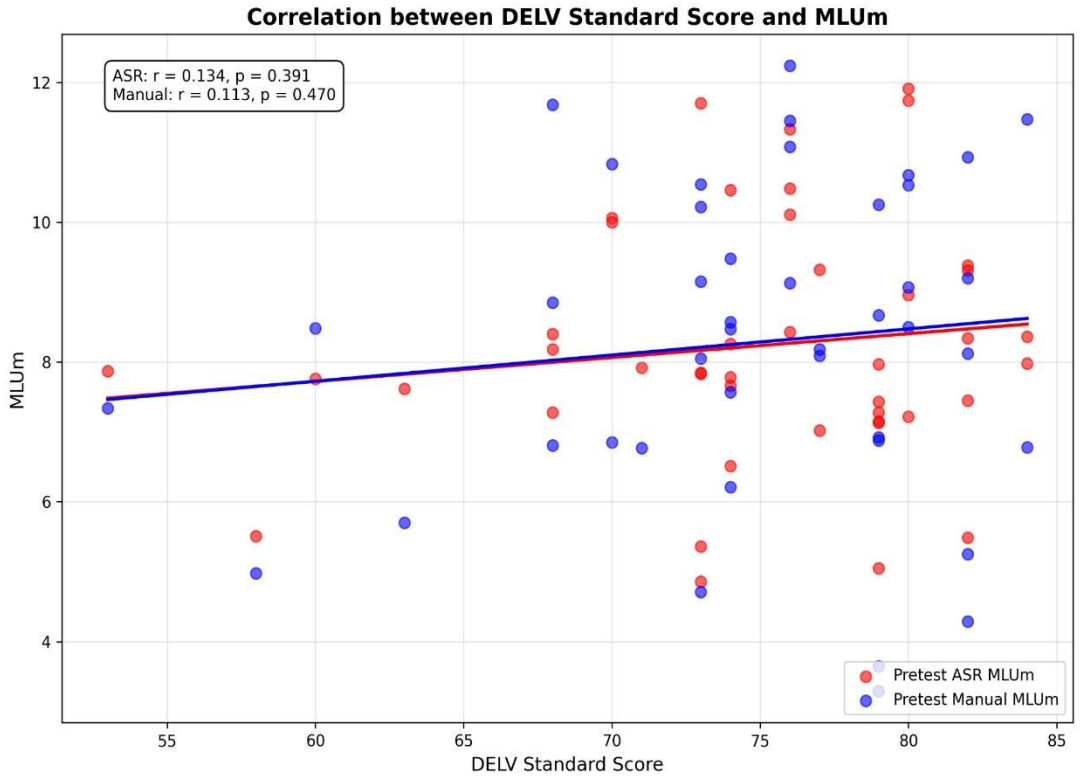

FIGURE VII. CORRELATION(S) BETWEEN PARTICIPANT DELV-NR STANDARD SCORE AND MEAN LENGTH OF UTTERANCE IN MORPHEMES (MLUM) METRIC MEASUREMENT FROM BOTH MANUAL (BLUE LINE AND DOTS) AND ASR (RED LINE AND DOTS) PRETREATMENT TRANSCRIPTION.

Correlations between pre-treatment ASR transcription and manual transcription MLUm analysis were compared to each participant's DELV composite standard score. This was done to better understand 1) the ground truth alignment between a child's standardized language assessment score and a common LSA metric (i.e., MLUm) 2) how the same metric in ASR transcriptions aligns with the standardized score. Both ASR MLUm and manual MLUm had very weak positive correlations with DELV standard scores (r=0.145; r=0.113, ASR;manual). Neither manual nor ASR transcriptions for MLUm calculations were strongly aligned with a child's DELV composite score, suggesting that both methods of transcription offer different insights to popular standardized assessments.

Question #4: How sensitive is ASR to morphosyntactic changes over the course of treatment for children with DLD, relative to manual LSA?

To understand how ASR can be used as a tool that measures progress over the course of treatment, we first conducted a paired t-test for pre-treatment and post-treatment effects in manual language metric analysis (i.e., MLUm, TNW, NDW) to capture a ground truth. Then, dependent on which metrics demonstrated significant pre-post growth, another paired t-test was calculated for that subset of ASR-transcription metrics. A final paired t-test was completed between pre- and post-treatment ASR-transcription MLSS metric, to better understand standard vs. Novel metric measurements from pre- to post-treatment.

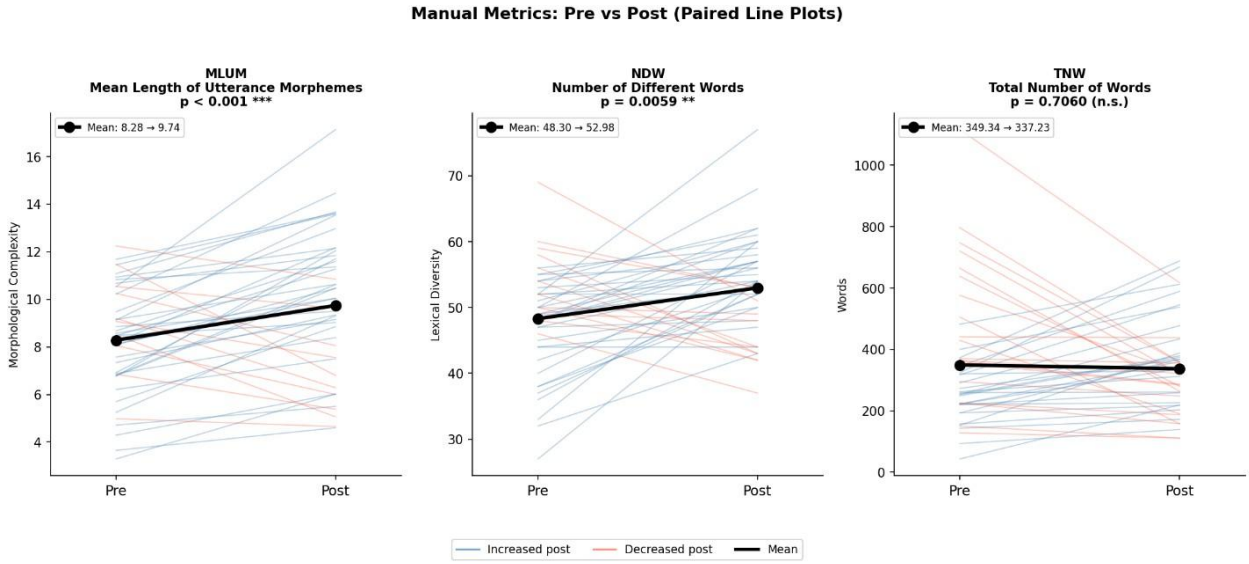

FIGURE VIII. PAIRED LINE PLOTS DEMONSTRATING CHANGE IN MANUAL TRANSCRIPTION METRIC MEASUREMENTS FOR MLUM, NDW, AND TNW FROM PRE- TO POST-TREATMENT. BOLDED LINE INDICATES MEAN CHANGE OF GROUP.

In our first set of paired t-tests to determine the ground truth of progress-monitoring in this study, manual-transcription metric analysis was compared from pre- to post-treatment. The paired t-tests reveal that MLUm increased significantly from pre- to post-treatment ($t(43) = -4.00$, $p<0.001$). Results also reveal that NDW increased significantly from pre- to post-treatment (t

(43) = -2.90, p = 0.006). Results, however, did not find significant change for TNW from pre- to post-treatment (t (43) = 0.38, p = 0.706). This means that manual transcriptions determine that children demonstrated a positive increase in morphological complexity and lexical diversity during the course of treatment, however, they did not demonstrate meaningful change in how many words they produced.

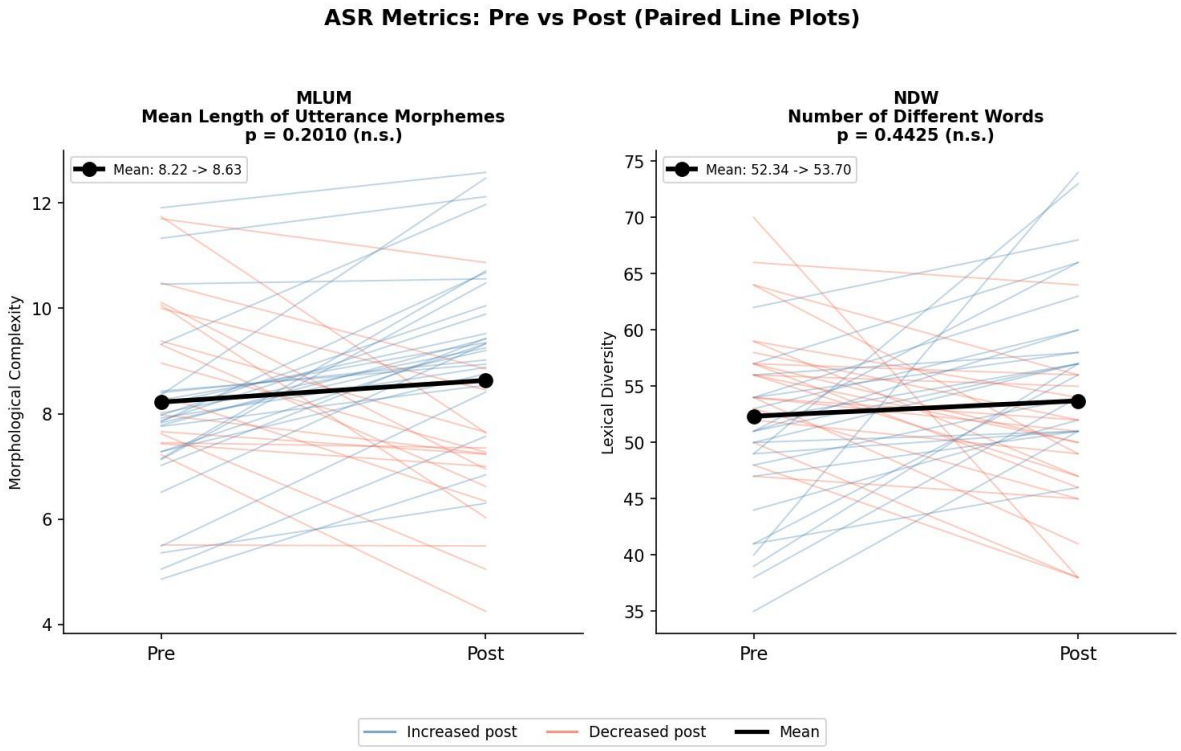

FIGURE IX. . PAIRED LINE PLOTS DEMONSTRATING CHANGE IN ASR TRANSCRIPTION METRIC MEASUREMENTS FOR MLUM AND NDW FROM PRE- TO POST-TREATMENT. BOLDED LINE INDICATES MEAN CHANGE OF GROUP

After the ground truth was established, two more paired t-tests were computed to track MLUm and NDWASR-transcription measurements from pre- to post-treatment. Results for these t-tests found no statistical significance or meaningful change for ASR's MLUm measurements from pre- to post-treatment (t (43) = -1.30, p = 0.201). Results also found that ASR's NDW measurements did not detect meaningful or significant change (t (43) = -0.78, p = 0.443). While ASR's MLUm and NDW also increased from pre- to post-treatment transcription, both

experienced smaller gains (ASR MLUm mean difference = 0.4127 morphemes; ASR NDW mean difference = 1.3636 different words) compared to manual-transcription MLUm and NDW. So, ASR transcriptions agree with metric gains as established through manual transcriptions; however, results demonstrate that ASR transcriptions are not capturing as many gains as manual transcriptions do. Thus, metric analysis in ASR transcriptions limited to MLUm and NDW to monitor progress over the course of treatment is not evidenced to currently replace the same analysis in manual transcriptions.

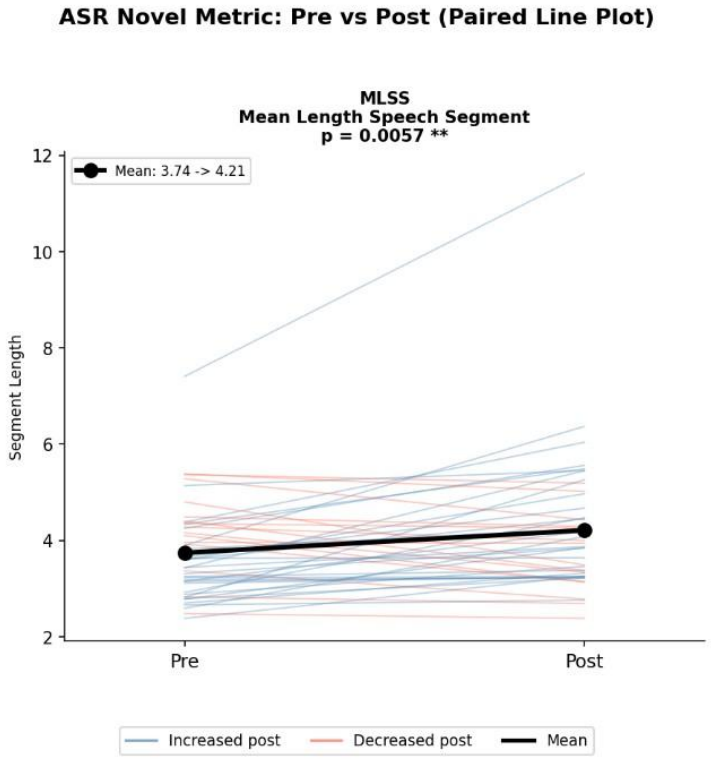

FIGURE X. . PAIRED LINE PLOTS DEMONSTRATING CHANGE IN ASR TRANSCRIPTION METRIC MEASUREMENTS FOR MEAN LENGTH OF SPEECH SEGMENT (MLSS) FROM PRE- TO POST-TREATMENT. BOLDED LINE INDICATES MEAN CHANGE OF GROUP.

In our final paired t-test analysis, we looked at how ASR-specific metric, MLSS, tracked changes from pre- to post-treatment transcriptions. Results from this t-test analysis discovered statistically significant differences in pre- to post-treatment transcriptions (t (43) = -2.908,

p=0.006). This suggests that ASR detected that participants demonstrated less fragmented and more connected speech segments, even if it did not detect the same growth in morphological complexity. MLSS metric was the only ASR-transcription metric to reach statistical significance, signaling that ASR metric analysis experienced greater sensitivity in measuring treatment gains for an acoustic-based analysis rather than standard morphological and lexical analysis.

CHAPTER 4: DISCUSSION

To assist clinicians with navigating novel ASR software, this study investigated what factors influence the accuracy of ASR software, LSA+, for narrative language samples from children with DLD at both the transcription and analysis stage. It explored what audio-related factors at the recording stage may influence transcription accuracy as well as what ASR-derived metrics may have influence over analysis accuracy. Without understanding what factors may be influencing ASR, 88 different narrative language samples transcribed by ASR achieved ~45% accuracy. From pre- to post-treatment transcription, accuracy levels for ASR improved by ~20%. After running statistical analyses to understand the influence of the three selected acoustic factors in this study, ASR accuracy was strongly correlated with perceived speech intelligibility of children in both pre- and post-treatment. In general, ASR transcription seems to perform more accurately for older children who have higher perceived levels of speech intelligibility. ASR also seems to be more robust to changes in background noise and audio quality. These results align with previous studies that have determined that ASR experiences most difficulty with transcribing speech as it becomes more unintelligible and unpredictable to what it may be trained on (Lüdtke et al., 2023; Lammert et al., 2025; Meera et al., 2025).

To observe how ASR measures progress over time, paired t-test analyses explored current ability for automated LSA to act as a progress-monitoring tool for clinicians. The participants in this study demonstrated significant gains in morphological complexity and lexical diversity from pre- to post-treatment, as demonstrated by metric analysis gains in manual transcription MLUm and NDW. ASR transcriptions did not track the same meaningful gains in their metric analysis from pre- to post-treatment. These findings align with this study's previous analyses exploring metric correlations at fixed points in time; ASR consistently demonstrated the ability to hear as many

words as humans, but it is not measuring syntactic complexity and vocabulary diversity to the same degree as humans do. Based on the bivariate correlation plots for ASR metrics and manual metrics, NDW metrics were more correlated when children were producing slightly more complex language; however, MLUm metrics decreased in correlation. NDW and MLUm metrics may not be appropriate metrics for ASR transcriptions used to monitor progress, but they do present abilities to provide insights at fixed points in time.

While ASR-transcription MLUm and NDW did not track similar gains, an ASR-transcription specific metric MLSS did. This highlights what metrics may perform better during analysis for ASR-transcription, when used as a progress-monitoring tool. Since results from the statistical analyses suggest discrepancies between how different ASR-metric analysis measures language gains, it is relevant to explore how the transcription preparation differs between metrics which may be able to explain some of the differences during analyses. Thus, we can start to better understand where in the LSA stages are discrepancies created (i.e., during ASR transcription or analysis?). MLU metric computation is bound by how utterances are determined during transcription. In our study, we instructed LSA+ transcription to determine utterance boundaries by allowing up to two clauses per utterance segment. Manual transcriptions follow SALT (Miller & Iglesias, 2012) transcription rules to segment utterances; these include basing utterances off grammar (C-units = main clause or main clause and subordinate), intonation/pauses, and speaker turns. While such standards for manual transcription exist, conventions still acknowledge that determining utterance boundaries becomes difficult as speech increases in complexity, so standards exist to assist human transcribers judgement while standardizing individual transcriptions for LSA. So, it is unsurprising that MLUm metrics performed more poorly as a

progress-monitoring tool for ASR, given the difficulties that exist with performing this metric during manual LSA.

In comparison, MLSS measures the length of children's speech segments by way of voice activity detection. This transcription method relies on a voice activity detection algorithm created by probability thresholds from pretrained models (Bredin et al., 2020), where such models have been validated against detection errors, false alarms, and detection misses. In this transcription method, MLSS metric does rely on a more standardized transcription process (than MLU), yet also a potentially more individualized transcription process as ASR voice activity detection prioritizes speech as segment boundaries, rather than clinical judgement and grammar. This becomes more relevant for LSA in the DLD population, as their grammar is more likely to be unpredictable for both humans and automated software. So, while manual LSA performs better as a progress-monitoring tool when relying on standard conventions and metrics, this does not discount the potential for ASR LSA to be used as a progress-monitoring tool. ASR's novel metric analysis, such as MLSS, may offer relevant and ASR-bound insights into speech and language development – particularly into language processing speech/fluency (i.e., MLSS metric).

Results from this study reveal introductory strengths and limitations of ASR in both the transcription and analysis stages for LSA+. In the transcription stage, poor intelligibility levels are consistently linked to high WER. In the analysis stage, even if transcription accuracy increases, ASR is still having trouble measuring standard LSA metrics (i.e., MLUm) compared to manual transcriptions. ASR experienced the highest correlations with TNW, demonstrating a strength in ASR hearing just as many words as humans do. Lower correlations for MLUm and

NDW suggest that ASR may not be hearing as many morphological boundaries or morphological variants as humans do. Since ASR consistently hears and transcribes just as many words as humans do but may not be tracking just as many morpheme boundaries, this may be the reason for the paired t-test results discovering that ASR does not track MLUm and NDW changes similar to humans. When examining progress by way of ASR-novel metric, MLSS, overall significant growth in longer continuous speech segments was measured, demonstrating one way ASR may be successfully utilized for progress-monitoring.

These findings support our hypothesis that ASR transcription accuracy is most impacted by speech intelligibility of children. They also support our hypothesis that ASR will measure language growth like manual transcriptions; the findings, however, do not support that ASR is measuring growth in language the same way humans do. ASR transcription and analysis, while imperfect, still seems to offer relevant and novel insights into specific language needs compared to reliance on standardized assessments. The strong correlation ASR and manual metrics shared compared to their weak correlation with the DELV standardized assessment demonstrates a clinical significance of LSA as a tool, including ASR.

CHAPTER 5: LIMITATIONS

The participants in this study were originally a part of a parent study that consisted of around 136 children. To investigate the longitudinal measurements of ASR, this study was only able to use complete data from 44 participants. Of these participants, 44.19% of the participants were the second youngest-aged participants (5;0-5;11, years; months). 27.19% of these participants were of the youngest age group (4;0-4;11, years; months). The youngest participants in this study were strongly correlated with poorer speech intelligibility ratings. While this study is limited by age diversity, analyses were previously done with 68 participants' pre-treatment samples. In this pilot analysis, MLUm, NDW, and TNW correlations between ASR and manual transcriptions achieved stronger correlations in counting MLUm (pilot analysis pre-treatment metrics: MLUm r=0.75, NDW r=0.66, TNW r=0.96). For greater insights into ASR analysis performance of all age groups, data analyses should be rerun with a larger sample size to observe potential changes in results.

It is also relevant to note that per parent study participant qualifications, children in this study had to be at least 50% intelligible by clinical judgement. While this study emphasizes the importance of speech intelligibility as an influential factor on ASR transcription, it is limited in understanding the full impact of lower speech intelligibility levels. Since results demonstrated significant impact of lower speech intelligibility levels for participants who were at least 50% intelligible, it is sensible to predict that ASR will produce less accurate transcriptions for children with even lower ratings.

When determining transcription accuracy for LSA+, it is relevant to highlight the circumstantial differences between ASR and human transcription. Human transcribers consisted of a team of 5 trained undergraduate students studying speech and language pathology. These students were

familiarized with the ENNI materials that were given to the children for narrative sample retrieval. They were also transcribing child speech from a zoom recording that had not been edited; thus, they benefitted from clinician adult speech that would occasionally gloss what the child had said (i.e., repeat back what they heard the child say for auditory confirmation). On the ASR transcription side, to reduce the labor of retrieving ASR transcription, all adult speech was cut from speech recording before being fed through LSA+. While this was done to prevent additional labor of needing to go back and edit the files for speaker diarization for accurate language analysis, this also meant that LSA+ was not privy to any context clues humans experienced during transcription. Realistically, ASR will always have less background information about language samples compared to humans, given that humans administered the language sample materials/questions. In that regard, it will always be necessary for humans to go back and double check any ASR transcription and analysis.

The current fundamental limitation ASR exhibits in the analysis stage was only explored for the three standard LSA metrics. When investigating how sensitive a particular system (e.g., humans/manual, ASR) can measure progress in treatment, research has found that it is most appropriate to use analysis metrics that directly correlate to the treatment targets (Gallagher & Hoover, 2020). The children in this study were originally a part of a parent study where they received recast therapy targeting passive structures and object relative clauses. Based on this study's results, we can predict that while children use more predictable forms of language, ASR becomes more accurate during transcription. We can also predict that since ASR had some discrepancies when calculating MLU and NDW during analysis, it may not always capture certain morphemes such as past tense -ed, which would likely increase if children were increasing their use of passive structures. From these predictions, we can conclude that ASR

analysis demonstrates a need for refinement in progress-monitoring, and that future research should rerun analysis with different metrics more specific to treatment targets. Research has also found that MLU is a weak analysis measurement in progress-monitoring (Gallagher & Hoover, 2020), so while MLU seems to be a weaker metric for ASR, that is not entirely unique to ASR. Future research should also explore what metrics/treatment targets ASR achieves the highest accuracy in calculating, compared to manual analysis.

Our research found ASR-specific metric, MLSS, captured the most language growth for ASR transcriptions over the course of treatment. The current study is limited in examining the full potential of using ASR as a progress-monitoring tool as there are other ASR-specific metrics that exist (e.g., pauses per word). Since children that are classified under DLD still exhibit varied language profiles, a singular metric to track progress for a group "fits all" approach should be reconsidered in future research. Instead, future research might benefit from moving away from group-level outcome metrics and more towards measurements that can better capture individual treatment gains.

CHAPTER 6: CONCLUSION

LSA offers valuable insights into language treatment targets for children with DLD. It also offers clinicians a way to monitor treatment-specific progress. While many clinicians have opted to reduce their use of LSA due to the labor involved, some clinicians have begun to rely on ASR tools to reduce labor and still reap benefits of LSA (Lammert et al., 2025). To provide clinicians with straightforward standards when using ASR for children with DLD, this research investigated factors that influence the efficacy of both ASR transcription and analysis. Results from analyses comparing pre-treatment and post-treatment manual and ASR transcriptions, as well as audio-related influences on transcription accuracy levels can provide clinicians with the following strengths, limitations, and considerations when using ASR for this population:

During step one of LSA – speech recording retrieval – one acoustic-related variable had the most impact on transcription

- ASR is most sensitive to changes in speech intelligibility rather than changes in background noise and audio-recording equipment. Backgrounds with nosier environments or audio files recorded with basic equipment are less likely to impact transcription accuracy. Clinicians seeking to obtain the most accurate transcription for children with DLD (who have lower levels of speech intelligibility) should opt for manual transcription or double check ASR transcription before the analysis stage.

During step three of LSA – analysis – metric selection for ASR matters for initial analysis and progress monitoring.

- TNW, NDW, and MLUm metric analysis for ASR transcriptions produced more accurate calculations when used for individual analyses rather than as a progress-monitoring tool.

TNW and NDW calculations in ASR transcriptions were overall more accurate than MLUm calculations, suggesting that manual intervention for transcriptions is required for more accurate ASR analysis of MLUm metrics. ASR-derived metric, MLSS, was the most insightful metric analysis for progress-monitoring. Clinicians may choose to use MLSS analysis to acoustically measure treatment gains.

These findings will begin to help guide clinicians on how to better navigate ASR to obtain the most accurate LSA for children with DLD. The expansive LSA process can still be altered in various ways, which furthers the question as to what factors can influence the efficacy of automated LSA for children with DLD. There are also remaining questions about potential ways automated LSA can continue to expand clinical utility of this process, while still saving clinicians time and labor. Future research should continue to explore what factors that clinicians can control during ASR transcription and/or analysis that impact the efficacy of LSA for children with

Appendix

Coding

Correlation graphs were created by scripts in Python to parse and manipulate the spreadsheet data and use that to create visual data. The libraries used include pandas, which was used to load and reshape the data, matplotlib, used for generating and saving the figures, and SciPy, for calculating the Pearson Correlation coefficient, (R value) and creating the linear regression lines. The Pearson correlation coefficient (r value) and corresponding confidence level (p value) were calculated and labeled for each graph. Any p value below .001 was rounded up to .001 and displayed as “p< .001”.

Bibliography

Agmon, G., Pradhan, S., Ash, S., Nevler, N., Liberman, M., Grossman, M., & Cho, S. (2024). Automated measures of syntactic complexity in natural speech production: Older and younger adults as a case study. *Journal of Speech, Language, and Hearing Research*, *67*(2), 545–561. https://doi.org/10.1044/2023_JSLHR-23-00009

Attia, A. A., Liu, J., Ai, W., Demszky, D., & Espy-Wilson, C. (2024). Kid-Whisper: Towards bridging the performance gap in automatic speech recognition for children VS. adults. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, *7*, 74–80. https://doi.org/10.1609/aies.v7i1.31618

Bååth, R., Sikström, S., Kalnak, N., Hansson, K., & Sahlén, B. (2019). Latent semantic analysis discriminates children with developmental language disorder (DLD) from children with typical language development. *Journal of Psycholinguistic Research*, *48*(3), 683–697. https://doi.org/10.1007/s10936-018-09625-8

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., Gill, M.-P., & ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) Barcelona, Spain 2020 May 4 - 2020 May 8. (2020). Pyannote.audio: Neural building blocks for speaker diarization. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7124–7128). https://doi.org/10.1109/ICASSP40776.2020.9052974

Elliot, C. D. (2007). Differential Ability Scales (2nd ed.). San Antonio, TX: Harcourt Assessment

Fox, C., Jones, S., Gillam, S. L., Israelsen-Augenstein, M., Schwartz, S., & Gillam, R. B. (2022). Automated progress-monitoring for Literate Language Use in Narrative Assessment (LLUNA). *Frontiers in Psychology*, *13*, 894478. https://doi.org/10.3389/fpsyg.2022.894478

Gallagher, J. F., & Hoover, J. R. (2020). Measure what you treat: Using language sample analysis for grammatical outcome measures in children with developmental language disorder. *Perspectives of the ASHA Special Interest Groups*, *5*(2), 350–363. https://doi.org/10.1044/2019_PERSP-19-00100

Guo, L.-Y., & Schneider, P. (2016). Differentiating school-aged children with and without language impairment using tense and grammaticality measures from a narrative task. *Journal of Speech, Language, and Hearing Research*, *59*(2), 317–329.

Lammert, J. M., Roberts, A. C., McRae, K., Batterink, L. J., & Butler, B. E. (2025). Early identification of language disorders using natural language processing and machine learning: Challenges and emerging approaches. *Journal of Speech, Language, and Hearing Research : JSLHR*, *68*(2), 705–718. https://doi.org/10.1044/2024_JSLHR-24-00515

Kapantzoglou, M., Fergadiotis, G., & Auza Buenavides, A. (2019). Psychometric evaluation of lexical diversity indices in Spanish narrative samples from children with and without developmental language disorder. *Journal of Speech, Language, and Hearing Research : JSLHR*, *62*(1), 70–83. https://doi.org/10.1044/2018_JSLHR-L-18-0110

Klatte, I. S., van Heugten, V., Zwitserlood, R., & Gerrits, E. (2022). Language sample analysis in clinical practice: Speech-language pathologists' barriers, facilitators, and needs. *Language, Speech, and Hearing Services in Schools*, *53*(1), 1–16. https://doi.org/10.1044/2021_LSHSS-21-00026

Koppy, A., O'Fallon, M., Huang, Y. T., & Van Horne, A. O. (2025). Distributional patterns in recast therapy do not systematically align with patterns in conversational data. *Journal of Communication Disorders*, *118*, 106576. https://doi.org/10.1016/j.jcomdis.2025.106576

Liu, H., MacWhinney, B., Fromm, D., & Lanzi, A. (2023). Automation of language sample analysis. *Journal of Speech, Language, and Hearing Research*, *66*(7), 2421–2433. https://doi.org/10.1044/2023_JSLHR-22-00642

Lüdtke, U., Bornman, J., de Wet, F., Heid, U., Ostermann, J., Rumberg, L., van der Linde, J., & Ehlert, H. (2023). Multidisciplinary perspectives on automatic analysis of children's language samples: Where do we go from here? *Folia Phoniatrica et Logopaedica*, *75*(1), 1–12. https://doi.org/10.1159/000527427

Meera, S. S., Swaminathan, D., Venkata Murali, S. R., Raju, R., Srikar, M., Shyam Sundar, S., Amudhan, S., Cristia, A., Pawar, R., Rao, A., Vasuki, P. P., Volme, S., & Mysore, A. (2025). Validation of the Language ENvironment Analysis (LENA) automated speech processing algorithm labels for adult and child segments in a sample of families from India. *Journal of Speech, Language, and Hearing Research : JSLHR*, *68*(1), 40–53. https://doi.org/10.1044/2024_JSLHR-24-00099

Miller, J. F., & Iglesias, A. (2012). *Systematic analysis of language transcripts (SALT)* (Version Research Version) [Computer software]. SALT Software, LLC.

Olswang, L. B., & Prelock, P. A. (2015). Bridging the gap between research and practice: Implementation science. *Journal of Speech, Language, and Hearing Research*, 58(6), S1818–S1826.

Ortiz, J., & Huang. (2024). *Automated Language Sample Analysis for Children with Developmental Language Disorder*. American Speech-Language-Hearing Association Convention, Seattle, WA.

Pavelko, S. L., Owens, R. E., Ireland, M., & Hahs-Vaughn, D. L. (2016). Use of language sample analysis by school-based SLPs: Results of a nationwide survey. *Language, Speech, and Hearing Services in Schools*, *47*(3), 246–258. https://doi.org/10.1044/2016_LSHSS-15-0044

Schneider, P., Dubé, R. V., & Hayward, D. (2005). *The Edmonton Narrative Norms Instrument*. University of Alberta Faculty of Rehabilitation Medicine website: www.rehabresearch.ualberta.ca/enni.

Seymour, H. N., Roeper, T. W., & de Villiers, J. (2005). DELV-Norm Referenced

SYSTRAN. (2025). faster-whisper (Version 1.2.1) [Computer software]. https://github.com/SYSTRAN/faster-whisper

Treffers-Daller, J., Parslow, P., & Williams, S. (2018). Back to basics: How measures of lexical diversity can help discriminate between CEFR Levels. *Applied Linguistics*, *39*(3), 302–327. https://doi.org/10.1093/applin/amw009

Vogel, A. P., & Morgan, A. T. (2009). Factors affecting the quality of sound recording for speech and voice analysis. *International Journal of Speech-Language Pathology*, *11*(6), 431–437. https://doi.org/10.3109/17549500902822189

Wilder, A., & Redmond, S. M. (2022). The reliability of short conversational language sample measures in children with and without developmental language disorder. *Journal of Speech, Language, and Hearing Research : JSLHR*, *65*(5), 1939–1955. https://doi.org/10.1044/2022_JSLHR-21-00628

Wilder, A., & Redmond, S. M. (2024). Updates on clinical language sampling practices: A survey of speech-language pathologists practicing in the United States. *Language, Speech, and Hearing Services in Schools*, *55*(4), 1151–1166. https://doi.org/10.1044/2024_LSHSS-24-00035

Winters, K. L., Jasso, J., Pustejovsky, J. E., & Byrd, C. T. (2022). Investigating narrative performance in children with developmental language disorder: A systematic review and meta-analysis. *Journal of Speech, Language, and Hearing Research*, *65*(10), 3908–3929.