12-2021

An Investigation Into the Feasibility of Streamlining Language Sample Analysis Through Computer-Automated Transcription and Scoring

Carly Fox
Utah State University

AN INVESTIGATION INTO THE FEASIBILITY OF STREAMLINING LANGUAGE SAMPLE ANALYSIS THROUGH COMPUTER-AUTOMATED TRANSCRIPTION AND SCORING

by

Carly Fox

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Disability Disciplines

Approved:

Sandra Gillam, Ph.D., CCC-SLP Major Professor	Ronald Gillam, Ph.D., CCC-SLP Committee Member
Lisa Milman, Ph.D., CCC-SLP Committee Member	Sarah Schwartz, Ph.D. Committee Member
Tyson Barrett, Ph.D. Committee Member	D. Richard Cutler, Ph.D. Interim Vice Provost for Graduate Studies

UTAH STATE UNIVERSITY
Logan, UT

2021

ABSTRACT

An Investigation into the Feasibility of Streamlining
Language Sample Analysis through Computer-
Automated Transcription and Scoring

by

Carly Fox
Utah State University, 2021

Major Professor: Sandra Gillam, PhD, CCC-SLP
Department: Special Education & Rehabilitation

The purpose of the study was to investigate the feasibility of streamlining the transcription and scoring portion of language sample analysis (LSA) through computer-automation. LSA is a gold-standard procedure for examining childrens' language abilities that is underutilized by speech language pathologists due to its time-consuming nature. To decrease the time associated with the process, the accuracy of transcripts produced automatically with Google Cloud Speech and the accuracy of scores generated by a hard-coded scoring function called the Literate Language Use in Narrative Analysis (LLUNA) were evaluated. A corpus of 255 narrative transcripts and 170 audio recordings of narrative samples were selected to evaluate the accuracy of these automated systems. Samples were previously elicited from school-age children between the ages of 6;0-11;11 who were either typically developing (TD), at-risk for language-related learning disabilities (AR), or had developmental language disorder (DLD). Transcription error of Google Cloud Speech transcripts was evaluated with a weighted word-error rate (WER_w). Score accuracy was evaluated with a quadratic weighted kappa (κ_{qw}). Results indicated an average WER_w of 48% across all 170 language sample recordings, with a median WER_w of 40%. Several recording characteristics were associated with transcription error including the codec used to recorded the audio sample, the presence of background noise. Transcription error was lower on average for samples collected using a lossless codec, and for those that did not contain background noise. Scoring accuracy of LLUNA was high across all six measures of literate language when generated from traditionally produced transcripts ($\kappa_{qw} = .66 - 1$), regardless of age or language ability (TD, DLD, AR). Adverbs were most variable in their score accuracy. Scoring accuracy dropped when LLUNA generated scores from transcripts produced by Google Cloud Speech ($\kappa_{qw} = .27 - .62$), however, LLUNA was more likely to generate accurate scores when transcripts had low to moderate levels of transcription error (up to about 50% WER_w). This work provides additional support for the use of automated transcription under the right recording conditions and automated scoring of literate language indices. It also provides preliminary support for streamlining the entire LSA process by automating both transcription and scoring, when high quality recordings of language samples were utilized.

(137 pages)

PUBLIC ABSTRACT

An Investigation into the Feasibility of Streamlining
Language Sample Analysis through Computer-
Automated Transcription and Scoring

Carly Fox

The purpose of the study was to investigate the feasibility of streamlining the transcription and scoring portion of language sample analysis (LSA) through computer-automation. LSA is a gold-standard procedure for examining childrens' language abilities that is underutilized by speech language pathologists due to its time-consuming nature. To decrease the time associated with the process, the accuracy of transcripts produced automatically with Google Cloud Speech and the accuracy of scores generated by a hard-coded scoring function called the Literate Language Use in Narrative Analysis (LLUNA) were evaluated. A collection of narrative transcripts and audio recordings of narrative samples were selected to evaluate the accuracy of these automated systems. Samples were previously elicited from school-age children between the ages of 6;0-11;11 who were either typically developing (TD), at-risk for language-related learning disabilities (AR), or had developmental language disorder (DLD). Transcription error of Google Cloud Speech transcripts was evaluated with a weighted word-error rate (WER_w). Score accuracy was evaluated with a quadratic weighted kappa (κ_{qw}). Results indicated an average WER_w of 48% across all language sample recordings, with a median WER_w of 40%. Several recording characteristics of samples were associated with transcription error including the codec used to recorded the audio sample and the presence of background noise. Transcription error was lower on average for samples collected using a lossless codec, that contained no background noise. Scoring accuracy of

LLUNA was high across all six measures of literate language when generated from traditionally produced transcripts, regardless of age or language ability (TD, DLD, AR). Adverbs were most variable in their score accuracy. Scoring accuracy dropped when LLUNA generated scores from transcripts produced by Google Cloud Speech, however, LLUNA was more likely to generate accurate scores when transcripts had low to moderate levels of transcription error. This work provides additional support for the use of automated transcription under the right recording conditions and automated scoring of literate language indices. It also provides preliminary support for streamlining the entire LSA process by automating both transcription and scoring, when high quality recordings of language samples are utilized.

DEDICATION

I would like to dedicate this work to my husband, Dr. Sharad Jones.

ACKNOWLEDGMENTS

I would like to thank my primary adviser, Dr. Sandi Gillam for her unwavering guidance and support through this process, as well as my entire committee for their valuable insights into this project. I would also like to thank my parents for encouraging me to pursue higher education and seek opportunities far from home. Finally, I'd like to thank my husband, Dr. Sharad Jones, who has been with me through all the ups and downs of my graduate career, and without whom I could not have finished with my sanity still (somewhat) intact.

CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

Language sample analysis (LSA) describes the practice of eliciting a representative sample of an individual's language, transcribing it to text, and then systematically analyzing indicators of typical and non-typical language usage (e.g., grammar, vocab, fluency). LSA can be utilized for a wide range of ages and for varying types of discourse (e.g., narrative, conversational, expository). This type of analysis is critical to conducting a comprehensive assessment in identifying children who at risk for language and literacy difficulties or who have developmental language disorders (AR/DLD; Evans, 1996; Pavelko et al., 2016). As compared to standardized, norm-referenced assessments, LSA can give a more nuanced account of a child's language skills, can be used to elicit language samples in authentic contexts (e.g., conversation or storytelling) and is less prone to cultural biases (Gutiérrez-Clellen & Simon-Cereijido, 2007; J. Heilmann et al., 2010; Laing & Kamhi, 2003). While standardized assessments provide strong construct validity, the inclusion of LSA adds ecological validity to a child's language profile (Botting, 2002). Clinician's utilizing LSA in conjunction with standardized assessments obtain higher quality information at the cost of increasing the quantity of time spent conducting assessment, however. It is the time-cost associated with conducting LSA that has made it difficult to bridge the research to practice gap for this important evidence-based practice.

Several large survey studies conducted over the past decades have confirmed that only a portion of SLPs utilize LSA as a part of their typical assessment protocol (Fulcher-Rood et al., 2018; Kemp & Klee, 1997; Pavelko et al., 2016; Westerveld & Claessen, 2014). Of the speech language pathologists (SLPs) who report not using LSA, the most cited barrier is that the practice is too time-consuming. Two of the most involved components of LSA are transcribing and then analyzing the language samples. It has been estimated that for each minute of audio about five minutes are needed to transcribe the language sample of a typically developing child (J. J. Heilmann, 2010; Miller et al., 2016). When the language sample is elicited from a child with impaired language abilities that estimate increases to seven to eight minutes per minute of audio, due to aspects such as disfluencies (e.g., false-starts, repetitions), unintelligible speech, and the use of ungrammatical utterances which makes the samples more difficult to accurately transcribe (Miller et al., 2016).

Language samples must also be long enough to be representative of a child's language abilities. By some estimates, as few as one to three minutes of audio is considered adequate for analysis (Tilstra & McMaster, 2007). More conservative estimates would recommend even longer samples (50-100 utterances), in order to get the best sense of a child's strengths and weaknesses since children's language use is inherently inconsistent and can vary across contexts, speakers, tasks, and environments (Bloom & Lahey, 1978; Evans & Craig, 1992; Miller & Iglesias, 2010). In the current project, a middle ground approach of including samples between 1-7 minutes was used, based on Heilmann et al.'s (2010) findings that estimates of certain language measures (productivity, lexical diversity, utterance length) were consistent across language samples that were one, three, and seven minutes in length. On top of the time spent transcribing language samples of variable lengths, manual analysis of transcripts adds more time to the process, with more complex analyses taking upwards of an hour to complete (Gabani et al., 2009). One can therefore imagine how clinicians might find conducting these procedures impractical given their busy schedules.

Simplification of LSA Procedures

Real-Time Transcription

There are, however, several means of reducing the time associated with both of these processes. In terms of transcription, survey data has revealed that close to half of SLPs who conduct LSA practice an expedited form of transcription, called real-time transcription (RTT; Fulcher-Rood et al., 2018; Kemp & Klee, 1997; Pavelko et al., 2016; Westerveld & Claessen, 2014). In RTT, the clinician (or a speech language technician) will transcribe the language sample as it is being elicited, instead of recording the sample and transcribing it at a later time (i.e., traditional transcription).

Some SLPs use RTT to reduce the time involved in LSA, however, the evidence for its efficacy is limited and mixed (C. B. Fox et al., 2021; J. J. Heilmann, 2010; Klee et al., 1991). Klee et al. (1991) examined the accuracy of RTT on 20 conversational, play-based language samples elicited from preschool-aged children with impaired language. Two SLPs with three years of transcription experience served as the transcribers in this study. The transcribers were first tasked with transcribing the conversational samples in real time and were then asked to use traditional transcription to produce the same recorded language samples 3.5 days later. The transcripts produced using traditional transcription were considered "gold-standard". The pairs of transcripts (RTT and traditional transcription) were then compared using correlation analyses to determine interrater reliability on total number of utterances (TNU), total number of words (NTW), number of different words (NDW), percent intelligibility and mean length of utterance (MLU). Results indicated strong interrater reliability across all five indices, $r(18) = .91 - .97, p < .005$.

While these results provide evidence for the efficacy of RTT for preschool-age, conversational language samples, a more recent study by Fox et al. (2021) did not find support for these findings for school-age children (7;5-11;9) whose language samples were elicited using narrative discourse sampling (e.g., narrative) as opposed to conversational sampling.

In his 2010 discussion on the Myths and Realities of Language Sample Analysis, Heilmann warned that SLPs practicing RTT outside of the research supported contexts should proceed with "high levels of caution" (p. 7), as it was unknown how accurate transcripts produced with RTT would be for older children displaying greater language productivity and complexity. Klee et al. (1991) similarly suggested that RTT would likely not be feasible with school-age children, stating that: "Presumably, there is a correlation between the child's level of language production and the difficulty of doing real-time transcription. The more advanced the child's language production, the more difficult will be the transcription" (p. 38). Survey data has indicated that RTT is being used with children outside the preschool age-range, however (Westerveld & Claessen, 2014). This is problematic given Fox et al.'s (2021) findings that transcription error for RTT by clinicians ranged between 11-83%, with transcription error increasing as a function of children's speech rate. Transcripts produced with RTT for samples produced by school-age children therefore have the potential to be invalid representations of their language abilities, which in turn defeats the purpose of collecting the sample in the first place. This mismatch of clinical recommendations and clinical practice highlighted the need to determine if there is a better alternative if RTT is not accurate in this context.

Computer-Assisted LSA Programs & Simplified Analysis Procedures

In addition to expediting the transcription process, efforts have also been made to streamline the coding and analysis processes involved in LSA. These solutions currently include the use of computer-assisted language sample analysis programs, as well as manual, but simplified analysis procedures. Currently used computerized systems will be discussed first, followed by summaries of simplified coding and analysis procedures.

Survey research has indicated that the Systematic Analysis of Language Transcripts (SALT; Miller & Iglesias, 2010) may be the most widely utilized computer-assisted program for LSA in the field of Speech Language Pathology (Fulcher-Rood et al., 2018), followed by the Computerized Language Analysis, which is also a commonly used for LSA (CLAN; MacWhinney, 2000). SALT requires that language samples be typed (or pasted) into a transcript file that is segmented into communication-units (C-Unit), which describe an independent clause or an independent clause and its dependent clause(s). Each C-Unit includes a speaker identifier at the beginning of each line (e.g., Child, Examiner, etc.). Several basic indices may be generated using only this basic formatting, such as total number of words (NTW), number of different words (NDW), total number of utterances (TNU), mean length of utterance in words (MLU-words) and type-token ratio (TTR). Additional analyses such as mean length of utterance in morphemes (MLU-m), percentage of grammatically correct utterances and percentage of syntactically complex utterances require the use of specific transcription conventions (i.e., mazing or marking of repetitions, false-starts, filler words, etc.), the removal of unintelligible utterances, and the use of morpheme segmentation rules (i.e., boys coded boy/s) to name a few.

SALT contains several proprietary language databases that clinicians may use to compare the results of their clients to other typically developing children/adolescents of the same age using a variety of sampling contexts, including play, conversational, narrative, expository and persuasive discourse. These databases are available for English and Spanish speaking children in some contexts (Pezold et al., 2020). SALT offers free training for clinicians who wish to use it, although the program itself must be purchased for a fee.

Another well-known, computer assisted language sample analysis program is CLAN, which unlike SALT, is available at no cost, allows the addition of user specific metrics for analysis, and does not require manual segmentation of the sample to obtain general language indices including MLU-m (Ratner & MacWhinney, 2016; Sagae et al., 2005). Like SALT, CLAN does require that the user code for repetitions, fillers, intelligible utterances, and grammatical errors. CLAN may be used with a total of 49 different languages, though all database comparisons are made to English (CHILDES), and include samples elicited from early childhood and preschool age children in play settings only (Pezold et al., 2020). Numerous tutorials exist on conducting specific types of analyses using CLAN (e.g., see, Pezold et al., 2020; Finestack et al., 2020).

In addition to computer assisted programs, simplified analysis procedures have been proposed to reduce the time involved in LSA. The Sampling Utterances and Grammatical Analysis Revised (SUGAR; Pavelko & Owens Jr, 2017) is one such procedure. SUGAR evaluates two measures of language productivity, including NTW and words per sentence (WPS), and two measures of syntactic complexity, including MLU-morphemes and clauses per sentence (CPS). SUGAR is intended for usage with samples that are 50 utterances long. Each of these measures may be obtained after the clinician has transcribed the sample using a basic set of segmentation rules (e.g., segment utterances at pauses in speech, do not include abandoned utterances; Casby, 2011). No coding is required, apart from inserting spaces between bound morphemes and contractions. MLU is calculated by summing the number of morphemes and dividing by 50. NTW is determined using the "word count" function in Microsoft word and WPS is calculated by dividing the NTW by the number of sentences. Finally, CPS is calculated by counting the number of clauses and then dividing by the number of sentences. SUGAR offers training videos and analysis tools free of charge, which can be found here, and also includes access to a database of typically developing children ages 3;0-7;11 in parent-child interactions. Both the computer-assisted LSA programs and SUGAR do still rely on transcription and some coding, which as previously stated, can be a time-consuming process that requires training. These factors serve as potential barriers to their implementation.

There is, however, one simplified coding procedure which does not rely on transcription or coding and is instead scored in real-time while listening to recorded language samples. This instrument, called the Grammaticality and Utterance Length Instrument (GLi; Castilla-Earls & Fulcher-Rood, 2018) can be used to produce two measures from a child's narrative retell, including grammaticality and utterance length. The grammaticality measure is determined by having the listener judge each utterance in a child's language sample as grammatical or not (based on their own knowledge of English grammar), and then calculating the percent of grammatical utterances (PGU). This measure was found to have adequate convergent validity with gold-standard PGU produced in SALT. The utterance length is determined by having the listener judge each utterance as fewer than three words, four to seven words, or more eight words and then sum the number of utterances in each length category (i.e., how many are $< 3, 4-7, 8+$). Utterance length was similarly found to have good convergent validity with MLU words calculated in SALT. The GLi has preliminary evidence for diagnostic sensitivity and specificity and therefore may be useful tool for clinicians to incorporate as a part of their LSA protocol, however, it can only be used to calculate two measures (utterance length and grammaticality) which may not meet all the needs of clinicians. Further, while these two measures (PGU, utterance length) were reliable with their respective measures produced in SALT (PGU, MLU) using C-Unit segmentation, their reliability with other segmentation conventions (e.g., DSS) is currently unknown.

Expedited transcription (i.e., RTT) and analysis may cut down on the time needed to conduct LSA as compared to utilizing traditional transcription or conducting analyses by hand, however, these time savings have not proven to be adequate as evidenced by the continued survey findings on the underutilization of LSA by SLPs (see e.g., Fulcher-Rood et al., 2018; Pavelko et al., 2016). If LSA is to be utilized in assessment and progress monitoring additional steps are needed to increase its utility.

As explained above, the first bottleneck in the LSA process is transcription, which up to this point has primarily been addressed through the use of RTT with preschool populations. The second bottleneck is coding and/or analysis which has been addressed through the use of computer-assisted and abbreviated manual analysis procedures. The purpose of this project was to address both issues by exploring the use of automated transcription and machine learning for the analysis of narrative language samples obtained from school age children. In the next sections an overview of machine learning is given, followed by an explanation of automatic speech recognition (ASR) and natural language processing (NLP) technology, which the current study investigated as a means of automating transcription and analysis. This review of the literature will conclude with a discussion of relevant clinical applications of these machine learning techniques and how the current project aimed to build on this body of work.

CHAPTER 2

Literature Review

Machine learning, sometimes referred to as statistical learning, refers to a class of algorithms used to form predictive models based on previous observations (Chollet, 2017). Linear regression, for example, is a type of machine learning model, albeit a simple one that is designed to handle linear, normally-distributed outcomes (e.g., you could predict someone's height based on their weight by plugging in their weight value to a regression equation computed from a number of height and weight observations). Regression is one of the simplest forms of machine learning, but there are many more options designed to make predictive inferences on more complex data types, like audio and text.

One aspect that is common to all machine learning methods is the requirement that data be used to train a model (Chollet, 2017). Training data are used to construct a model, based on "learned" parameters, such that predicted observations can be generated with some level of accuracy. Again, returning to the regression example of predicting height from weight, the slope of the regression line is a learned parameter, which tells us that for every one unit increase in weight, there is an associated x increase/decrease in height. Also common to machine-learning methods is the issue of over-fitting, which occurs when the model has learned "noise" present in the data (i.e., idiosyncrasies specific to the given dataset) that is not reflective of the actual signal of interest (i.e., the true pattern). Models that are overfit will have high predictive accuracy on the dataset it is trained on, but may have poor predictive accuracy on unseen data. For example, a model predicting height from weight could have an accuracy of 95% on the data that was used to train the model, but only have a predictive accuracy of 64% when applied to new observations of weight (not in the training data). This stark difference in predictive accuracy between training and unseen "test" data, would be indicative that the model was overfit.

Several methods may be employed to prevent overfitting (Chollet, 2017). Ideally, datasets should be split into training, validation, and test sets, however, this is only feasible with large datasets. The *training set* is used to construct the model, the *validation set* is used to tune the hyperparameters of the model in an attempt to improve accuracy, and the test set is used to determine the accuracy of the model on unseen data. The test set should only be utilized once the final parameters have been determined, otherwise overfitting can occur. When datasets are smaller, cross-validation serves as another viable alternative to help prevent overfitting. There are different types of cross-validation (e.g., leave-one-out, k-fold, etc.), but generally they are resampling methods that involve splitting a dataset into k groups and then iteratively training the model on $k-1$ groups, each time leaving out one of the k groups as a hold-out test set. This sampling is done without replacement. The average of each model's predictive accuracy on each of the k test sets is used to determine the overall model accuracy.

Beyond these basic commonalities, there are many machine learning algorithms that vary widely in their underlying components. Relevant to the issue of streamlining transcription is a subfield of machine learning known as automatic speech recognition (ASR), which can perform automatic speech-to-text conversion. Relevant to the issue of streamlining analysis is natural language processing (NLP). NLP is a subfield of computational linguistics focused on computerized text analysis. NLP often utilizes machine learning for tasks such as part-of-speech (POS) tagging, next word prediction, and syntactic parsing, though this list is by no means exhaustive. It should also be noted that certain NLP processes can be conducted without machine learning, using hard-coded rulesets designed by domain experts (e.g., linguists). Of the NLP varieties that do utilize machine learning, many of the underlying mechanisms are also common to ASR. Generally speaking, however, ASR is a more complex operation in that it deals in two different modalities (i.e., audio and text), while NLP deals only in text, and thus ASR requires some additional mechanisms. Machine learning has not yet become prevalent in speech language pathology research, so it would be helpful to 'lift' the proverbial hood and 'see how the engine works' before getting into a discussion of ASR and NLP systems.

Neural Networks

Most modern ASR and NLP systems are constructed by combining several types of neural networks. Neural networks are a general form of machine learning algorithms that utilize interconnected artificial "neurons" to learn underlying patterns in data that then output a decision (numerical value, classification, recognition, etc.). They are referred to as a 'general form' because they are capable of approximating any function. A basic neural network takes input from x variables, learns their parameters (weights and bias) through one or more hidden layers, and then determines the output. The intermediate layers are called "hidden" because the values within these layers are not directly observable the same way input or output values are.

The basic building block of the neural network is called a perceptron or an artificial neuron (Nielsen, 2015; Rosenblatt, 1958). There are several components to the neuron including the inputs x, their associated decision weights ω, and a bias value b. The weight ω, determines the relative importance of an input (e.g., when classifying a child as DLD, how important is their mean length of utterance vs their lexical diversity), while the bias pushes toward a particular decision (e.g., TD vs

DLD); these are the learned parameters. A basic neuron can output either a zero or one, with a higher bias making the output more likely to "fire" or be one. The neuron generates the output by summing the product of the input values x and their weights ω with the bias b, given by the Equation 2.1 below:

$$z = \ \omega x \ + \ b \tag{2.1}$$

When there is only a single input, and the Equation 2.2:

$$z = \sum_j \omega_j x_j + b \tag{2.2}$$

when there are multiple inputs. This equation should look familiar, as it is a more general form of simple regression (i.e., weight is equivalent to slope, bias to the intercept). The neuron algorithm in isolation is not well-suited to learning however, since even small adjustments made to weights or the bias are cascading and can lead to large changes in the output.

Activation Functions

Activation functions are applied to the basic neuron algorithm to make outputs more suitable to specific types of problems, the same way link functions are utilized in generalized linear models (GLMs) to transform underlying distributions (Nielsen, 2015). Figure 2.1 depicts a basic single neuron with three inputs and the associated weights and bias of each connection. The activation function is applied to produce the activated output value a.

The sigmoid activation function is used in binary classification problems, where output values range between 0 to 1. The sigmoid function is defined in the follow-

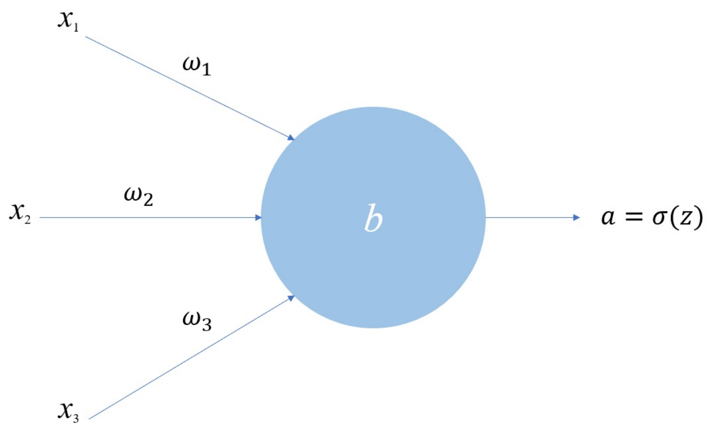


Figure 2.1: Structure of single artificial neuron x represents input variables, ω represents variable weights, b represents bias, $a = \sigma(z)$ represents the output.

ing Equation 2.3:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{2.3}$$

Which when applied to the output z, results in the following transformation defined in Equation 2.4:

$$\frac{1}{1 + \exp\left(-\sum_j \omega_j x_j - b\right)} \tag{2.4}$$

The sigmoid activation function is equivalent to the sigmoid link function used in logistic regression. Logistic regression can therefore be thought of a specific variant of the sigmoid neuron, or similarly as a very basic neural network. An additional purpose of applying an activation function is that it allows for subtle changes to weights and bias that similarly result in subtle changes to the output. This makes for a better learning algorithm than a basic neuron. As in logistic regression, a sigmoid activated output value will fall between zero and one, see Figure 2.2. The determination of the output as either a zero or one is decided by a threshold value which defaults to .5, but can be increased or decreased to bias the output more towards either decision.

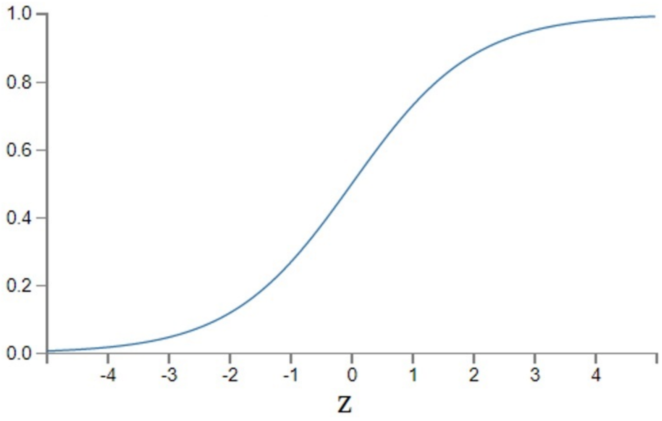


Figure 2.2: Sigmoid function where Z represents the z-score, source: here

Of note, there are a variety of activation functions that can be applied depending on the task. Softmax, for example, is another activation function commonly utilized for multi-class classification problems, where the output is instead defined in terms of a probability distribution. Softmax is commonly used in character-mapping (i.e., matching speech to a letter/grapheme) for speech recognition.

Feedforward Neural Networks

Feedforward Neural Networks (FFNN) are composed of multiple layers of connected neurons, including an input layer, hidden layers, and an output layer (Nielsen, 2015; Sanger, 1989). They are called “feed forward” neural networks because communication between neurons occurs in only one direction, from the input layer to the hidden layer(s) to the output layer. The number of nodes in the input layer corresponds to the number of inputs (e.g., all the words in a text, all the pixels in an image, all the 10 ms clips in an audio file). All inputs must be stored within a tensor (e.g., a 1-D tensor is a vector, 2-D is a matrix, etc.) of n-dimensions, which is accomplished through *feature extraction* (Chollet, 2017). The

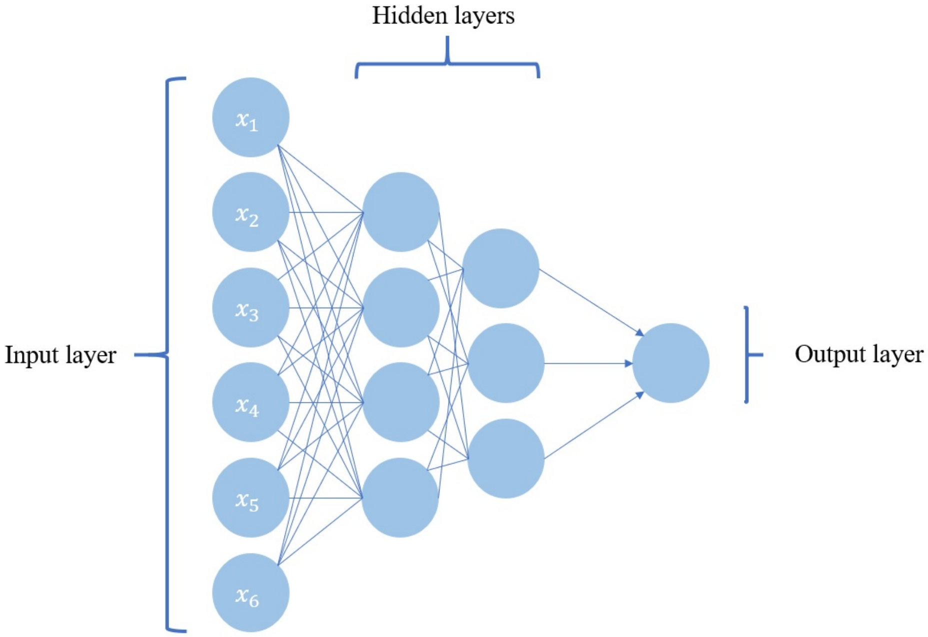


Figure 2.3: Basic feedforward neural network architecture where x represent the input variables.

method of feature extraction varies depending on the raw form of the data (e.g., image, text, audio, etc.), but all act to convert raw data to useable representations without sacrificing key information. The hidden layers contain one or more layers of connected neurons. Information is passed between layers in succession, updating parameters through each layer. The number of hidden layers and neurons within each layer are predefined by the user, which are called *hyperparameters*, but a neural network with multiple hidden layers is called a deep neural network (DNN). The output layer can be made up of any number of neurons, depending on the task (e.g., one for binary classification, as many as there are classes for multi-class classification, etc.). This layer is where the final value(s) of the model are determined. The following schematic represents a four-layer FFNN with two hidden layers, where each connecting line represents a distinct weight, see Figure 2.3.

A basic FFNN contains inputs, a hidden layer(s), and outputs, and it learns the parameters of the network to minimize the loss (i.e., error) of the output. Learning occurs in a “trial-and-error” type fashion through a process called *backpropagation* where an *optimizer* is used to determine the direction and magnitude of changes to the parameters (i.e., weights and biases), such that the loss function (i.e., error) of the model is minimized.

Loss Functions

The goal of any predictive model is to minimize the error between the predicted and observed values/classes, in a neural network the model is error is represented by the loss function (Nielsen, 2015). Different kinds of loss functions can be used in neural networks, depending on the task at hand, but typically all loss functions share two common properties: 1) the loss value is positive, and 2) the smaller the loss function, the better the model performance. A well-known loss function is mean-squared error (MSE), which is used in OLS regression. The MSE is given by Equation 2.5:

$$MSE = \frac{(y - a)^2}{2} \tag{2.5}$$

Where y, represents the observed outcome and a represents the predicted outcome. The lower the MSE, the closer the predicted values are to the observed values.

Backpropagation & Optimizers

Backpropagation is the means through which weights ω_k and biases b_k are updated after each time the loss function is calculated (Rumelhart et al., 1986). Beginning from the first iteration of the model the following steps take place: 1) the weights and biases (i.e., parameters) are set for each neuron in the network, typically through random initialization, 2) a forward pass through each layer is completed, and 3) the error (e.g., MSE) of the observed output is computed. The error from the output layer is then propagated through the layers of neural network, beginning with the final layer and then moving backwards. Parameter adjustments made to each neuron within a layer are impacted by their connections to neurons from the previous layer.

The magnitude (i.e., how much) and the direction (i.e., positive, negative) of the parameter adjustments is determined through the optimizer. Stochastic Gradient Descent (SGD) is a commonly used optimizer that takes a step-wise approach to move the loss function towards a minimum (Bottou & others, 1991; Nielsen, 2015). SGD works by repeatedly sampling inputs (known as mini-batches) at random from the larger training set, training the model (i.e., initializing parameters, computing the loss function, backpropagation), and then re-sampling additional mini-batches until all training inputs have been utilized – referred to as one epoch of training (i.e., like k-fold cross-validation). Training can then continue over a predetermined number of epochs, continuing the process of adjusting parameters until it appears the loss function has been minimized. A simple way of thinking about SGD is as a means of communicating to the model that the given parameter adjustments led to more or less error. SGD is commonly visualized as a ball being pushed in a hyperdimensional surface representative of the loss function. The ball can be pushed in different directions and for greater or shorter distances, with the end goal of reaching the lowest point in the loss surface (i.e., the global minimum). If the distance traveled by the ball through each step is too small, it can settle at a local minimum (i.e., a low loss, but not the lowest loss). If the distance traveled by the ball is too large, it might skip over the global minimum, see Figure 2.4. Thus, SGD is a step-wise training approach to minimizing loss, where setting an effective

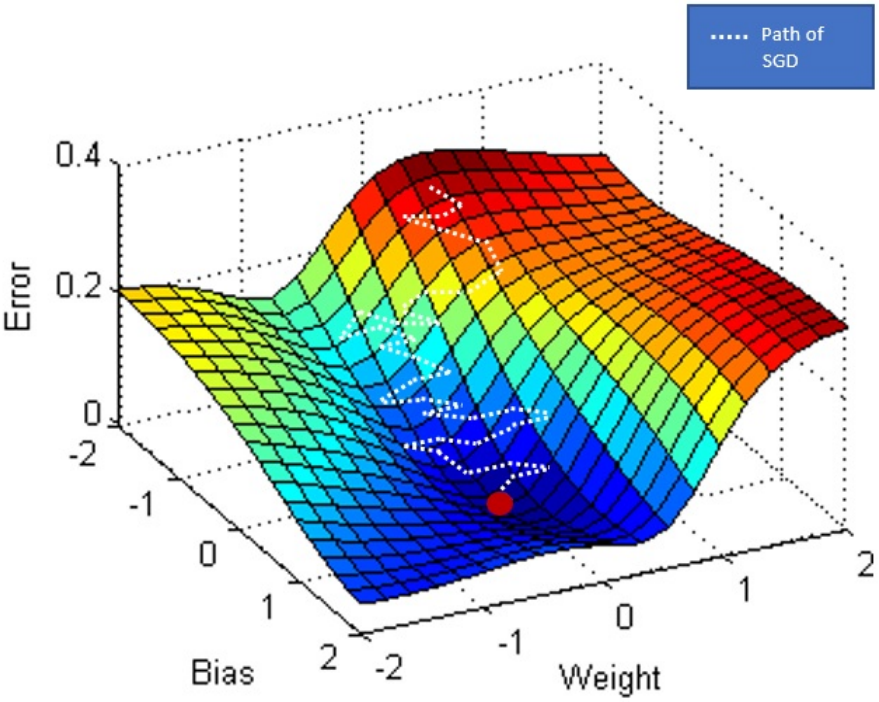


Figure 2.4: A hypothetical path for stochatisc gradient descent within the loss function.

distance for each step, or the learning rate, is critical. The learning rate is another type of hyperparameter that must be set and tuned by the user.

Deep Neural Networks (DNN)

Thus far, the shallow FFNN has been described, where there is a forward pass through the model, which then updates via backpropagation. However, there are many types of neural networks that build off this basic architecture. An extension of FFNNs are called "deep" neural networks (DNN), also known as "deep learning". Deep learning describes neural networks for which there are multiple hidden layers;

a feedforward neural network with 3 or more layers would be considered a DNN, for example (Nielsen, 2015). Increasing the number of hidden layers allows the model to handle more complex functions without adding more neurons to layers. This concept is described as adding “depth” to the model, instead of “width”, and is used to reduce computational demands. DNN models are utilized for complex data analysis tasks, such as speech recognition. In the following sections, the DNNs commonly utilized by ASR systems and some NLP tasks will be explained.

Recurrent Neural Networks (RNN)

Recurrent neural network (RNN) is a variety of DNNs designed to handle sequential data (Goodfellow et al., 2016; Rumelhart et al., 1986). Speech, for example, is a type of sequential data that would be difficult to process using a basic feedforward network. Feedforward networks handle data that are static or sampled at one time-point, while RNNs are designed to handle dynamic data that are time-dependent. RNNs are constructed such that information learned from prior time-points or “states” can inform the current state, making it a useful architecture for speech- and language-related tasks.

The general structure of an RNN is depicted in Figure 2.5 and can be thought of as a collection of several FFNNs. FFNNs are trained on data with a single time-point and learn parameters (i.e., weights and biases) from the input layer towards the output layer. RNNs, on the other hand, are trained on data with multiple time-points and learn parameters in two directions: forward towards the output at each time-step and horizontally (left to right) across each time-step, such that the prior state can inform the current state. Here, h represents the hidden state at each time point t. At each time step h receives information from two sources: the prior hidden state and the current input x. Weights are passed through these connections,

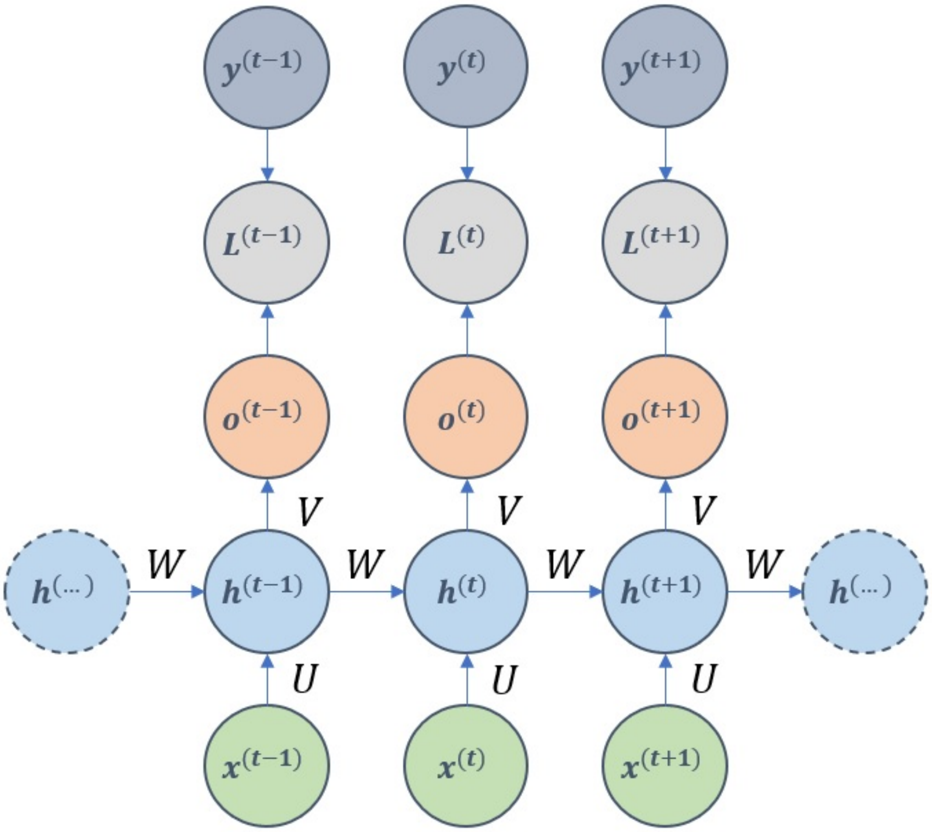


Figure 2.5: Basic RNN Structure, see description in text below for interpretation of the figure and labels.

where W represents weights of the hidden-to-hidden layer connections and U for weights of the input-to-hidden connections. At each time step, the hidden state also produces an output o through V, which represents weights of the hidden-to-output connections. The output value is then activated (e.g., with softmax) to the appropriate form, so that it can be compared to the observed output value y of the corresponding time-step. The loss function L is then calculated between the observed and predicted output values for each time-step.

The RNN updates through forward propagation of each time-step, such that each state is sequentially computed upon the completion of the prior state. Updates are completed at each time-step based on the following Equations 2.6:

$$\begin{aligned}
a^{(t)} &= b + Wh^{(t-1)} + Ux\left(t\right) \\
h^{(t)} &= tanh\left(a^{(t)}\right) \\
o^{(t)} &= c + Vh^{(t)} \\
\hat{y}^{(t)} &= softmax\left(o^{(t)}\right)
\end{aligned} \tag{2.6}$$

Where a at time t is represented by a linear equation summing the bias b, the product of the prior hidden state $h^{(t-1)}$ and weights W (hidden-to-hidden connection), and the product of current input value x and weights U (input-to-hidden connection). The current hidden state h is defined as a activated through hyperbolic tangent ($tanh$). The output o of the current time step is defined by the sum of the bias c and the hidden state h multiplied by the weights V (hidden-to-output connection). Finally, the predicted observation $\hat{y}$ is the output o with a softmax transformation, which allows the output to be evaluated on a normal distribution.

RNNs can be utilized for a number of speech- and language-related tasks, such as part-of-speech tagging. When predicting the POS of the current word ($x^{(t)}$), where t represents the current time-step, the model has access to the previously labeled word(s) through the hidden representation of the prior time-step ($h^{(t-1)}$), which serves as an additional input to the hidden representation of the current time-step ($h^{(t)}$) to determine the most likely POS tag ($o^{(t)}$).

Bidirectional RNNs

While a basic RNN only runs in one direction, taking past information to make future decisions, it can also be valuable to use information from an entire input sequence (i.e., both "past" and "future") to decide about the current time-step (Goodfellow et al., 2016; Schuster & Paliwal, 1997). In ASR for example, one can imagine that accurate speech-to-text mapping can benefit from knowledge of the whole speech sequence, both at the phoneme and word-level. When considering speech-to-phoneme mapping, co-articulation impacts the way speech sounds are produced, such that the pronunciation of the current phoneme is impacted by adjacent (prior or following) phonemes (e.g., the pronunciation of /p/ within “cu*p*” is different from within “su*p*er”). At the word-level, the likelihood of a particular string of characters can be dependent on other words (prior and following) in the sentence (e.g., “the bird flew away”, is more likely than “the *Bert flew* away” or “the *Bird-flu* away”). Therefore, a more useful RNN architecture for ASR is a bidirectional RNN that utilizes two separate sub-RNNs. One that runs forward (past-to-future) and one that runs backward (future-to-past). These separate RNNs each pass in their respective directions and do not overlap (i.e., there are no connections between these sub-RNNs). Each then independently provide information to the output. This architecture is visualized in Figure 2.6, where $h(t)$ represents the sub-RNN running forward in time, and $g(t)$ represents the sub-RNN running backward in time.

Long Short-Term Memory (LSTM) Networks

While it is clear that having a network that can store information connecting current states with future and past states is quite useful for problems such as speech recognition, that storage does not come for free, and it is not unlimited in capacity. The connection between the current state and a state from the “distant” past or future is called a long-term dependency. RNNs are prone to a phenomenon known as the vanishing/exploding gradient that results from long-term dependencies. Essentially, the weights of connections for long-term dependencies will either become exponentially large (explode) or small (vanish) as the span between connected time-steps increases. In either case, very large or very small weights

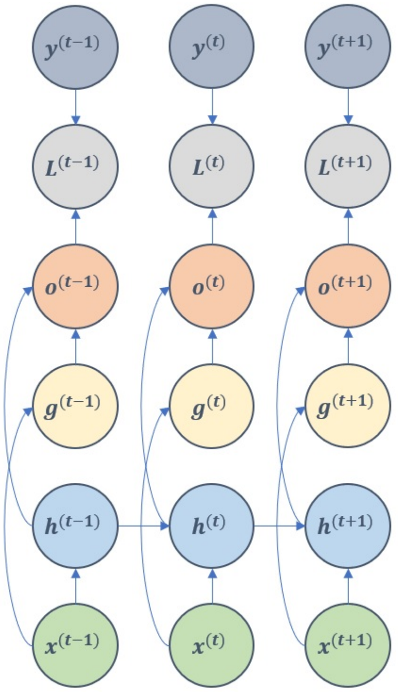


Figure 2.6: Bidirectional RNN architecture (computational graph).

severely hinders the success of SGD which in turn halts learning (Bengio et al., 1994; Hochreiter, 1991).

A solution to this issue comes from the unique architecture of the long short-term memory (LSTM) network. The LSTM updates what is held in memory at each time-step, allowing the model to utilize or “remember” information over a larger number of time-steps by “forgetting” information that is no longer relevant (Hochreiter & Schmidhuber, 1997). There are several pieces added to the architecture of an RNN that allow for this updating. One critical piece is called the cell-state, which can be thought of the memory storage available to the current time step. At each time step, the cell-state (or memory) can be updated, this updating is controlled through a series of gates. Each gate contains a sigmoid layer, which scales the output to between 0 and 1, where 0 means no information is let through and 1 means all information is let through. Values between can be thought of as a gate partially open, with larger values corresponding to a gate that is open more

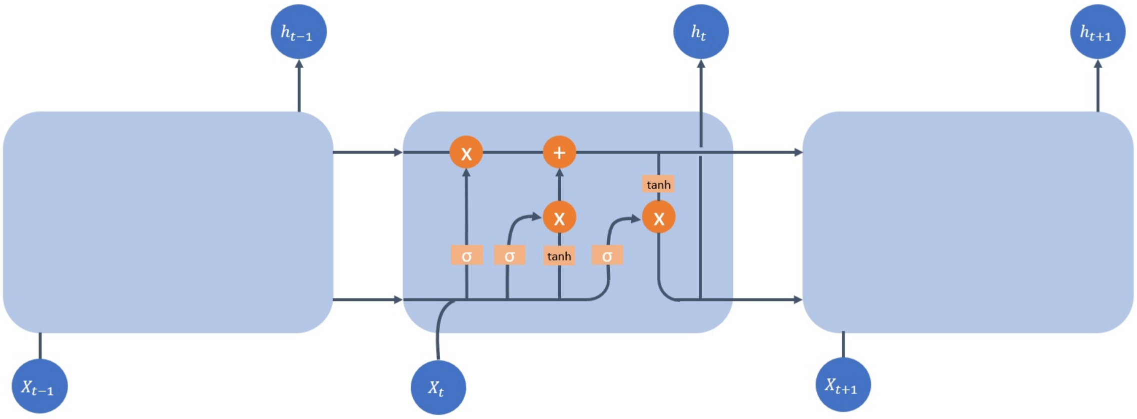


Figure 2.7: Long short-term memory (LSTM) architecture.

"widely", allowing more information through. The LSTM has three such gates, one which determines how much information to forget from the previous state, one which determines how much information will be added to the cell-state (memory) from the current time-step, and one which determines how much information will be passed on to the next time-step, see Figure 2.7.

These gate systems allow the cell-state (memory) to constantly update by maintaining relevant information and forgetting information that is no longer useful to the model. By making the memory system more efficient the model can store necessary information across longer-term dependencies. This can be important to both ASR and NLP, as predicting the most likely word might be reliant on words from prior sentences, not just words from the current sentence (e.g.,*The girl had been a vegetarian for years. Now, the sight of meat/feet/mead made her stomach turn*).

The Attention Mechanism

Common to both ASR and a number of NLP tasks is the encoder-decoder architecture, which is composed of RNNs. The encoder RNN takes as input some type of sequence (e.g., a text string or audio features) and outputs a hidden state, which is then fed into the decoder. The decoder RNN takes this hidden state and decodes it to the appropriate format (e.g., a translated text string or transcription) before outputting the final product. This basic architecture bottlenecks information between the encoder and decoder, which can be problematic for tasks that require access to greater context (e.g., dealing with homonyms or semantic ambiguity), however. The attention mechanism removes this bottleneck, mimicking the biological process it is named after (Bahdanau et al., 2014). The attention mechanism allows the decoder to access information from *all* hidden states at each time-step, and through training it learns to "focus" on (i.e., increase the weight of) hidden-states from the encoder that are relevant to the current time-step. One can imagine that having access to all hidden states can be useful for "local" disambiguation (e.g., coarticulation, determining word boundaries, etc.) as well as more "global" disambiguation (e.g., distinguishing homonyms, semantic ambiguity, etc.) in performing speech-to-text conversion.

Automatic Speech Recognition (ASR)

This next section describes ASR as it diverges from NLP, since again, ASR has the added complexity of dealing with different modalities of input (audio) and output (text). A large component of the added complexity comes from having to extract numerical features representative of the raw data, which are a series of audio signals. Audio preprocessing and feature extraction are important early steps in the process of performing speech-to-text conversion.

Data Preparation

Audio Preprocessing

It is a given that not all audio is of equal quality. When recorded in a natural setting, such as a classroom or clinic, recordings are subject to certain level of background noise that can vary quite widely. In addition, the audio quality of a language sample can be impacted by additional variables like the type of recording device used, the speakers, volume, the speaker's distance from the mic, and whether one or more speakers were captured. Audio preprocessing helps to remediate some of the audio degradation caused by these real-world complications. When more than one speaker is captured in a recording (e.g., a child and an examiner/clinician) the transcription accuracy can be negatively impacted. If the audio is recorded in a stereo format with two channels, one for each speaker, the audio channels can be split into two mono format audio files and transcribed separately to improve transcription accuracy. The sampling rate of the recording represents the captured frequency range, with most audio files being recorded at 16, 32, 44.1 or 48 kHz. The sampling rate is critical to speech intelligibility, and transcription accuracy can be negatively impacted when the sampling rate falls below 16 kHz. The bit depth of the recording is another important piece in speech intelligibility, relating to the range of volume and the signal-to-noise (i.e., distinguishing speech from background noise) ratio of the audio. The minimum recommended bit depth for speech recognition of an audio file is 16 bits. Both the sampling rate and bit depth of an audio file can be converted to fall within these optimal ranges, to hopefully increase transcription accuracy.

Feature Extraction

Common to both classic ASR models and more modern DNN-based models is the process of feature extraction, which is necessary to extract information from audio signals which is then stored in a tensor. Two common forms of acoustic feature representations are Mel filter banks and Mel-frequency cepstral coefficients (MFCC), which are used to generate a unique set of features for each segment or window of audio (Jurafsky & Martin, 2020). Feature extraction from audio utilizes a sliding window to segment an audio sample into smaller units. A typical window size is about 25 milliseconds (ms), sampled about 10ms apart. By setting a sliding window, the dynamic changes in pronunciation of phonemes (due to co-articulation) can be captured within overlapping window segments (of 25ms); see Figure 2.8.

Features can be extracted from each window through several methods, including Mel filter banks or MFCC. A high-level explanation is that for each window of audio there is an associated acoustic signal, represented by a sine wave of frequency (Hz) over time. For each discrete signal, a *power spectrum* can be calculated to determine which frequencies present across the frame. However, not all frequencies are equally important; humans are more attuned to lower frequencies, which indicate different articulatory features (e.g., voicing) important to perceiving/recognizing speech. In order to place greater focus on these lower frequency ranges, a filter is applied on a Mel scale, which places a fine-grained filter on frequencies closer to 0 that becomes broader at higher level frequencies, see Figure 2.9. The changes in filter-width are used such that even small variations in power at lower frequency ranges can be captured, while only large variations in power are captured at the less informative, high frequency ranges.

The values captured within each filterbank are then summed to evaluate the level of energy present within different frequency ranges (e.g., vowels are produced with frequencies primarily between 1-2 kHz). For certain models, this information

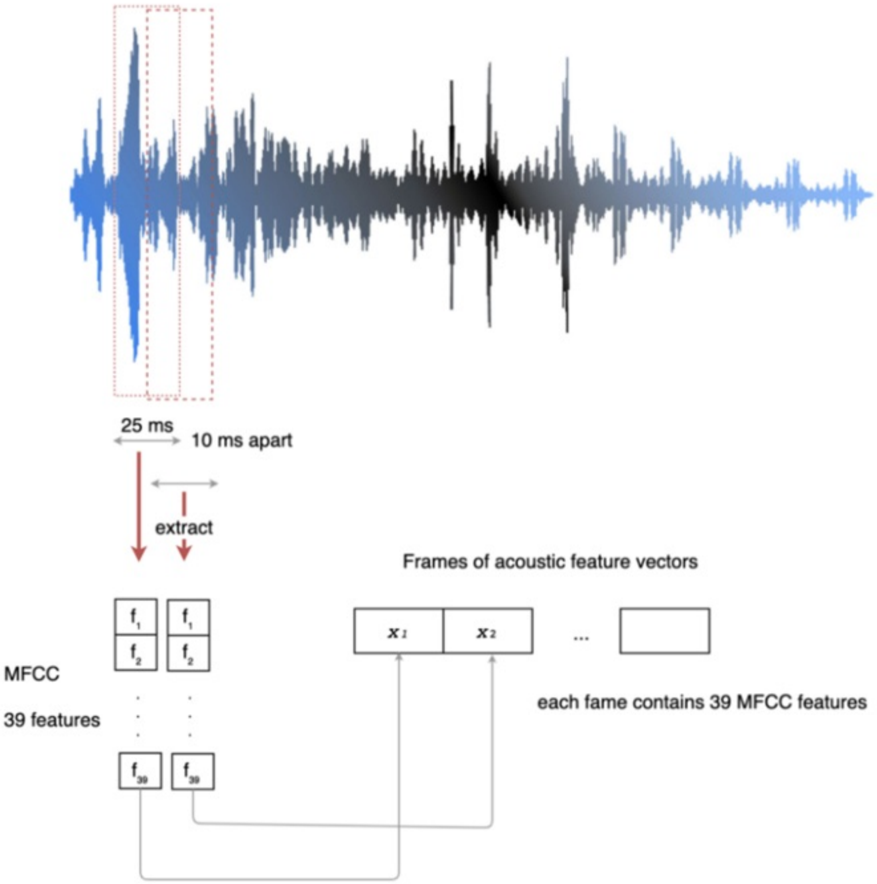

25 ms
10 ms apart
extract
Frames of acoustic feature vectors
f_1
f_2
f_{39}
MFCC
39 features
x_1
x_2
...
each fame contains 39 MFCC features

Figure 2.8: Acoustic feature extraction. Source: https://jonathan-hui.medium.com/.

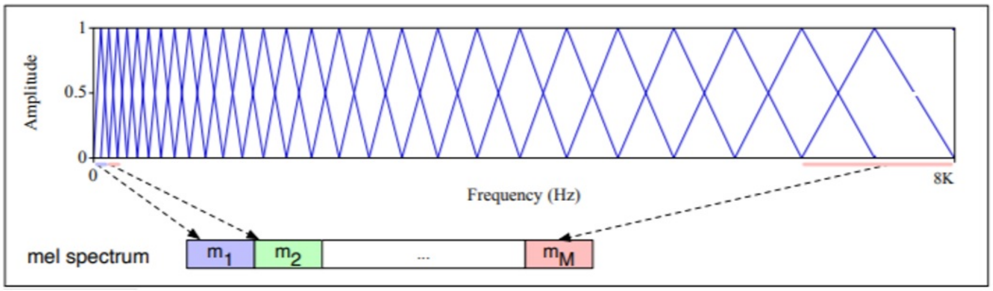


Figure 2.9: Mel filter bank. Source: https://web.stanford.edu/ jurafsky/slp3/26.pdf.

from the Mel filter bank is sufficient to use as the feature set. More classic ASR models require further compression however and thus utilize MFCC. MFCC requires several additional steps, including applying an inverse discrete Fourier transform (IDFT) to decorrelate the features so that they meet the underlying assumptions of the acoustic model. The end result is a series of feature sets representative of each window of audio from the sample, that is utilized by the model to decode the most likely phoneme/character representation.

ASR Architecture

Most state-of the-art ASR models utilize some combination of DNNs. However, ASR existed prior to the widespread adoption of DNNs, and it is easier to understand modern ASR systems by first describing pre-DNN ASR systems (Schalkwyk, n.d.). This is because classic models are more linguistically driven in their design, whereas modern models are more computationally driven. For example, while classic ASR models first map audio features to phonemes, and then utilize compiled pronunciation tables to determine the most likely word, modern ASR models are end-to-end, meaning they can directly map audio features to character-level text representations. It may be less obvious how such models can accomplish this direct mapping, so the architecture of classic ASR models will briefly be discussed before diving into what is considered state-of-the-art in speech recognition.

Classic ASR Design

Classic ASR systems consist of three independently trained models: 1) an acoustic model, 2) a pronunciation lexicon model, and 3) a language model, which all feed into a decoder that outputs the corresponding text (Hui, n.d.; Schalkwyk, n.d.). The acoustic model takes as input a series of acoustic features extracted from short segments of audio (~25 milliseconds) and uses a probability-based model to map features to the most likely phonemes , given the known features of distinct phonemes (i.e., frequency range). The pronunciation lexicon provides input about the likelihood of particular phoneme combinations given a known lexicon (essentially a word bank) of word-pronunciations. The language model is separately pretrained on a large corpus of texts to learn the joint probabilities of different word combinations. With this information it can provide input on the probability of a predicted word given one (bigram) or more (n-gram) previous words. Essentially, the acoustic and pronunciation model find the most probable mapping of the observed audio to speech sounds and individual words, while the language model provides information towards likely (2+) word combinations based on some basic learned grammatical knowledge. The decoder then takes input from all three models and searches for the most likely sequence of words, see Figure

2.10.

Using such an approach, models are optimized to take as input audio signal and find the most likely sequence of words, based on the combined outputs of the acoustic, pronunciation and language models. The output text sequence forms the hypothesis transcript, which is compared to an accompanying reference (i.e., ground-truth) transcript. The accuracy of the reference transcript(s) is determined using word error rate (WER), a metric of word misclassification. The WER is calculated as the total number of word insertions (i.e., adding a word that is not present in the reference transcript), word substitutions (i.e., substituting one word for another), and deletions (i.e., failing to transcribe a word in the reference transcript) and then dividing by the total number of words in the reference transcript. WER is the most common evaluation metric across all ASR systems (Park et al., 2008), and typically the WER on both the training and test corpus are presented to determine how accurate a model is.

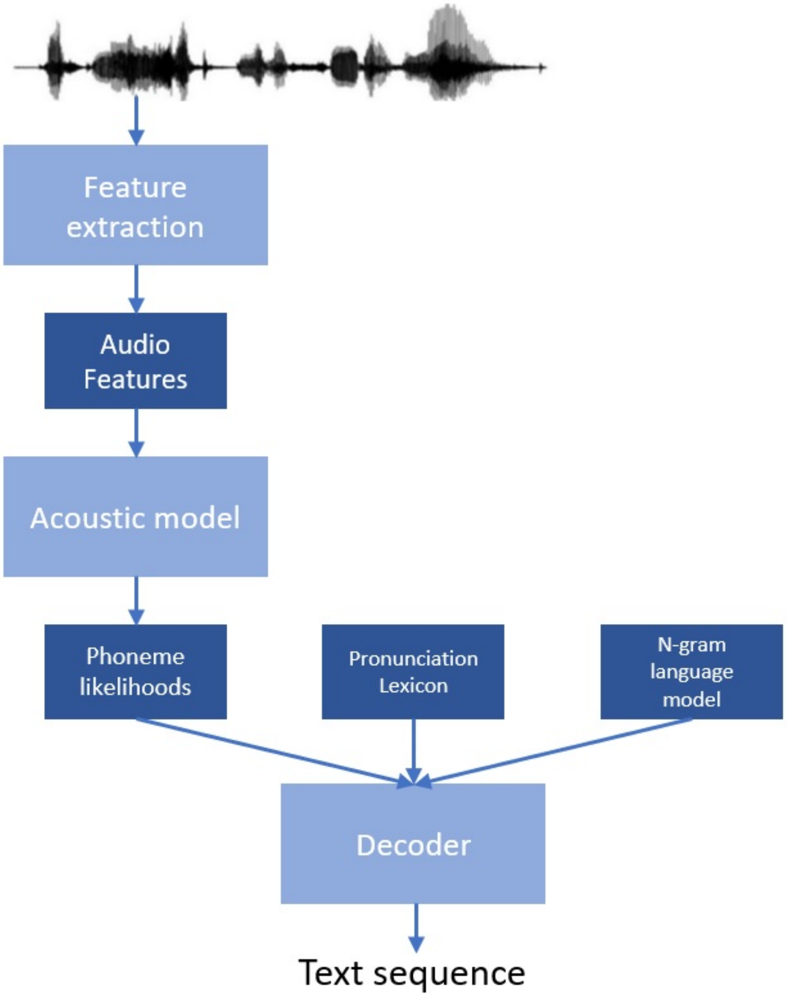


Figure 2.10: Classic ASR system architecture with acoustic, pronunciation, and language model.

State-of-the-Art ASR Design

Listen, Attend, and Spell (LAS) is a state-of-the-art ASR model that can be used for long-form audio transcription (Chan et al., 2015). LAS utilizes an encoder-decoder architecture comprised of a listener, which encodes the audio data to a higher-level representation, and a speller which decodes the representation into a string of characters. The encoder takes as input mel filter bank features and processes them through three layers of pyramidal bidirectional RNNs. Pyramidal RNNs are utilized to encode many time-steps (i.e., hundreds of audio frames) from the input sequence into a vector h of higher-level dense representations. This vector h is fed into the decoder, which is composed of a LSTM with an attention mechanism (remember attention can learn through training to attend to relevant inputs). The current decoder state (cell-state) s_i is computed from the previous decoder state s_{i-1}, the previous output character y_{i-1}, and the previous context c_{i-1}. The attention score is computed based on the current decoder state s_i and the encoder state h (inputs), which produces the current context c_i. The context determines which components of h the model should attend to. The character probability distribution (i.e., the probability that the output is any given character) is then computed with a FFNN with inputs from both the decoder state s_i and the context c_i, see Figure 2.11. Beam-search decoding is applied to find the most likely character sequence, using a maximum likelihood loss function, with the most likely character sequence having the highest score (i.e., probability) and the lowest loss. A final adjustment is made to the score by introducing the language model, trained on a large corpus of English texts to up-weight character sequences that are linguistically/grammatically probable (e.g., *That's their house on the corner* vs *that's there house on the corner*).

The LAS model underlies the long-form audio implementation of Google Cloud Speech. Long-form audio refers to audio clips that exceed 1 minute in length, which

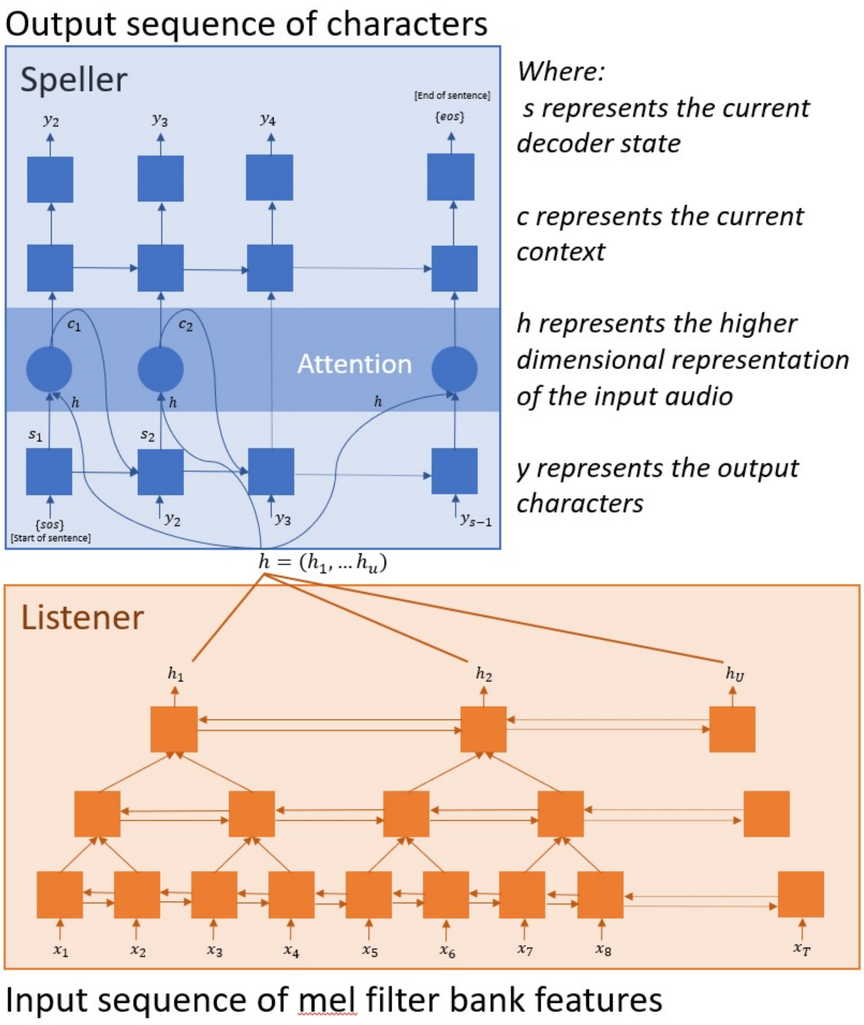

Output sequence of characters
Speller
y_2
y_3
y_4
[End of sentence]
$\{eos\}$
c_1
c_2
Attention
h
h
h
s_1
s_2
$\{sos\}$
[Start of sentence]
y_2
y_3
y_{s-1}
$h = (h_1, \ldots h_u)$
Listener
h_1
h_2
h_U
x_1
x_2
x_3
x_4
x_5
x_6
x_7
x_8
x_T
Input sequence of mel filter bank features
Where:
s represents the current decoder state
c represents the current context
h represents the higher dimensional representation of the input audio
y represents the output characters

Figure 2.11: Listen, Attend, and Spell Model based on Chan et al. 2015

best describes the length of audio samples used in LSA. Google Cloud Speech is accessed through the cloud and therefore places minimal computing requirements on the user. The cost to use Google Cloud Speech is also quite low ($.009/15 seconds of audio), which may make it a viable option for clinicians to use for automatic transcription.

Applications of ASR to Child Speech

As complex as it is to take an audio file and convert it to speech, even greater complexity is added when completing this process for child speech. In comparison to adult speech, which the majority of ASR systems are trained on, child speech contains high levels of both acoustic/spectral variability, caused by the continued development of articulators and vocal tract, and linguistic variability, caused by the continued development of language skills. In addition to the difficulty posed by highly variable speech data, there is also a scarcity of speech corpora (i.e., audio data and corresponding transcripts) available to train child-specific ASR systems. Several methods have been proposed to deal with these limitations, including data augmentation, speaker normalization and transfer learning.

Data augmentation describes the practice of slightly perturbing the dataset to produce a new dataset that corresponds to the same labels, that also increases the size of the dataset two (or more) fold. For example, it is common in ASR research to apply speed perturbations, such that the audio signal is transformed creating new audio data, but the corresponding transcript is unchanged. Speaker normalization, in the context of child ASR, attempts to normalize the speech signal so that it more closely approximates adult speech. A common type of speaker normalization is called vocal tract length normalization (VTLN), which normalizes the spectral features of child speech, making it less variable (Potamianos et al., 1997). Transfer learning describes a method of training a model on a large source of data to learn good weight and bias values, and then retraining the final layer of the model on some smaller set of data to slightly adjust the learned parameters (Torrey & Shavlik, 2010). In this case, transfer learning would involve first training a model on large adult speech corpora and then training the final layer of model on child speech. This is done so that the system learns a good encoder-decoder and language model from the large amount of adult training data, but then is hopefully specialized to better handle child speech from the smaller set of child data.

Such processes have been successfully applied in prior work (e.g., see Shivakumar & Georgiou, 2020; Gale et al., 2019; Shahnawazuddin et al., 2020; Yadav & Pradhan, 2021), where WERs as low as 9% for unseen test sets of typically developing children (4-14 years) and 26% for children with impaired language abilities (kindergarten through third grade) have been reported. It is also possible however, that with a robust enough ASR system, such methods would be unnecessary. In an investigation of state-of-the-art cloud-based ASR system, Rodrigues et al. (2019) found evidence that Google Cloud Speech was not significantly impacted by speaker age, having included speech samples from both typical adults (18-51 years) and children (8-14 years). To the authors knowledge, only limited work has investigated the accuracy of Google Cloud Speech on children with impaired language abilities (e.g., C. B. Fox et al., 2021).

Natural Language Processing (NLP)

NLP generally describes different forms of text analysis, the scope of which will not be covered here. Instead, several aspects of NLP relevant to automating the analysis portion of LSA will be discussed, including tokenization, cleaning, part-of-speech tagging, stemming, and string functions, see 2.12.

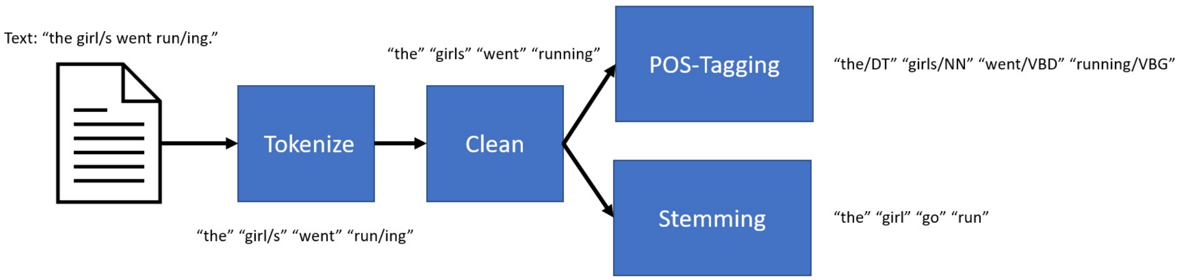

Text: “the girl/s went run/ing.”
Tokenize
“the” “girl/s” “went” “run/ing”
Clean
“the” “girls” “went” “running”
POS-Tagging
“the/DT” “girls/NN” “went/VBD” “running/VBG”
Stemming
“the” “girl” “go” “run”

Figure 2.12: NLP Preprocessing Procedures

Tokenization

Tokenization is the first step in any NLP task because it is required to turn raw text into a usable format for processing. Tokenization refers to the process of splitting up raw text data into smaller, meaningful units called tokens. Tokens can take the form of characters, words, or sub-words (also called n-grams). Tokenization at the word-level is quite common and is determined by the space characters that lie between words (e.g., the black dog “the” “black” “dog” based on the two spaces falling between the three words). Tokenization can also occur at the character level, treating each individual letter, number, or special character as a token, or at the sub-word level to break down free and bound morphemes, for example, into individual tokens (e.g., greatest → “great” “est”). The current study utilized word-level tokenization.

Cleaning

Cleaning describes the steps taken to prepare tokenized text for processing. This may include removing unwanted annotations (e.g., codes) in the text, if applicable. Some formatting adjustments may also need to be made, such as replacing additional punctuations (e.g., commas, question marks, etc.) with periods and segmenting utterances. These replacements and deletions can be performed automatically using a series of functions called regular expressions or regexes. Another common formatting change is stemming, which describes the process of removing inflectional endings like plural s, past-tense ed, possessive ‘s, and so on, reducing words to their root (Jurafsky & Martin, 2020). This can be useful when comparing a reference (i.e., ground truth) and hypothesis (e.g., ASR produced) transcript, such that minor differences in tense, for example, can be ignored when evaluating transcript similarity.

Part-of-Speech Tagging

After the corpus has been cleaned, part-of-speech tagging may be implemented to classify the types of words used in a text. A part-of-speech tagger tool is designed to either replace or appended the part-of-speech to the end of each word in a text (e.g., cat → NOUN, or cat/NOUN). Part-of-speech taggers are typically trained on a large, labeled corpora (i.e., texts that have been manually tagged for parts-of-speech) and are then able to generate part-of-speech tags on unseen texts with varying levels of accuracy (Jurafsky & Martin, 2020). Part-of-speech tagging has numerous applications in NLP, however, in the current study it was used to automatically identify certain literate language features (e.g., adverbs) in language samples. This application will be described in greater detail in later sections. Part-of-speech taggers are typically included as one function in an NLP tool-kit. For example, NLTK (Bird, 2006), which is an open-access machine-learning based NLP-toolkit with an available Python implementation, includes a part-of-speech tagger. Its part-of-speech tagger uses a FFNN classifier, but details relating to its specific architecture are not available. The accuracy of a part-of-speech tagger usually depends on how similar the unseen texts are to the training corpora (e.g., a tagger trained on Wallstreet Journal articles will likely be less accurate on narrative texts than other journal articles).

String Functions

String functions, which are built into Python, make creating functions to match existing language assessment tools possible. String functions can be used to identify specific patterns in a text and return a score or value. For example, one could create a function to count the number of instances of adverbs marked with the ADV tag or create a function to identify complex sentences (i.e., sentence containing a subordinate clause or non-canonical sentence) by specifying patterns of various independent clause types (e.g., noun phrase, verb phrase, simple declarative) that are followed by a clause headed with a subordinating conjunction tag. For example, if a pattern of noun phrase + verb phrase + subordinating conjunction was coded as string indicating a complex sentence, then the sentence *the small boy* (noun phrase) *ran quickly* (verb phrase) *because* (subordinating conjunction) *he was scared*, could be identified as a complex sentence using string matching. These rulesets can vary in complexity depending on the language measure of interest; however, such string functions provide the opportunity to automate the analysis portion of LSA by operationalizing the measures of interest. More qualitative measures (e.g., holistic measures of language sample quality) cannot be scored using hard-coded rulesets, and require additional machine learning based solutions that will not be discussed here.

Applications of NLP to Child Language

NLP applications have been utilized both in the direct prediction of language impairment from children's transcribed language and in automatically computing language assessment tools. Both types of applications offer the opportunity to assess the language abilities of children more quickly. Gabani et al. (2009) examined the accuracy of both language (i.e., n-gram models) and machine learning models in predicting language impairment (LI) status in adolescent children (13-16 years) based on a collection NLP features extracted from spontaneous narrative language samples. The corpus included both story-telling and personal narratives from 99 adolescents with typical language and 19 with LI. Language models were trained to classify children as LI based on the perplexity value of the POS combinations present in a child's narrative transcript. Perplexity is the inverse of probability, such that lower probability POS combinations are associated with higher perplexity values, meant to represent the fact that children with LI use less grammatically correct sentences. Machine learning models were trained on a greater number of features, including language productivity (e.g., MLU, NTW), morphosyntactic skill (e.g., subject-verb agreement), vocabulary (e.g., NDW), fluency (e.g., repetitions and revisions), and language model perplexity values. Cross-validated results indicated that machine learning models outperformed the language models for both the story-telling and personal narrative conditions, with a F1 score (i.e., a type of classification accuracy metric) of 72.22% and 56.25%, respectively.

Hassanali et al. (2012) improved upon these results by adding in additional features generated in Coh-Metrix, an open-source text analysis tool, including readability, situation model (e.g., causal features, temporal features), word features (e.g., frequency of content words), syntactic features (e.g., use of connectives, number of noun phrases), referential (e.g., number of adjacent utterances with argument overlap). Their best performing model achieved an F1 score of 91.4% for story-telling and 66.7% for personal narratives. While the performance in classifying LI status from personal narratives remained low, Hassanali et al. (2012) found that machine learning methods trained on NLP features could be utilized to classify LI status from story-telling samples with high levels of accuracy. These results provide evidence that integrating NLP into predictive models could be used help clinicians quickly screen for children with TD and LI status, however, additional work is needed to extend these findings to younger age ranges whose usage of these language features may significantly differ from adolescents. For example, syntactic knowledge continues to develop throughout childhood, with the incorporation of causal connectives in narrative discourse not appearing until age nine or later in TD children (Berman & Nir-Sagiv, 2007).

Additional efforts have been put towards automating language assessment tools, such as the index of productive syntax (IPSyn; Scarborough, 1990). The IPSyn is designed to measure the development of particular syntactic forms in expressive language. The IPSyn measures four main syntactic constructs, including noun phrases, verb phrase, questions and negations, and sentences for a total of 60 items that are scored based on their number of unique instances. Several automated scoring systems exist for the IPSyn, the most recent being the automatic computation of the IPSyn system (AC-IPSyn; Hassanali et al., 2014). The AC-IPSyn uses both automatic POS tagging and syntactic parsing, in combination with hard-coded rule-sets to automatically identify the 60 IPsyn structures and then compute the associated scores. The point-by-point accuracy was evaluated on two datasets, one which included 20 transcripts elicited from young typically developing (TD) children (2-3 years-old) and one which included 20 transcripts from 10 TD children and 10 children with LI (6 years-old). Results indicated that AC-IPSyn was able to compute scores with high-levels of accuracy across both datasets, 96.9% and 96.4%, respectively. The authors concluded that use of this automated assessment tool could result in highly accurate scores that significantly cut down on the time spent conducting manual analysis. The current study built upon this work by utilizing similar methods (e.g., part-of-speech tagging and syntactic parsing) to automatically assess additional linguistic indices of interest to analyzing narrative language samples (i.e., literate language).

In isolation, neither ASR nor NLP are sufficient in fully addressing the time-costs of LSA, since ASR only addresses expedited transcription while NLP only addresses expedited analysis. In combination however, they offer the opportunity to streamline the entire assessment process (apart from elicitation, of course). Researchers in speech language pathology and related fields have already begun to recognize the utility of these systems for automating certain education and language assessments, but only limited work has examined their implementation for LSA with even fewer examining automated transcription.

Applications of ASR & NLP to Child Language & Literacy Assessment

In a recent study, Lileikyte et al. (2020) developed a system intended to identify children who were at-risk for speech and/or language delays by automatically calculating language productivity (i.e., total words) from transcripts produced through ASR. Audio was recorded and then manually transcribed across 33 children (2.5-5 years old), totaling 15 hours of recordings, which were collected using Language Environmental Analysis (LENA) recording units. All recordings took place during child-child and adult-child interactions occurring throughout routine preschool activities.

The authors trained a baseline ASR system and compared its accuracy to a deep neural network-ASR (DNN-ASR) system trained with and without data augmentation (e.g., speed and tempo perturbations made to the original audio so as to double or triple the size of the corpus). Their best performing ASR model (DNN-ASR trained on augmented data) achieved a poor level of transcription with an average WER of 63.74%. Even with high transcription error however, the system was able to generate reliable estimates of children's language productivity (word count). While their ASR model proved poor for transcription purposes, it could still be used to reliably analyze language productivity likely because transcription errors did not significantly impact word count. Thus, Lileikyte et al. (2020) found that their ASR system had potential utility in helping to identify at-risk children within an early childhood environment based on their language productivity.

Other studies have investigated the use of ASR for scoring various educational assessments, without directly examining accuracy in transcription (Asgari et al., 2016; Gale et al., 2019; Nese & Kamata, 2020). In their study, Nese & Kamata (2020) examined the feasibility of using an open-sourced ASR toolkit *Bavieca* to automatically score the Curriculum-Based Measure of Oral Reading Fluency (CBM-R; Deno, 1985), which is an oral reading proficiency test administered to elementary school children. The CBM-R requires children to read grade-level narrative texts while an examiner records the number of word-errors they make within a 60-second time frame. The total number of words correct per minute (WCPM) is calculated by subtracting the number of incorrect words from the total number of words read. Traditionally this score has been computed by hand and because this test is administered in real-time, it is prone to administration and scoring mistakes (Christ & Silberglitt, 2007; Cummings et al., 2014). In their study, the accuracy of traditional CBM-R scoring done in real-time by trained scorers and those obtained using ASR were compared against criterion scores produced by trained scorers from audio recordings.

A total of 13,766 audio recordings across 902 students were examined by grade level (2-4) in this study. Results indicated that ASR tended to underestimate WPCM, with the average estimated WPCM for ASR ranging between 83.7-106.0, while average estimates for criterion scores ranged between 100.7-123.2. These differences in average WPCM were statistically significant across each grade-level, $p < .00125$. By comparison, WPCM estimates produced using traditional scoring were closer to criterion estimates and were only significantly different for second grade scores ($SE = 1.8, p < .00125$). The authors concluded that while using ASR for scoring the CBM-R may result in slightly less accurate results than traditional hand-scoring, the time and cost-saving benefits associated with automatic scoring should be weighed against this loss.

A more relevant study on the use of ASR as it relates to language transcription for children was conducted by Gonzalez Villasanti et al. (2020) who developed and examined an automatic tool called the Classroom Interaction Detection and Recognition system, which was designed to transcribe and then code for child-directed speech (CDS) interactions within preschool classrooms (CIDR; Gonzalez Villasanti et al., 2020). CIDR utilizes several Amazon web service modules, including their computer vison service *Amazon Rekognition* (used for facial recognition) and their ASR service *Amazon Transcribe*. Both are state-of-the-art competitors with Google Cloud suite products.

Visual data and language samples were collected from 13 children (2;11-4;10) using wearable LENA recording devices and head-mounted cameras over the course of two one-week periods. All children were recruited from the same preschool classroom from a non-profit childcare center. Recordings took place during non-group activities (structured and unstructured) where the focal child engaged with adults and their peers (e.g., free-play). Focal children's interactions with adults and peers were coded by trained research staff to identify CDS, which were then manually transcribed in SALT. CIDR utilizes a deep-learning model trained to integrate facial feature and speech activity data to detect CDS interactions directed towards the focal children. Results indicated that CIDR correctly classified CDS with adults at 81.1% of the time and 86.1% of the time with peers. ASR performance for speech-to-text conversion of CDS interactions with adults and peers was subpar however, with WERs above consistently 60%. As a result, the authors concluded further work was required to improve ASR specific to child-produced speech.

Fox et al. (2021; under review) added to this body of research by conducting two pilot studies that investigated the automation of separate pieces in the LSA process using ASR and NLP applications. These pilot studies served to lay the groundwork for the present project and are described in detail below. The first study investigated the use of ASR in transcribing narrative language samples, and the second was designed to test the accuracy of a scoring system designed to evaluate specific literate language features important to the development of scholarly oral and written discourse in school-age children.

Pilot Studies

Pilot Study 1: Automated Transcription

In the first pilot study, Fox et al. (2021) compared the transcription accuracy of Google Cloud Speech ASR against RTT performed by SLPs and trained transcribers on 42 narrative language samples of school-age children (7;5-11;9) with DLD. Narrative samples were 1-3 minutes in length and were elicited via one of three prompts, including a narrative retell and two semi-spontaneous (i.e., some scaffolding was in place based on model stories from other sections of the test administered) narratives with either a sequenced or single-scene picture prompt. All prompts were administered as a part of the Test of Narrative Language, which is a norm-referenced assessment of school-age children's narrative listening comprehension and oral narrative production abilities (TNL; Gillam & Pearson, 2004). A total of seven SLPs and seven trained transcribers participated. Participants in each group were assigned a random selection of six unique language samples to transcribe in real-time (i.e., as they listened to it). The same 42 language samples were also transcribed automatically with Google Cloud Speech. Transcripts produced with both expedited methods (ASR and RTT) were compared against a reference corpus that was produced using traditional transcription that served as the gold-standard. Transcription error was evaluated with a weighted word-error rate (WER_W).

Results indicated that ASR-produced transcripts had significantly less transcription error than either source of RTT (SLP or trained transcriber, $WER_W M \approx 42\%, SD \approx 20\%$), with an average WER_W of 30% (SD = 10%). In addition, moderation analysis indicated that speech rate had a significant impact on RTT error, but not ASR transcription error On average, language samples of children who spoke at below average rates (less than 75 words per minute) had lower transcription error when produced with RTT However, for samples spoken at 75 words or more per minute, ASR had significantly lower error than RTT with the disparity between the two expedited methods growing as a function of speech rate. As compared to Gonzalez Villasanti et al. (2020) who found WERs consistently above 60%, transcription error of ASR-produced transcripts was considerably lower, which may be attributed to differences in speaker age-range (i.e., Gonzalez Villasanti et al. investigated preschool-age children as compared to school-age) and audio/speech quality (e.g., Gonzalez Villasanti et al. recorded children in a naturalistic setting with background speakers instead of a quiet room).

To evaluate the clinical utility of the expedited transcription methods, ASR- and RTT-produced transcripts were input into SALT to generate four common LSA indices, including total number of utterances (TNU), mean length of utterance in words (MLU-words), number of total words (NTW), and number of different words (NDW). Accuracies of the generated scores were evaluated by examining their linear correlations with scores generated from the reference corpus. Pearson coefficients indicated strong correlations between ASR- and reference-transcript scores on all four indices $r(40) = .87 - .99$, while correlations between RTT- and reference-transcript scores were more moderate $r(40) = .66 - .82$. These results provided preliminary evidence that ASR could be used to automatically transcribe language samples for school-age children and reliably generate four quantitative indices. It was unknown the degree to which transcription error (M = 30% WER) would impact LSA indices that examine usage of specific linguistic devices (e.g., adverbs, conjunctions, etc.), however. It was expected that transcription error may have a more negative impact on these indices, given that the errors of substitution (i.e., incorrectly transcribing the words, not the number of words) were most common for Google Cloud Speech in the pilot study. As previously discussed, word substitutions would not impact measures of productivity like TNU or NTW, but had the potential for greater impact on measures of language quality, like literate language, which are scored based on the usage of particular words or phrases.

Pilot Study 2: Automated Scoring

In a second pilot study, C. Fox et al. (2021) examined the feasibility of automatically generating scores from manually transcribed narrative samples for a set of specific language features important for the generation of scholarly oral and written discourse. These features, often referred to as "literate language" features, included coordinating and subordinating conjunctions, metacognitive and linguistic verbs, adverbs, and elaborated noun phrases (ENP). See Table 2.1 for definitions. A system called *Literate Language Use in Narrative Assessment* (LLUNA) was designed to automatically score each of the six indices of literate language individually. The functions underlying LLUNA were written to match the scoring scheme of an existing progress-monitoring tool, the *Monitoring Indicators of Scholarly Language*, which assigns a score of 0-3 to each of the six measures based on their usage within a spontaneous narrative (MISL; Gillam et al., 2017). Five of the six measures are scored based on the number of unique instances within the narrative sample, where a score of 0 indicates no usage of the literate language convention and a score of 3 indicates three or more unique instances of the convention (e.g., *and, but*, and *so* for coordinating conjunction). ENP is scored based on the number of elaborative words that precede a noun in a given noun phrase, such that a score of 0 indicates a noun in isolation (e.g., *Dog* ran home), while a 3 indicates that three or more elaborative words precede a noun (e.g., *The big black dog* ran home). Post-noun modification was not addressed in this study, as it is generally not consistently used until students are 10 years of age and older (Eisenberg et al., 2008).

In order to match the scoring scheme of the MISL, a combination of hard-coded string matching functions, part-of-speech tagging (using NTLK), and a manually specified syntactic coding scheme was used to automatically score each index individually. The usage of these components for scoring each literate language index is described in the following sections.

Three indices, including subordinating conjunctions, mental verbs, and linguistic verbs, utilized only string matching in their scoring functions because they were the least susceptible to semantic ambiguity. Semantic ambiguity occurs when words have different meanings/parts-of-speech depending on the given context. In the case of these three literate language indices, this is rarely an issue (e.g., *said* is always a linguistic verb, because is always a subordinating conjunction). Given the straightforward nature of identifying these indices, these scoring functions were

Table 2.1: Literate Language Measures Defined

Term	Definition
Coordinating Conjunction	Words that connect two independent clauses, such as *and, but* and *or.*
Subordinating Conjunction	Words that connect an independent and dependent clause, such as *because, therefore*, or *when.*
Linguistic Verbs	Verbs that indicate dialogue, such as *yelled, said* or *whispered.*
Mental Verbs	Verbs that indicate cognitive state, such as *wondered, thought*, or *decided.*
Adverbs	Verb modifiers of degree, time, manner or place.
Elaborated Noun Phrase	Noun phrases that contain a set of modifiers that elaborate on the given noun, e.g., *The big black dog.*

Note. A full explanation for each microstructure element from the MISL can be found in Gillam, Gillam, Fargo, Olszewski & Segura (2017): *Monitoring Indicators of Scholarly Language: A Progress-Monitoring Instrument for Measuring Narrative Discourse Skills.*

first developed by compiling a constrained list of common, age-appropriate words from each index. Once the wordlists were compiled, rulesets were coded in Python by designing functions unique to each literate language index. These functions first utilized word-level tokenization to convert the raw text of each transcript into word tokens. String matching was then specified to find overlap between the word tokens within the transcript and the words contained in each indices' word-bank. The number of unique instances of overlap was then assigned a corresponding score (i.e., zero for no instances, one for one unique instance, and so on) that was capped at three to match the MISL rubric.

The function for scoring coordinating conjunctions was developed in a similar manner but required additional specifications. As with the previously discussed indices, a wordlist was compiled to contain the typical coordinating conjunctions utilized by this age group. These are relatively limited in scope and only included *for, and, nor, but, or, yet* and *so.* Once compiled into a list, the coordinating conjunctions scoring function was written to tokenize the transcript text to words, extract the first word from each C-unit, determine the number of unique instances of overlap between the set of word tokens and the specified wordlist, and then assign a score between 0-3. The main difference in the coordinating conjunctions function then, was that it was designed to only examine the first word of each C-unit. This was designed to prevent counting instances of these words when they were not being used as coordinating conjunctions. The word "and" for example, can be used to list off objects, such as in the sentence: *she like apples, and bananas, and oranges.* Similarly, other words present in the specified wordlist do not always serve to coordinate conjunctions. By only examining the first word in each C-unit, the likelihood of misclassifying a coordinating conjunction decreased, since the C-unit structure splits utterances at the instance of a coordinating conjunction (e.g., "the girl went home and then she ate dinner" → "the girl went home. And then she ate dinner.").

The final literate language indices, adverbs and ENP employed a part-of-speech tagger, from the *OpenNLP* package in R (Hornik, 2019; RCore-Team, 2021). This was necessary as these indices were most susceptible to scoring error through semantic ambiguity. Adverbs cover a wide variety of words, many of which serve as different classes given the context (e.g., *like* can be a preposition, adverb, conjunction, noun, verb, or adjective). The part-of-speech tagger function from *OpenNLP* served as a better solution to identifying adverbs than using predefined word bank. This part-of-speech tagger was trained on a large corpus of newspaper articles however, so part-of-speech tagging was not 100% accurate; though it maintained adequate accuracy upon examination of transcripts. The part-of-speech tagger function from *OpenNLP* was implemented through its open-source package in R and was used to tag adverbs within each narrative. The number of unique adverbs were then counted through string matching and assigned a corresponding score.

Elaborated noun phrase also employed the part-of-speech tagger function to replace the words within each narrative with their part-of-speech tag (e.g., *the quick girl ran* → determiner adjective noun verb). ENP is scored based on the number of modifiers that precede a noun, and modifiers can include only determiners, numbers, pronouns, adverbs, and adjectives. In addition, these modifiers can only occur in certain orders, for example, in a grammatically correct sentence, a determiner cannot follow an adjective within the same noun phrase (e.g., *silly a girl* is not grammatical). With these constraints in mind, a syntactic scoring scheme was designed by compiling possible permutations of word classes that could precede a noun and result in a score of a 0 (e.g., noun in isolate) to 3 (e.g., three or more modifiers preceding a noun). Once the part-of-speech tagger had been applied to the corpus and the scoring scheme constructed, the function was designed to parse the text for overlap between the tagged word tokens in the transcript and the syntactic scoring scheme to identify the largest string of modifiers preceding a noun and then assign the corresponding score.

These combined functions formed the comprehensive tool, LLUNA. In order to determine the accuracy and reliability of LLUNA, it was evaluated upon fifty narrative transcripts randomly selected from a large normative sample consisting of elementary school-age children between the ages of 5;0 and 9;11. Each of the fifty samples were also evaluated manually by a team of trained, but non-expert scorers and one expert. The non-experts had received training on the identification of literate language features in narrative language samples and had met at least 85% interrater reliability with the expert, but had only around 1 year of experience. Non-experts were included to gauge the reliability of a MISL hand-scorer (e.g., research assistant, SLP). The expert had over five years of experience with MISL scoring, and so their evaluation of language samples was used to determine the accuracy of scores generated automatically via LLUNA. Accuracy was evaluated by calculating the interrater reliability between scores manually produced by the expert and LLUNA scores using a weighted kappa metric (Cohen, 1968).

Acceptable levels of accuracy ($\kappa_{qw} \geq .60$) were achieved across all six literate language indices (coordinating conjunctions, $\kappa_{qw} = .78$; subordinating conjunctions; $\kappa_{qw} = .88$, mental verbs, $\kappa_{qw} = .89$; linguistic verbs, $\kappa_{qw} = .89$; adverbs, $\kappa_{qw} = .79$; ENP, $\kappa_{qw} = .74$). In addition, LLUNA achieved higher levels of scoring reliability as compared to trained scorers on four of the six measures, including subordinating conjunctions, mental verbs, linguistic verbs, and adverbs, though there was room for improvement on ENP and coordinating conjunctions. The pilot of LLUNA therefore showed promise as a method for automatically evaluating literate language indices in narratives, however, it was a goal of the current project to get LLUNA's scoring accuracy on par with or more reliable than trained scorers across all six indices.

LLUNA was also only tested on manually produced transcripts. Therefore, it was also of interest to determine whether LLUNA could maintain accuracy in scoring aspects of literate language when transcripts were generated from ASR, as the use of ASR in combination with LLUNA could substantially improve the efficiency of LSA.

Purpose & Rationale of the Study

The current project aimed to build on the findings from two pilot studies by Fox et al. where the feasibility of automating separate components (transcription, scoring) of the LSA process on small samples ($N = 42, N = 50$, respectively) were investigated. The primary goal of this project was to investigate the feasibility of streamlining the LSA process by combining ASR and LLUNA to automatically transcribe and then compute six literate language indices from audio recordings of narrative language samples elicited from school-age students between the ages of 6;0 and 11;11 with impaired language abilities (AR/DLD). A number of aims were set in order to best address this goal.

Aim 1

The first aim of this project was to replicate and extend the findings from the Fox et al., 2021 pilot to a larger and more diverse set of language samples, in terms of age-range and language ability. This was accomplished by determining the level of transcription accuracy that could be achieved by Google Cloud Speech on the oral narratives of school-age (6;0-11;11) children with impaired language abilities (DLD), including those who were at risk for language and literacy disabilities (AR).

Though there is evidence that impaired speech can negatively impact transcription accuracy (Calvo et al., 2021; Espana-Bonet & Fonollosa, 2016; Young &

Mihailidis, 2010), it was unknown how *language* ability might affect Google Cloud Speech transcription, which partially relies on a pretrained language model in its generation of transcripts. In the pilot study, Fox et al., 2021 utilized samples of children with DLD between the ages of 7;5 and 11;10. It was hypothesized that in addition to including samples from students with DLD from a larger age range (6;0-11;11), that including samples from students designated as AR for language and literacy difficulties might further address this aim by adding diversity to the evaluated language samples. This hypothesis was based on the assumption that language analyses may be conducted on samples obtained from AR students as part of the diagnostic process because they are commonly referred to SLPs for assessment.

A secondary focus of aim 1 was to examine the potential associations between different language sample recording characteristics (perceived audio quality, background noise, recording type) and transcription error. Recall that in Fox et al. (2021), only high quality recordings were included, which were not representative of the way that recordings might be obtained in clinical settings. It was therefore deemed important to determine the degree to which recording characteristics might be associated with transcription error, so that clinical recommendations regarding the use of ASR in authentic practice might be made. Therefore, recordings were obtained from different sources that utilized a variety of audio recording devices/systems and elicitation settings (quiet, background noise etc.).

Based on the recommendations for best outcomes listed by Google Cloud Speech, it was expected that samples recorded with less background noise and minimal to no compression audio encoding (specified by the codec) would have the highest transcription accuracy. Similarly, it seemed a logical assumption that higher perceived audio quality, based on human ratings, might correspond to higher transcription accuracy.

Aim 2

The second aim of this project was to increase LLUNA scoring accuracy in general, and for a wider, more diverse population of students (language status, age range). While the pilot study of LLUNA (Fox et al., under review) was promising, there were aspects of scoring that might be improved. For example, in the pilot project, LLUNA scoring for coordinating conjunctions and ENPs had kappas of .78 and .74, respectively. While this was considered to be above the acceptable level of interrater reliability ($\kappa_{qw} = .60$) based on common standards of automated scoring (Dikli, 2006), these values were lower than what was achieved by trained scorers ($\kappa_{qw} = .86, .78$). In the current project, it was a goal to have LLUNA achieve scoring accuracy that was on par with or exceeded that of trained MISL hand-scorers for each of the six literate language indices.

It was hypothesized that updating the part-of-speech tagging model underlying LLUNA, as well as expanding/altering its wordlists and syntactic coding scheme would increase its scoring accuracy for coordinating conjunctions and ENP, as well as the other indices (subordinating conjunctions, adverbs, mental and linguistic verbs). In the pilot of LLUNA, scoring accuracy was evaluated on a normative sample of school-age children, meaning samples were primarily representative of typically developing children. In the current project, LLUNA was evaluated on data from children who were AR for language and literacy difficulties or who had DLD. Children who are AR or have DLD generally demonstrate poorer vocabularies, and greater levels of grammatical and syntactic errors. This had the potential to impact the scoring accuracy of LLUNA by reducing the variety of literate language items it was tested on. It was therefore necessary to not only determine whether the levels of accuracy reported in the pilot of LLUNA could be improved upon, but whether they could generalize to the narrative language samples of AR/DLD children.

Aim 3

The third aim of this project was to evaluate how transcription error in ASR-produced transcripts impacted LLUNA scoring accuracy. It was expected that transcription accuracy would be variable across language samples, therefore it was of interest to see how the degree of error in transcription might impact LLUNA scoring accuracy.

Given findings from the Fox et al. (2021) pilot study, it was expected that there would be errors in transcription, but it also seemed likely that there could be an acceptable range of error that still resulted in reliable LLUNA scoring accuracy. Even with the average WER of 30% found in the pilot, four quantitative metrics including NTW, NDW, TNU, and MLU-words could be generated with a high level of reliability ($r(40) = .89 - .97$). Of all types of transcription error (deletion, substitution, and insertion), Google Cloud Speech was found to be most prone to errors of substitution (e.g., really → chilly). Because the literate language indices are scored based on either the number of unique instances or number of modifiers, errors of deletion would likely pose the greatest threat to accurate LLUNA scoring. Errors of substitution may have less of an impact on LLUNA scoring as long as the same word type is substituted. It was expected that LLUNA may therefore be robust to transcription error, up to a point. It was necessary to determine the point, or more likely, range of transcription error that LLUNA could tolerate before its scoring accuracy suffered.

Collectively, these three aims were intended to address the larger goal of determining the feasibility of streamlining the LSA process through automated transcription (through ASR) and scoring (through LLUNA) on a diverse set of language samples elicited from school-age children (6;0-11;11) with impaired language abilities (AR and DLD). The proposed aims were addressed by the following research questions:

1. What level of error (as measured by WER_w) is present in transcripts produced by Google Cloud Speech ASR on the oral narratives of school-age children (6;0-11;11 years) who are at-risk for language and literacy difficulties (AR) or who have developmental language disorder (DLD)?
2. Are subjective ratings of audio recording quality (as measured by mean opinion score), background noise rating, or audio encoding/decoding type (i.e., codec) associated with transcription error?
3. Are updates made to LLUNA's scoring functions associated with improvements in accuracy (as measured by κ_{qw}) across the six measures of literate language for oral narratives elicited from school-age children?
4. Do accuracy levels generalize to the oral narratives of school-age children with impaired language abilities (AR, DLD)?
5. How does LLUNA score accuracy (as measured by score error) vary across different levels of transcription error from ASR-produced transcripts?
6. Is there an acceptable range of transcription error (WER_w) that is associated with maintained scoring accuracy?

CHAPTER 3

Methodology

Language Samples

A total of 255 narrative language samples were utilized in this study, elicited across an equal split of three groups of English-speaking monolingual elementary school-age children who ranged in age from (6;0-11;11 years). Language samples of students were included who were judged to be typically developing (TD; $n = 85$), at-risk for language and literacy difficulties (AR; n = 85), and to have developmental language disorders (DLD; $n = 85$). Because language abilities can vary widely amongst this developmental range, language samples were split into two age groups, one which included samples elicited from children ages 6;0-8;11 ($n = 150$) and one for ages 9;0-11;11 ($n = 105$). The differences in sample sizes between age groups was due to the availability of narrative language samples meeting inclusionary criteria requirements, however, as statistical tests were not conducted on age group differences, this was not deemed a critical issue.

Each group of language samples was collected as a part of other studies related to school-age children's narrative language abilities. Across each study, the Test of Narrative Language (TNL) was used to elicit three narrative language samples per child (Gillam & Pearson, 2004). The first edition of the TNL was used to collect the TD and DLD language samples, while the second edition of the TNL was used to collect the AR, as well as additional DLD language samples (TNL-2; R. Gillam & Pearson, 2017). Both editions of the TNL contain three prompts for eliciting different types of oral narratives, including a retell, a semi-spontaneous narrative with a sequenced picture prompt, and a semi-spontaneous narrative with a single-scene picture prompt. The main difference between the first and second edition is in the retell prompt, thus all narrative retells were excluded from the potential pool of language samples.

Sample Selection Process

Language samples between 1 and 7 minutes in length were deemed to be of sufficient length to obtain valid and reliable measures to answer the research questions that were posed (J. Heilmann et al., 2010; Tilstra & McMaster, 2007). Therefore, all samples shorter than one minute were excluded from the potential pool of samples. The upper limit was deemed less critical, given that duration was not found to be a significant predictor of transcription accuracy in the pilot study (C. B. Fox et al., 2021). However, an upper limit of seven minutes was still set, based on the upper limit used in Heilman et al. (2010), where they established that measures calculated from seven minute language samples were consistent with language samples of one or three minutes.

When possible, random selection of samples was utilized on the remaining eligible pool of language samples across each language ability (TD, AR, DLD) and age group (younger; older) in an attempt to evenly distribute language sample characteristics. Only the DLD group for the 6;0-8;11 age-range was not compiled through random selection, due to a scarcity of language samples that were a minute or longer; the remaining sample (80%) utilized random selection. Demographic variables including age, gender, and ethnicity are reported as descriptive statistics to present information on the narrator of the language sample, see 3.1. It was expected that language sample characteristics including duration of sample, productivity (e.g., NTW), intelligibility, and complexity would differ across groups, so ad-

Table 3.1: Narrator Demographic & Language Sample Information

Index	TD ($n = 85$)	AR ($n = 85$)	DLD ($n = 85$)
Female	49 (57.6%)	36 (42.3%)	32 (37.6%)
Male	36 (42.4%)	42 (49.4%)	51 (60%)
No Response	0 (0%)	7 (8.3%)	2 (2.4%)
White	81 (95.3%)	41 (48.3%)	45 (52.9%)
Latino	0 (0%)	21 (24.7%)	20 (23.6%)
2+ Ethnicities	0 (0%)	4 (4.7%)	6 (7.2%)
African American or Black	4 (4.7%)	6 (7.1%)	2 (2.3%)
No Response	0 (0%)	7 (8.2%)	5 (5.9%)
American Indian	0 (0%)	3 (3.5%)	3 (3.5%)
Native Hawaiian	0 (0%)	2 (2.3%)	2 (2.3%)
Asian American	0 (0%)	1 (1.2%)	2 (2.3%)
Alien Story	61 (71.8%)	61 (71.8%)	60 (70.6%)
LFS Story	24 (28.2%)	24 (28.2%)	25 (29.4%)
Sample Characteristic	$M(SD)$	$M(SD)$	$M(SD)$
Age	8;10 (1;7)	8;9 (1;4)	8;9 (1;1)
Total Utt.	20.9 (14.4)	18.8 (8.6)	15.2 (6.5)
% Intelligible Utt.	1.0 (0.0)	1.0 (0.0)	1.0 (0.1)
MLU words	7.7 (1.5)	7.6 (1.5)	7.5 (1.6)
Total words	143.4 (80.1)	123.6 (54.0)	99.5 (42.9)
Different words	72.4 (24.9)	60.8 (19.2)	52.4 (17.6)
Duration (sec)	91.8 (35.4)	101.2 (41.5)	85.2 (29.4)

Note. TD = typically developing, AR = at-risk for language and literacy difficulties, DLD = developmental language disabilities, NR = non-response.

ditional descriptive statistics are also provided by group in 3.1.

In the following sections the screening and selection process for narrative transcripts is described by language ability and age group.

Age-Range: 6;0-8;11

Typically Developing Language Samples

Criteria for typically developing (TD) status was determined through standard score on the TNL. The score range to be considered TD was between 95-115 points, or up to 0.33SD units below the mean score.

The TD corpus was sourced from a larger normative sample of 1,782 children aged 6;0-15;11. Exclusionary criteria were applied to reduce these samples down to an eligible pool, this was accomplished through the following steps. Exclusion of samples was first made based on participant age; a total of 284 participants fell within the 6;0-8;11 age range. For the purpose of the current project, language samples were also excluded if they were elicited from bilingual children, children whose first language was not English, and children who had disability status(es). Based on these criteria, the language samples of a total of 48 participants between 6;0-8;11 were excluded, bringing the eligible pool down to 236. An additional 87 participants were excluded as they fell outside of the TNL score range (95-115 points) classified as TD. Finally, two language samples (two spontaneous narratives from the TNL) from each participant of the remaining eligible pool of 149 were screened for length (n = 298) and samples that were less than one-minute in duration were excluded; no samples were longer than seven minutes. This resulted in an eligible pool of 92 potential language samples. Based on the remaining language samples, 50 were randomly selected to represent the 6;0-8;11 age-range. Audio recordings for this group were not utilized, as the transcription error of Google Cloud Speech on TD children was not of interest in this study, thus, only the selected transcripts were gathered.

At-Risk Language Samples

At-risk for language and literacy difficulties (AR) status was first determined through an initial screening process used in the parent study from which these language samples were sourced. This included two assessments: the comprehension subtest of the Gates MacGintie Reading Test (GMRT; MacGintie et al., 2007) and the Test of Narrative Language (TNL-2; Pearson & Gillam, 2017). Children were first administered the Gates reading comprehension subtest, and those whose scores fell below the 33rd percentile were then administered the TNL-2. Children whose scores again fell below the 33rd percentile on the TNL-2 were considered AR. For the purpose of the proposed study, additional criteria were set to ensure that language samples selected as AR were distinct from those selected as DLD. This included a specific standard score range for the TNL-2 between 83-94 points, or between $0.34 - 1SD$ units below the mean score.

Samples used in the AR corpus were sourced from a larger sample of narratives collected from a parent study of 357 children (6;0-10;5). From this study, 222 participants fell within the 6;0-8;11 age-range. Language samples of children whose first language was not English or who were bilingual were excluded, reducing the potential participant pool to 152 participants. Of the remaining participants, 87 met the criteria for AR, with the remaining classified as DLD (n = 65). Excluding the narrative retell from the TNL-2, each child told two stories across three time-points, including at pretest, posttest, and at a five-month follow-up, though some attrition at post and follow-up testing occurred. Language samples of children who participated in a treatment group were not included in the potential pool beyond pretest. All remaining language samples were screened for length, where samples shorter than one-minute in duration were excluded; no samples were longer than seven minutes. This resulted in a final eligible pool of 54 narrative samples, as most samples fell below the one-minute mark. From this pool of 54, the final 50 language samples were selected at random to represent the AR corpus in the 6;0-8;11 age-range. Both the audio and transcripts were utilized to address all the posed research questions.

Developmental Language Disorder Language Samples

Developmental language disorder (DLD) classification was based on the participants' TNL standard score. In order to be considered DLD, language samples must have been elicited from children whose standard score fell at or below 82 points, or less than 1 SD below the mean score.

Samples used in the DLD corpus were sourced from a larger pool of samples collected as a part of two parent studies ($N = 357, N = 117$). A total of 65 participants from the previously described parent study ($N = 357$) met the criteria for age-range, lingual status, and DLD classification (see the previous subsection). Each participant told two stories (two spontaneous narratives from the TNL-2) across three testing points, though samples from children in the treatment group were not considered beyond pretest. After excluding samples which were shorter than one minute, 46 narrative samples were found to meet all eligibility criteria; no samples were longer than seven minutes. The second parent study included 117 children (7;0-11;5), all of whom had been previously classified as monolingual English-speakers with DLD. As different criteria for DLD status were utilized in this parent study however, the TNL standard score range (less than 82 points) was applied to this sample. Based on the remaining criteria of DLD classification, age-range (6;0-8;11) and duration (1-7 minutes), only four language samples were found to meet all inclusionary criteria. Between both parent studies, exactly 50 samples for DLD children between 6;0-8;11 could be utilized for the current study, meaning that random selection was not implemented. Both audio and transcripts were utilized to address all the posed research questions.

Age-Range: 9;0-11;11

Typically Developing Language Samples

The same criteria were used to establish TD status (i.e., TNL standard score of 95-115 points), as previously described for the 6;0-8;11 TD group.

The typically developing (TD) corpus for the 9;0-11;11 age-range were selected from a (previously discussed) normative sample of 1,782 children aged 6;0-15;11. A total of 344 participants fell within the age range of 9;0-11;11. Language samples were excluded if they were elicited from bilingual children, children whose first language was not English, and children who had disability status(es). Based on these criteria, the language samples of a total of 47 participants were excluded, bringing the eligible pool down to 297. Participants whose TNL standard scores fell outside the specified (95-115) range were also excluded, reducing the eligible pool to 193 participants. Finally, two language samples from each participant of the eligible pool were screened for length ($n = 386$) and samples that were less than one-minute in duration were excluded. This resulted in an eligible pool of 137 potential language samples. Based on the remaining language samples, 35 were randomly selected to represent the 9;0-11;11 age-range. Only transcripts were utilized for the TD (9;0-11;11) corpus, as it was not of interest to investigate transcription error on TD language samples.

At-Risk Language Samples

AR status was again determined through an initial screening process used in the parent study, including a reading comprehension test and the TNL-2, and additional criteria based on the participant's TNL-2 standard score. The specific TNL-2 standard score range for AR classification was between 83-94 points, or between

$.34 - 1SD$ units below the mean score.

The AR corpus for the 9;0-11;11 age-range were sourced across two larger samples of narratives collected from the previously described parent studies, which collectively included 474 children between 6;0-11;5. Across both studies, 143 participants fell within the 9;0-11;11 age-range and were monolingual English-speakers. Depending on the parent study the participant was sourced from, two to six language samples were produced by each child. These language samples were screened for length, where samples shorter than one-minute in duration were excluded; no samples were longer than seven minutes. This resulted in a final eligible pool of 41 narrative samples, as most samples fell below the one-minute mark. From this pool of 41, 35 language samples were selected at random to represent the AR corpus for the 9;0-11;11 age-range. Both audio and transcripts were utilized to address all the posed research questions.

Developmental Language Disorder Language Samples

DLD classification was again based on the participants' TNL standard score. In order to be considered DLD, language samples must have been elicited from children whose standard score fell at or below 82 points, or less than $1SD$ below the mean score.

SD below the mean score. The DLD corpus for the 9;0-11;11 age-range were sourced across the two larger parent studies previously described. Across both studies ($N = 474$), 143 participants fell within the 9;0-11;11 age-range and were monolingual English-speakers. All remaining language samples were screened for DLD status and length, where samples shorter than one-minute in duration were excluded; no samples were longer than seven minutes. This resulted in a final eligible pool of 45 narrative samples, as most samples fell below the one-minute mark. From this pool of 45, a final 35 language samples were selected at random to repre-

Table 3.2: Total Language Samples by Group

Language Group	6;0-8;11	9;9-11;11	Total
TD*	50	35	85
AR	50	35	85
DLD	50	35	85
Total	150	105	255

Note. * Only transcripts, not audio.

sent the DLD corpus for the 9;0-11;11 age-range. Both audio and transcripts were utilized to address all the posed research questions.

Once all language samples were selected ($N = 255$) for each group (TD = 85; AR = 85; DLD = 85) and age-range (6;0-8;11 = 150; 9;0-11;11 = 105), samples were prepared for analysis. This began with preprocessing all audio samples ($n = 170$), sourced from the AR and DLD corpora. In the following sections, audio preprocessing and measures of audio recording characteristics are described, see Table 3.2.

Audio Preprocessing

Several steps in audio preprocessing were utilized in accordance with recommended best practice for Google Cloud Speech ASR. As previously discussed, Google Cloud Speech works best with lossless codecs (i.e., computer program used to encode/decode digital audio signal) like Free Lossless Audio Codec (FLAC) or LINEAR16, which is used in .wav, .aiff, and .raw file formats. Other "lossy" codecs (i.e., some audio information is discarded in order to make a file smaller) like MPEG (.mp3 or .m4a file formats) and Windows Media (.asf or .wma file formats) are less ideal as they tend to reduce speech recording quality. Since the original codec could not be altered, it was considered as a factor that might impact transcription error. Google Cloud Speech could only accept certain audio file formats however, so it was necessary to transcode audio files originally recorded in .mp3, .m4a, and .wma formats to a .wav format. This conversion, as well as all other preprocessing steps were completed in Adobe Audition (Version 14.2).

Once all recordings were in .wav file formats, they were converted to single channel (mono) from stereo, when necessary, and de-noised (i.e., the filtration/removal of some undesired background noise). From there, recordings were cut such that only the child's story was included; all other speech pertaining to the examiner or preamble to the story was removed. Finally, when necessary, recordings were up-sampled to a 44,100 Hz and 16-bits, such that all recordings had the same sample rate (i.e., number of audio samples recorded per second; related to perceived audio quality) and bit depth (i.e., number of bits of information in a sample; related to audio resolution). Of note, neither sampling rate nor bit-depth were evaluated as recording characteristics due to their lack of variation within the samples used in the current study.

Language Sample Recording Characteristics

Mean Opinion Score for Audio Quality Rating

In the pilot study investigating Google Cloud Speech for language sample transcription, Fox et al. (2021) excluded samples rated as having "poor" audio quality. This was done to examine the accuracy of ASR under ideal conditions to determine the feasibility of using this technology for language sample transcription. Since that feasibility has been established, in the current project, it was of interest to determine how well Google Cloud Speech performed on a range of different audio qualities, in order to better replicate how language samples may be recorded in a natural clinical environment. This determination was considered critical in order to be able to make recommendations to SLPs once this technology is ready for clinical implementation. It was therefore necessary to include samples from a range of audio qualities in order to best address the association between different characteristics of recordings and transcription error. Several measures of audio recording characteristics were utilized to examine the construct of audio quality.

The first was a subjective rating scale of audio quality called a Mean Opinion Score (MOS; ITU-T P.85, 1994), which is a commonly used metric in research on speech and audio quality. This metric was utilized to examine whether listener's perceptions of audio quality were associated with the degree of transcription error. The MOS metric utilizes a 5-point Likert scale ranging from 1, meaning "bad quality", to 5, meaning "excellent quality" and was used to rate the audio files utilized in the study. Because listener ratings of audio quality areis inherently subjective, this metric takes the average of several raters to represent each recording's MOS score, instead of treating any rating as the "true" quality score; see Appendix for full MOS scale.

A total of three independent raters were tasked with listening to all 170 audio recordings across both age (6;0-8;11 and 9;0-11;11) and language groups (AR and DLD). All raters were asked to independently listen to each recording and assign a score of 1-5. The ordering of audio recordings was randomized across each rater to prevent the effects of rater drift (i.e., changes in how a rater evaluates a construct overtime). Before beginning the scoring, raters were instructed to listen to two examples of recordings with "bad" speech quality (score of 1) and two examples of recordings with "excellent" speech quality (score of 5) in order to orient the MOS scale. Raters also met prior to beginning the task to discuss the aspects of speech quality that should be considered, based on the goals of the given study. These included aspects such as speaker volume (e.g., is the speaker loud enough to be understood without difficulty?), noise-level (e.g., does the speech sound contain static or other sound artifacts that degrade the quality of the speech sound?), and background noise (e.g., are there sounds in the background that do not pertain to the primary speaker, such as other speakers, papers shuffling, chairs moving, etc.?) which all seemed to have reasonable likelihood of impacting transcription error. Once all three raters had completed their scores, they were averaged to produce a unique MOS value for each language sample.

MOS values were not considered “true” measures of audio quality, but rather as a proxy of how an average clinician might judge the quality of a language sample recording. It was important to include this measure to gain insight into whether a subjective rating of audio quality was at all associated with how well Google Cloud Speech performed transcription. A lack of association would indicate that clinicians can likely not rely on their personal judgements of audio quality when deciding whether or not to use ASR for transcription and may therefore need to rely on more objective measures. As the MOS is inherently an aggregate metric (i.e., its an average score between several raters), interrater reliability was not calculated. However, in order to evaluate the construct validity of the MOS measure, the convergent validity between ratings was established via a Pearson correlation matrix. Correlations between each of the raters scores ranged between .71-.72, indicating convergent validity across subjective ratings of audio quality (Carlson & Herdman, 2012; Gregory, 2007).

Background Noise Rating

The second measure of audio quality was a rating of background noise presence. In order to investigate the relationship between audio quality and ASR transcription accuracy, the recordings utilized in this study varied in terms of their degree of control vs naturalness of environment. Some samples were collected in a quiet room with just the child and examiner (i.e., controlled environment), while others were recorded with other children, examiners, or background noises present while the child told their story (i.e., naturalistic environment). While MOS score provided a subjective rating of listener perceived audio quality, it was of interest to directly address the presence of background noise in a more objective manner. The relationship between background noise and transcription error was also considered as a measure of ecological validity, that could be used to inform clinical practice in recording samples.

In order to evaluate this factor, two raters independently coded for the presence of background noise within the recording. All language samples were independently double-coded as either 0 for "no background noise" or 1 "presence of background noise". Prior to coding, scorers met to discuss what constituted a code of 0 or 1, see Appendix. In addition, they met after coding the first 10 recordings to ensure alignment. While this factor was more easily operationalized as compared to audio quality generally, there was still room for subjectivity in terms of how loud and/or consistent a sound should be to be considered background noise. In order to ensure the reliability of this factor, point-by-point interrater reliability was calculated across raters, by dividing the number of agreements by the total number of recordings. Reliability was calculated at 80.59%. Any discrepancies in scores were discussed between the two raters and resolved through consensus, after which the final code was assigned.

Codec of Recording

A range of codecs (i.e., software used to encode/decode audio signal) were utilized across and within the parent studies, including WAV (e.g., .wav file formats), MPEG (e.g., .mp3 and .m4a file formats), and Windows Media (e.g., .asf and .wma

Table 3.3: Audio Recording Characteristics

Recording Characteristic	$M(SD)/N(\%)$
Age of Narrator	8;9 (1;2)
Duration (seconds)	92.7 (36.4)
Windows Media	44 (25.9%)
MPEG	74 (43.5%)
WAV	52 (30.6%)
Sampling Rate: 16,000 Hz	8 (4.7%)
Sampling Rate: 32,000 Hz	19 (11.2%)
Sampling Rate: 44,100 Hz	143 (84.1%)

Note. N = 170, 85 = AR, 85 = DLD

file formats). Guidelines related to Google Cloud Speech recommend against the usage of these lossy codecs (e.g., MPEG and Windows Media) for performing speech-to-text conversion, however, given that there was no way to change the original codec of the language samples included in this study, it became of interest to investigate the relationship between codec and transcription error. It was considered critical to communicate to SLPs which codecs to use/avoid in order to ensure optimal results, as well as address the degree to which codec was associated with transcription error. Codec information was extracted from each language sample recording automatically using the mutagen module in Python (Reiter, 2016; Van Rossum & Drake, 2007).

This information, in addition to the other audio recording characteristics, can be found in Table 3.3. In the following sections, methods pertaining to the evaluation of automated transcription, are described. This includes a discussion of the steps taken to utilize Google Cloud Speech and to prepare the ASR transcripts for analysis. A description of the gold-standard reference corpus against which ASR transcripts were compared is also summarized.

Automated Transcription Evaluation

Google Cloud Speech ASR

Once all 170 AR and DLD language sample recordings were preprocessed, they were transcribed using Google Cloud Speech. This process involved a series of steps including 1) enabling the Google Cloud Speech application program interface (API) on the Google Cloud website, 2) uploading the de-identified language sample recordings to Google's secure cloud-storage, and 3) accessing the API using the Python client for Google Cloud Speech, which additionally required 4) providing a user-specific key such that all speech-to-text conversion could be charged to the correct account. Google Cloud Speech utilizes the Listen, Attend, Spell (LAS) system on the backend to perform speech-to-text conversion, which is described in Chapter 2. Once processed, the text output for each language sample recording was compiled into a long-format dataframe containing a column for the language sample ID, story type, utterance number, and the transcribed utterance. Individual text files pertaining to each transcript were then exported from Python for conversion and then analysis. Each text file was labeled with the sampleID_story_ASR, such that each ASR transcript could be matched to the corresponding reference transcript.

Transcript Conversion Procedures

In order to prepare the ASR transcripts for error analysis, several manual steps were undertaken to convert the text so that it matched the format of the ground-truth reference corpus. This was necessary because Google Cloud Speech does not follow the same segmentation rules as the reference corpus (C-Unit segmentation), and WER could not be calculated (i.e., it will produce an error) if there were not the same number of utterances between a reference and ASR transcript. This made it necessary to manually correct punctuation within the ASR transcript to match the C-unit segmentation that was utilized in the reference corpus. In addition, it was not uncommon for ASR transcripts to be missing utterances present in their corresponding reference transcript. In order to maintain alignment between the ASR and reference transcripts, a placeholder "X" character was inserted to indicate a missing utterance, where necessary, across all ASR transcripts.

A random 20% of ASR transcripts were selected for conversion by an independent evaluator to ensure formatting was performed reliably. Point-by-point comparison resulted in an average of 92% interrater reliability. Other changes to ASR transcripts included altering spelling variants of proper nouns (e.g., Carlie → Carly) or numbers/times (e.g., 7:30 → seven thirty) to match the reference corpus. It's important to briefly note that these manual corrections to ASR transcripts were only necessary to conduct the analysis of transcription error. Such changes would not be necessary to utilize LLUNA with ASR transcripts unless punctuations were to fall in the middle of an elaborated noun phrase.

Reference Corpus

The reference corpus served as a gold-standard comparison to the ASR-produced transcripts, in order to determine the level of transcription error present. The majority of the 170 selected language sample recordings ($n = 151$) were previously transcribed in SALT as a part of their original studies. Across all studies, a portion of transcripts were selected for independent double-transcription to verify their accuracy. For the AR sample, all language samples were double-transcribed, indicating 96.16% reliability. For the DLD sample, a random 20% were double-transcribed, indicating 98.6% reliability. In each study, discrepancies were discussed between transcribers and resolved. The remaining 19 language samples were independently double-transcribed resulting in 98.91% reliability.

Several preprocessing steps were necessary to prepare the reference corpus for comparison against the ASR corpus. While Google Cloud Speech output plain-text, the reference corpus was originally transcribed in SALT and thus contained a number of unwanted annotations (e.g., mazes of repetitions, revisions, false-starts, filler words; markings for unintelligible and abandoned utterances; morpheme segmentation). These annotations, along with any examiner utterances present in the reference transcript, were removed from the reference corpus using a series of regular expressions in R. Reference transcripts produced in SALT also contained all stories told by each participant (i.e., each participant's transcript contained three stories), so it was necessary to split the transcripts into different files by stories. An open-source R function called csv2text was used to perform this process automatically (RCore-Team, 2021). All reference transcripts were then exported as text files and labeled by their participant ID and story (i.e., sampleID_story_ref).

Automated Scoring Evaluation (LLUNA)

In the next sections, the methods pertaining to evaluation of automated scoring through LLUNA are described. As a reminder, language sample transcripts were evaluated across three groups, including TD ($n = 85$), AR ($n = 85$), and DLD ($n = 85$). TD samples were included in this portion of the study to evaluate whether changes made to the pilot implementation of LLUNA resulted in gains in accuracy, as well as to help establish whether the levels of accuracy seen for TD language samples generalize to language samples elicited from children with language impairments (AR/DLD).

Reference Scores

All 255 selected language sample transcripts were hand-scored using the microstructure section of the *Monitoring Indicators of Scholarly Language* (MISL) rubric, to serve as the reference (“ground truth”) scores (see Chapter 2). These reference scores were compiled to determine the accuracy of scores generated by LLUNA, based on their correspondence with trained MISL hand-scorers. All 85 TD language samples ($n = 85$) were scored as a part of a previous study. MISL scorers received training on how to identify literate language structures in narrative language samples. Scorers were required to achieve at least 85% point-by-point interrater reliability with an expert scorer across five language samples. All TD language samples were independently double-scored, and achieved a point-by-point interrater reliability of 85.2%. The remaining 170 language samples for the AR and DLD groups were scored by two trained scorers. In order to ensure the reference scores were closest to the ground truth, and to be consistent with the manner in which the samples for TD students were scored, 100% of the samples were double-scored. This both established the interrater reliability level between hand-scorers and allowed for the correction of errors through consensus between the two scorers. Separate rounds of scoring were completed for each age-range after between 35-50 narratives had been scored, to ensure rater drift did not occur over time.

Interrater reliability for each of the six literate language indices was evaluated through a weighted kappa so that interrater reliability levels between hand-scorers could be compared to LLUNA scoring accuracy levels. For the 6;0-8;11 age-range quadratic weighted kappa values ranged from .62-.92, with adverbs once again standing out as the least reliably scored index. For the 9;0-11;11 age-range, a similar pattern was seen with quadratic weighted kappa values ranging between .65-.95. See Table 3.4 for the interrater reliability information on each of the six mea-

Table 3.4: MISL Hand Scorer Interrater Reliability

Index	6;0-8;11 (κ_{qw})	9;0-11;11 (κ_{qw})
Coordinating Conjunctions	.91	.95
Subordinating Conjunctions	.79	.73
Mental Verbs	.92	.92
Linguistic Verbs	.90	.87
Adverbs	.62	.65
ENP	.83	.76

Note. 6;0-8;11 n = 100, 9;0-11;11 n = 70. k = quadratic weighted kappa.

sures. These levels of interrater reliability were on par, and in most cases slightly higher, than what was reported for MISL hand-score reliability in Fox et al. (under review), where kappas ranged from .52-.86.

The interrater reliability of hand-scored adverbs was consistently lower than the other literate language indices, which made it necessary to resolve discrepancies between the two scorers to ensure LLUNA generated scores were being compared to a set of reference scores representative of the ground truth. This required the two scorers to meet and resolve discrepancies through consensus. While adverbs were the greatest source of discrepancies, other literate language indices were also reviewed to ensure the accuracy of the reference scores. Once the scorers came to an agreement about each of the discrepancies, the finalized MISL scores were saved in a separate Excel spreadsheet for use in the analysis phase as the ground truth reference scores.

LLUNA Modifications

Several modifications were made to LLUNA with the goal of potentially increasing its scoring accuracy for use with students between the ages of 6;0-11;11. To ensure that the accuracy of LLUNA scores was generalizable and not, in a sense, "over-fit" to the language samples utilized in the study after making modifications, the reference transcripts and their corresponding MISL scores were randomly split into a validation ($60\%, n = 100$) and test ($40\%, n = 70$) set. The validation set was used to establish the initial accuracy of LLUNA on the reference transcripts, and then used to re-evaluate this accuracy once adjustments had been made. The "test set" was used to answer the questions posed in the project.

The first change made to LLUNA was expanding wordlists to include higher level vocabulary for mental and linguistic verbs, as well as subordinating conjunctions. In the pilot study of LLUNA, a younger age-range (5;0-9;11) was utilized with a significantly smaller sample size ($N = 50$). Narrative samples in the current project's validation sets contained more instances of complex vocabulary including words like *insisted, realized*, and *however* that were not previously included amongst the predefined wordlists in the pilot study. Examples of more complex vocabulary words were added to make LLUNA more generalizable to an older age-range.

The pilot implementation of LLUNA utilized an older part-of-speech tagger available through the OpenNLP package in R (Hornik, 2019). In this study, a newer, updated tagger implemented through the Python programming language was used to help ensure that the model was state-of-the-art (Bird, 2006; Van Rossum & Drake, 2007).

In the pilot study, it was also determined that LLUNA sometimes misclassified elaborated noun phrases identified in the reference transcripts. For example, Adverb + Determiner + Adjective + Noun = *quickly the nervous boy* from *he ran quickly the nervous boy said.* In addition, its syntactic scoring scheme was missing permutations containing possessives, which led to underscoring phrases like *the principal's office* as an ENP of one (*the principal*) instead of a two. Possessives were therefore added to the syntactic scoring scheme that was used in LLUNA to assign ENP scores.

Each time LLUNA was modified, accuracy on the validation set was re-calculated to see if gains in accuracy had been made. Once modifications resulted in no further gains in accuracy, LLUNA's design was finalized. The test set was used to determine whether the validation accuracy generalized to a new set of language samples (also from the reference corpus). The test set accuracy was then used to address research questions three and four.

Data Analysis

The current study posed six research questions to address three posed aims. The data analyses used to address each of these questions and their associated aims are split into separate subsections below and include 1) automated transcription accuracy, 2) automated scoring accuracy, and 3) streamlined transcription and scoring accuracy.

Automated Transcription Accuracy

The first research question aimed to determine the level of transcription error produced by Google Cloud Speech ASR on the oral narratives of school-age children (6;0-11;11) who had impaired language abilities (AR and DLD). In order to measure transcription error, a weighted word-error rate (WER_w). WER is the standard accuracy metric for evaluating ASR systems (Park et al., 2008). The WER was calculated by summing the number of insertions (i.e., adding in an additional word not present in the reference transcript), substitutions (i.e., replacing a word from the reference transcript with another), and deletions (i.e., removing a word present in the reference transcript) and then dividing by the total number of words in the reference transcript.

The WER is calculated at the clause level, so when examining the WER of an entire transcript, one can report the WER_w in order to account for the length of each clause when calculating the overall WER. By using a WER_w, omitting an entire clause that is only a couple words long is not weighted the same as omitting a clause that is 10 or more words long. The WER_w acts as a weighted average of the WER pertaining to each clause. This metric is calculated by multiplying each individual clauses' WER (WER_{c_i}, where $i \in 1, \ldots, n$ and n is the number of clauses) by the number of words in that clause (L_{c_i}) divided by the total words in the transcript (L_T). Each weighted value (one per clause) is then summed to produce the weighted WER (WERw), see Equation 3.1 below.

$$WER_w = \sum_{i=1}^{n} WER_{c_i} \frac{L_{c_i}}{L_T} \tag{3.1}$$

Clause-level WER was computed using an open-source WER function in Python (Van Rossum & Drake, 2007). An additional function was then created by the author in Python to calculate the weighted average based on clause length in accordance with Equation 3.1, resulting in the WER_w for each transcript. The possible range of WER_w is 0-100%, where 0% indicated perfect transcription with no error and 100% indicated all words were incorrectly transcribed and/or deleted.

While it was expected that transcription error might vary based on language sample (e.g., language status, age) or recording characteristics (subjective audio quality, background noise, codec), hypothesis testing was not implemented to address research questions one or two because of the exploratory nature of the current project. In order to avoid p-hacking (i.e., inadvertently or purposefully running statistical tests until significant or interesting findings are obtained) or other erroneous conclusions regarding causality, no statistical tests were run or considered necessary to address these research questions.

Instead, the WER_w of the overall sample is first presented through descriptive statistics (mean, median, SD, range) and as a histogram illustrating the distribution of WER_w, which addressed the first research question. Then, visualizations of scatter plots stratified by different language sample characteristics were used to evaluate associations with transcription error to address research question two. Importantly, these visualizations and descriptive statistics could not be used to establish causal relationships.

Automated Scoring (LLUNA) Accuracy

The third research question aimed to determine whether modifications to LLUNA would be associated with increased scoring accuracy on a new corpus of TD language samples. The fourth research question then addressed whether these accuracy levels would then generalize to language sample transcripts elicited from children with impaired language abilities (AR and DLD). The primary goal of these analyses was to establish the accuracy of scores generated automatically by LLUNA, which was accomplished through evaluating the interrater reliability between LLUNA generated and reference (ground truth) scores, using a quadratic-weighted kappa metric.

Quadratic weighted kappa K_{qw} is a metric commonly used in analyzing the interrater reliability between automated scoring systems and hand-scores (Dikli, 2006). It is a recommended metric for multi-class classification problems because K_{qw} is able to take the distance between classes into account, meaning that a difference of one point between the reference and CAMS generated (i.e., observed) scores is not weighted the same as a difference of two or three points (Ben-David, 2008). In addition, K_{qw} weights the probability of correct classification by chance alone

(Cohen, 1968). K_{qw} is given by Equation 3.2 below:

$$K_{qw} = \frac{P_o - P_e}{1 - P_e} \tag{3.2}$$

Where P_o represents the proportion of observed agreement, given by Equation 3.3:

$$P_o = \sum_i \sum_j W_{ij} Pij \tag{3.3}$$

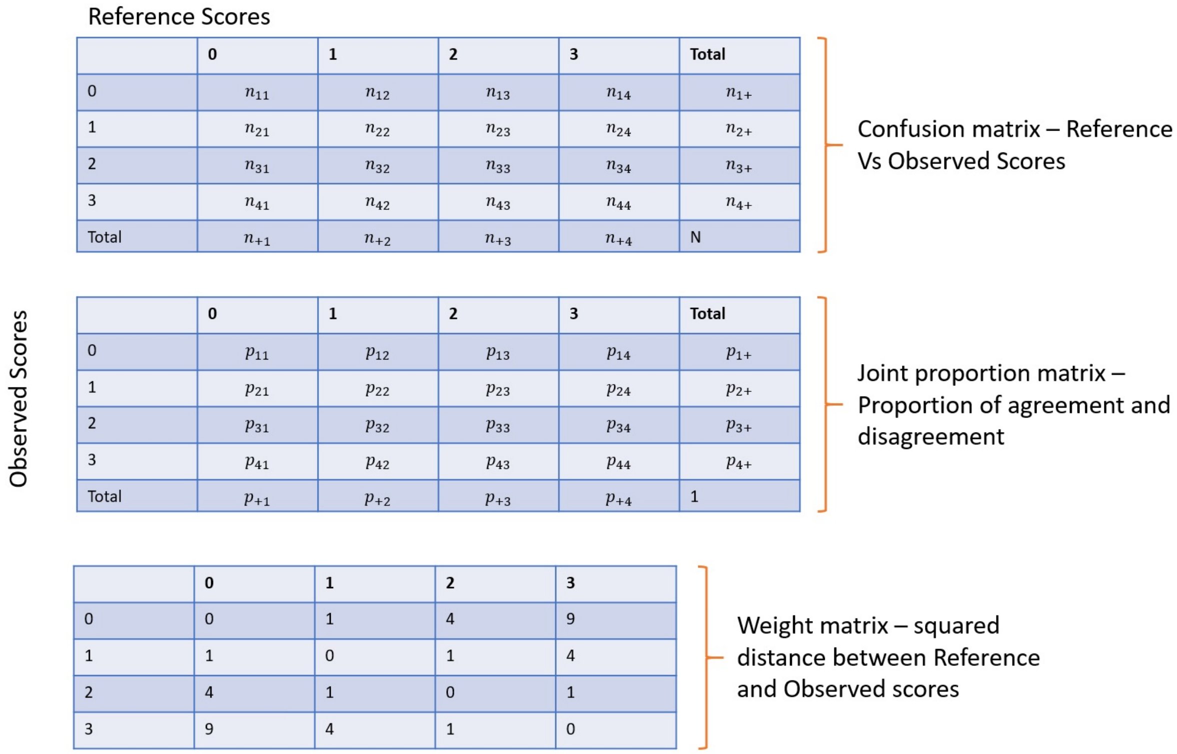
Here, W_{ij} is a quadratic weight matrix with values ranging between zero and nine. The diagonal of the weight matrix are zeros, which corresponds to scorer agreement. The farther off the diagonal, the higher the weight value, which corresponds to the squared distance of the scorer disagreement; such that a disagreement of one point receives a weight of one, two points a weight of four, and three points a weight of nine. P_{ij} is a joint proportion matrix where the value of each cell is calculated by dividing the cell count of the ith row and the jth column from the confusion matrix of reference and observed scores by the total number of cells N, see Figure 3.1.

P_e represents the proportion of expected chance agreement, given by Equation 3.4:

$$P_e = \sum_i \sum_j W_{ij} P_{i+} P_{+j} \tag{3.4}$$

Which is the sum of the weighted product of row (i) and column (j) marginal proportions.

The possible values of K_{qw} range between zero (meaning no agreement) to one (meaning perfect agreement). Within the automated scoring literature, a K_{qw} of .60 or above is considered an acceptable level of interrater reliability (Dikli, 2006). A more meaningful comparison was the interrater reliability between trained MISL


Reference Scores
Observed Scores
Confusion matrix – Reference Vs Observed Scores
Joint proportion matrix – Proportion of agreement and disagreement
Weight matrix – squared distance between Reference and Observed scores

	0	1	2	3	Total
0	n_{11}	n_{12}	n_{13}	n_{14}	n_{1+}
1	n_{21}	n_{22}	n_{23}	n_{24}	n_{2+}
2	n_{31}	n_{32}	n_{33}	n_{34}	n_{3+}
3	n_{41}	n_{42}	n_{43}	n_{44}	n_{4+}
Total	n_{+1}	n_{+2}	n_{+3}	n_{+4}	N

	0	1	2	3	Total
0	p_{11}	p_{12}	p_{13}	p_{14}	p_{1+}
1	p_{21}	p_{22}	p_{23}	p_{24}	p_{2+}
2	p_{31}	p_{32}	p_{33}	p_{34}	p_{3+}
3	p_{41}	p_{42}	p_{43}	p_{44}	p_{4+}
Total	p_{+1}	p_{+2}	p_{+3}	p_{+4}	1

	0	1	2	3
0	0	1	4	9
1	1	0	1	4
2	4	1	0	1
3	9	4	1	0

Figure 3.1: Confusion, joint, and weight matrices for calculating K_{qw}.

hand-scorers, as well as to the previously reported kappas from the LLUNA pilot study. Scoring accuracy was evaluated separately on both the validation and test set for the TD group, which addressed research question three, and then the language impaired group (AR/DLD), which addressed research question four. These kappa values are presented in tables to allow for comparison between the accuracy levels of the modified LLUNA version created in the current study and accuracy levels of LLUNA and hand-scorers from the pilot study.

Automated Transcription and Scoring Accuracy

The final research questions aimed to determine the feasibility of combining ASR and LLUNA to streamline automated transcription and scoring. The fifth research question addressed the degree to which LLUNA scoring accuracy varied by transcription error, while the sixth research question as determined whether there was an acceptable range of transcription error that maintained accurate LLUNA scores. As with research questions three and four, scoring accuracy was evaluated via Kqw. These kappa values indicated the accuracy of LLUNA scores generated from ASR transcripts (ASR + LLUNA), as determined by their overlap with the reference (ground truth) scores. Once kappa values had been calculated, an array of box-plots were used to visualize association between the transcription error (word-error rate) and the LLUNA's scoring error (i.e., the exact difference between the LLUNA and reference score) for each literate language measure. These plots provided insight towards the median and interquartile range of transcription error associated with accurate LLUNA scores (i.e., a difference of 0).

CHAPTER 4

Results

Research Questions 1 & 2: ASR Transcription Error

The first research question aimed to determine the degree of transcription error produced by Google Cloud Speech on the language sample recordings of school-age children (6;0-11;11) who have impaired language abilities (AR and DLD). This was evaluated using a weighted word-error rate.

Out of all 170 narrative language samples, which included both age-groups (6;0-8;11 and 9;0-11;11) and both levels of language impairment (AR and DLD), the mean WER_w was 48% (SD = 27%). The median WER_w of 40%, was more representative of the distribution given its right skew, however. The range of scores was large, covering most of the possible values (i.e., 0-100%) with WER_w values falling between 8-100%, see Figure 4.1. This meant that within the sample of 170 narratives, some were transcribed nearly perfectly by Google Cloud Speech, while others were not transcribed at all (i.e., 0% = perfect classification, 100% = complete misclassification). Specifically, 67 (39.3%) of the samples had a WER_w at or below 33%, 56 (32.7%) samples had a WER between 34-66%, and 47 (28%) samples had a WER_w of 67% or higher.

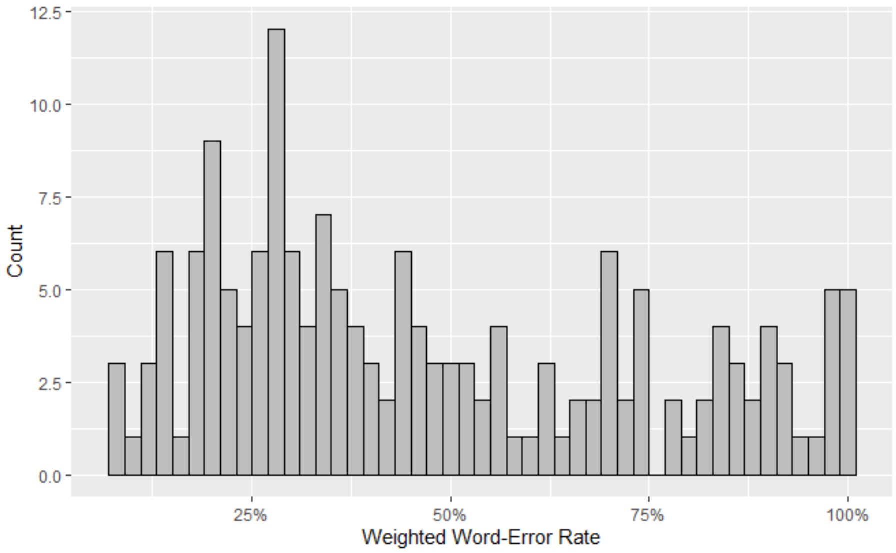


Figure 4.1: Weighted Word-Error Rate (WER_w) across all narrative samples, $N = 170$. Bin-width = .02.

Table 4.1: Descriptive Statistics of Word-Error Types within Utterances

Error Type	$M(SD)$	Median	Min	Max
Deletions	2.37 (3.33)	1	0	26
Substitutions	1.28 (1.45)	0	0	11
Insertions	0.12 (0.49)	1	0	11

Note. Deletions, substitutions and insertions are summed to calculate the word-error rate.

Error analysis was used to evaluate break down of transcription error by word-error type, including deletions, insertions, and substitutions, see Table 4.1. On average, deletions were the most common type of word-error per utterance ($M = 2.37, SD = 3.33$), followed by substitutions ($M = 1.28, SD = 1.45$), and then insertions ($M = 0.12, SD = 0.49$). Deletions also had the widest range, with between 0-26 words deleted per utterance, while the range for substitutions and deletions only ranged between 0-11 words per utterance. Deletions were therefore the most common and most variable type of word-error.

In the current study, ages ranged from 6;0-11;11, but were split into two distinct age-groups since stages of language development vary heavily amongst this

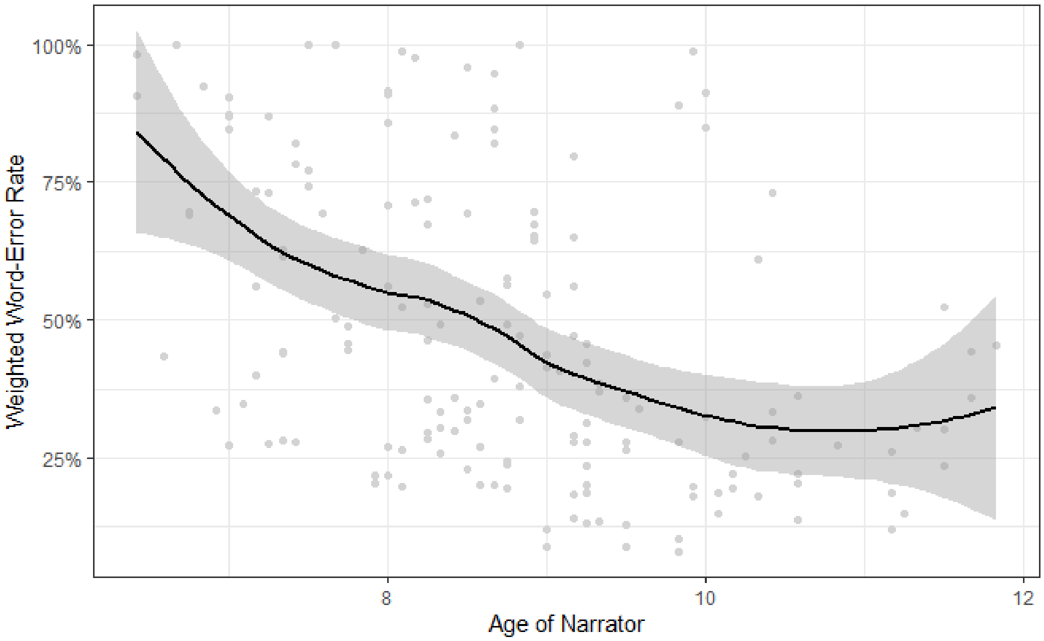


Figure 4.2: Scatter plot of age of narrator by WER_w, fit with a quadratic line for visual aid.

range. Age was therefore treated as both a continuous (all ages) and categorical (age-groups) variable. An examination of the transcription error by (categorical) age-group revealed that the mean WER for narrative samples from the 6-0-8;11 age-range was 58% (SD = 26%), while the mean WER for the 9;0-11;11 age-range was 33% (SD = 21%). Similarly, when age was plotted as a continuous variable against WER there was a decreasing trend in WER as age increased that leveled out at the older ages (10-12 years-old), see Figure 4.2.

While age appeared to play a role in transcription error, it is important to highlight two points. The first is that the scatter plot was noisy (i.e., high variance in error), which called into question the representativeness of the fitted line. The second, was that samples were collected with variable recording characteristics that were not evenly distributed amongst age-groups. For example only seven samples from the 6;0-8;11 age-range (7%) were recorded with a lossless codec (i.e., WAV) as compared to 65.7% (n = 46) of the samples from the 9;0-11;11 age-range. This was addressed by research question two, which was to determine whether different recording characteristics (perceived audio quality, background noise, and codec) were associated with transcription error. Age remains on the x-axis of the following visualizations to help clarify whether the role of age on transcription error held when accounting for other factors.

The first language sample recording characteristics of interest was the subjective audio quality rating. This was measured via the mean opinion score (MOS) for each recording. The average MOS across all 170 narrative samples and was 3.7 ($SD = 0.9$), which fell amongst the more moderate range of perceived audio quality, with a score of four indicating “good” audio quality and a score of three indicating “medium” audio quality. The average MOS ratings did not differ across age-groups (6;0-8;11 and 9;0-11;11) as both groups had an average MOS of 3.7 ($SD = 1.0; SD = 0.8$, respectively). A scatter plot of the relationship between transcription error and MOS indicated a potentially small trend of higher MOS ratings being associated with lower WER, however, there was again high error variance see Figure 4.3. This is impacted by the imbalance of ratings, with few samples receiving a score of two or lower.

Based on these findings, perceived audio quality, as measured by MOS, did not provide much insight into transcription performance by Google Cloud Speech. A reevaluation of the perceived audio quality ratings as a Median Opinion score was more meaningful because it allowed for the treatment of the score as a true categorical variable. As can be seen in Figure 4.4, there was a negative trend between median opinion scores and transcription errors. It was noted, however, that there were several distinct outliers in transcription error when the median opinion score was five, meaning “excellent” audio quality.

Stratification of the boxplot in Figure 4.4 by codec helped explain these outliers, see Figure {fig:figure18}. This boxplot provided evidence that the noise (i.e.,

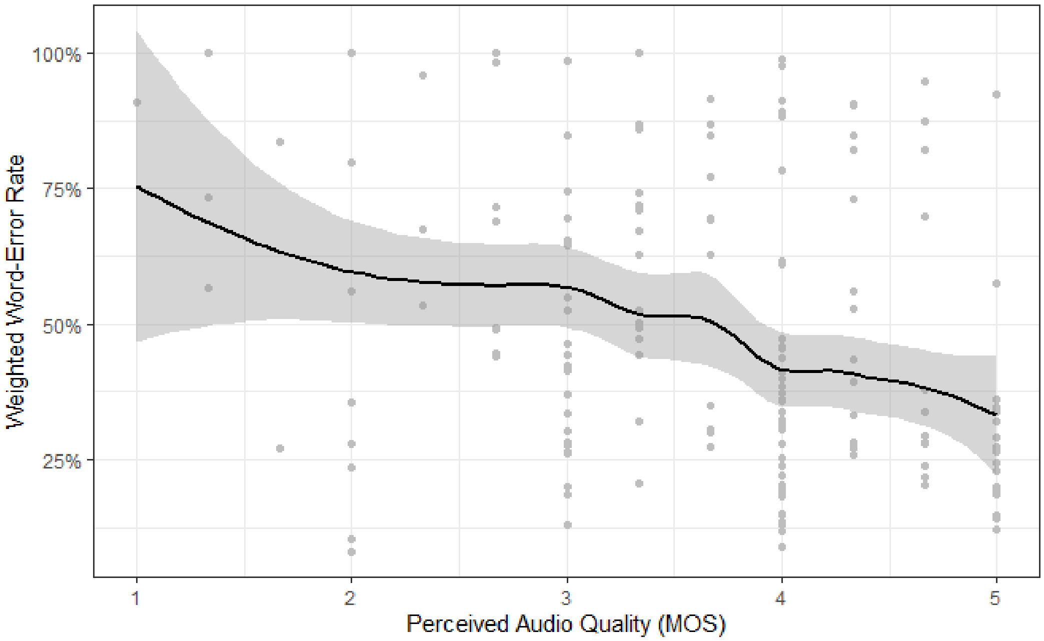


Figure 4.3: Scatter plot of mean opinion scores (MOS) by weighted word-error rate.

high error variance) present within the relationship between perceived audio quality (MOS) and transcription error was attributable (at least in part) to the codec used. For both the Windows Media (ASF) and WAV codecs, there was not a clear relationship between perceived audio quality and transcription error.

When examining WAV, for example, the median WER for samples whose audio quality was rated as a 2, meaning "poor" was lower than the of scores 3-5 (representing higher perceived audio quality). Similarly, there was no pattern of association between WER and perceived audio quality (is this MOS) for narratives that were recorded using the Windows Media codec, which applied high levels of compression. Only samples recorded using an MPEG codec, which also applies compression to recordings, revealed a negative trend between perceived audio quality and transcription error (i.e., the higher the perceived quality the lower the transcription error).

Because codec appeared to play an important role in evaluating the relationship between listener perceived audio quality and transcription error, it was also

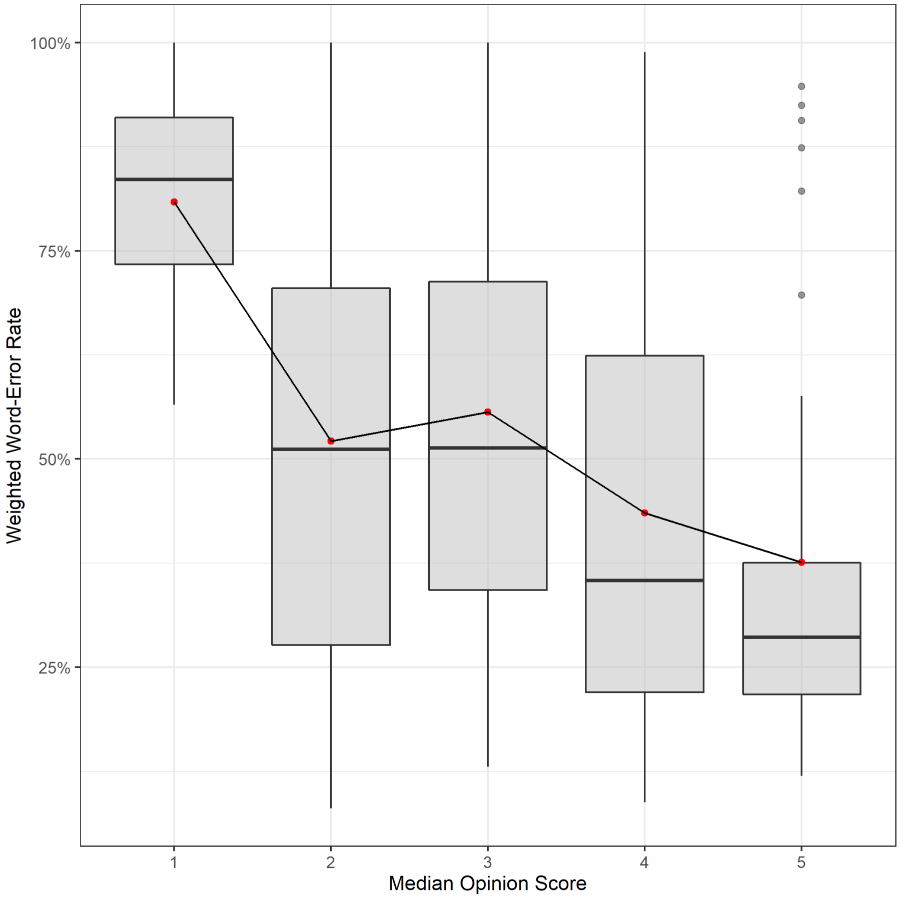


Figure 4.4: Boxplot of median opinion scores by transcription error, as measured by weighted word-error rate. The plotted line shows the mean word-error rate value associated with each median opinion score.

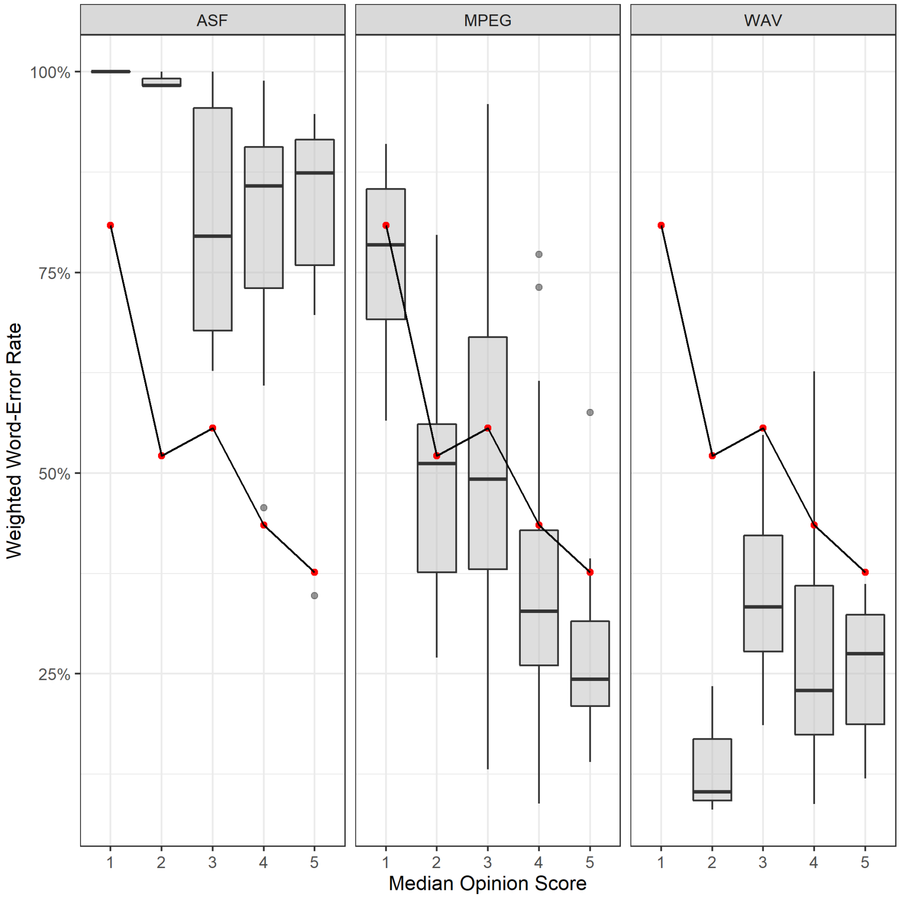


Figure 4.5: Boxplot of transcription error by median opinion score, facetted on codec. ASF = Windows Media.

used in a facetted plot to examine the relationship between background noise and transcription error.

In Figure 4.6, it can be seen that among Windows Media (ASF) codecs, narrative samples rated as “not containing background noise” had a slightly lower transcription error rate ($M = 78\%, SD = 18\%$) and a small, decreasing trend of WER as age increased. The same trend was observed for files recorded using Windows Media (ASF) when background noise was present ($M = 85\%, SD = 13\%$). MPEG samples were associated with lower transcription errors when samples were rated as not containing background noise ($M = 32\%, SD = 17\%$) as compared to those rated as containing background noise ($M = 48\%, SD = 21\%$). The strongest trend between age and WER was present for the MPEG recorded files rated as containing background noise, with WER decreasing as age increased. Finally, for narrative samples recorded with WAV, too few samples were rated as having background noise to make a meaningful comparison, but there was no relationship between age and transcription error when no background noise was present.

Another factor which had the potential to impact transcription error was the degree of language impairment of the speaker (AR, DLD). Overall, results indicated that there was only a small difference in the transcription error between the two language ability groups, where the average WER_w of AR language samples was 46% ($SD = .27$) and the average WER_w of the DLD language samples was 50% ($SD = .27$).

Research Questions 3 & 4: LLUNA Scoring Accuracy

Research question three aimed to determine whether modifications to LLUNA were associated with higher scoring accuracy (as measured by quadratic weighted kappa) than were reported in the pilot study when applied to samples obtained

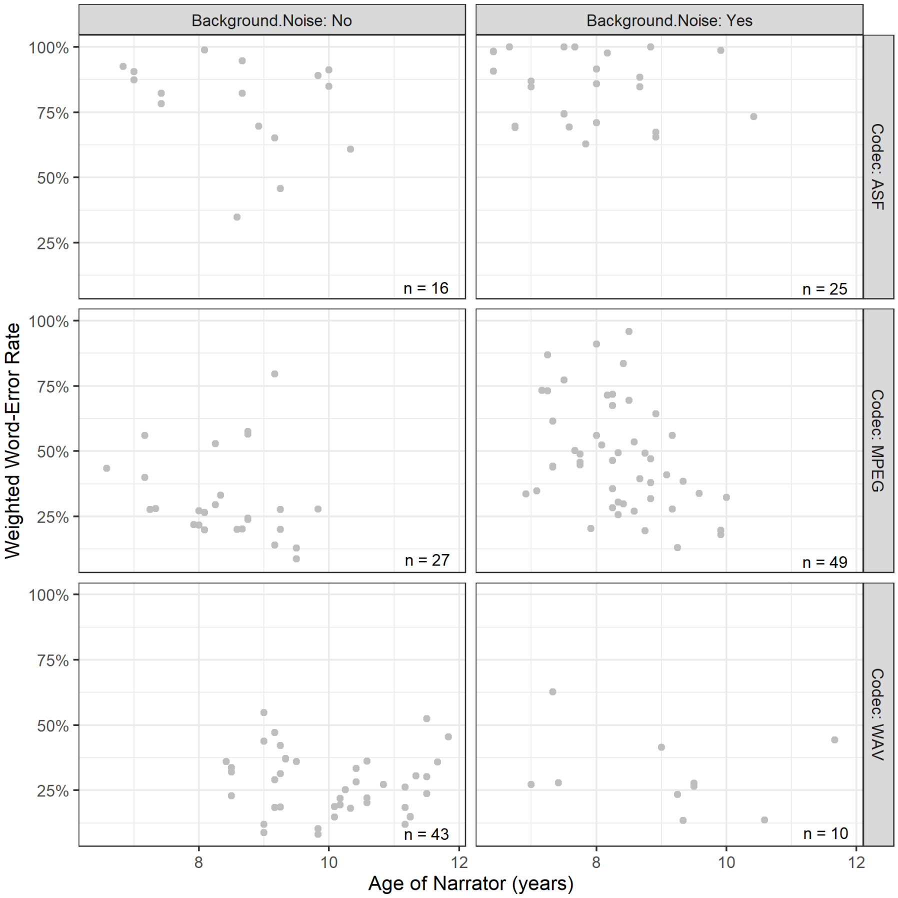


Figure 4.6: Scatter plot of WER by Age of Narrator, faceted by codec and background noise rating.

Table 4.2: Validation and Test Accuracies of LLUNA for TD (κ_{wq})

Index	$LLUNA_{2.0v}$	$LLUNA_{2.0t}$	$LLUNA_{1.0}$	Hand-Scores
CC	.94	.93	.78	.86
SC	.84	.90	.88	.71
Mental Verbs	93	.92	.89	.83
Linguistic Verbs	.96	.91	.89	.75
Adverbs	.95	$*$	.79	.52
ENP	.83	.75	.74	.78

Note. Validation and Test Accuracies of LLUNA for TD Narrative Samples, κ_{qw} = quadratic weighted kappa,$LLUNA_{2.0v}$ = modified LLUNA validation set, $LLUNA_{2.0t}$ = modified LLUNA test set. CC = coordinating conjunctions, SC = subordinating conjunctions, ENP = elaborated noun phrase. Validation set, $n = 51$, Test set, $n = 34$. *κ_{qw} was 0 because all ground truth scores in the validation set were threes. Point-by-point accuracy for adverbs was 97.06%. Pilot LLUNA and Hand kappa values taken from Fox et al. (under review), $N = 50$.

from school-age TD children. Research question four evaluated whether potential scoring accuracy gains were maintained in language samples obtained from school-age children with DLD or who were AR for language and literacy difficulties.

The first table presents the accuracy scores generated from the modified LLUNA (LLUNA 2.0) on the reference transcripts of the TD corpus ($n = 85$) across all ages (6;0-11;11), see Table 4.2. This provides a comparison against the pilot findings from Fox et al. (under review), which evaluated a normative sample of 50 children ages 5;0-9;11. In this case, the two age-ranges were not evaluated separately since the primary goal of including the TD narrative samples was to determine whether the pilot findings could be replicated and/or improved on a new set of similar language samples (i.e., both TD) after LLUNA had been modified, and also generalize to a slightly older age-range (6;0-11;11). In addition, the kappas observed between trained and expert hand scorers from the pilot study were also included for comparison of the reliability of LLUNA scores against trained hand-scorers.

The scoring accuracy of modified LLUNA was generally similar amongst the validation and test sets of the TD samples, though the kappa observed for ENP in the test set was .08 points lower than in the validation set, see Table 4.2. Test set accuracies of the modified LLUNA ranged between .75-.93, meaning all measures were generated with acceptable levels of accuracy ($\kappa_{qw} > .60$). The test set accuracies of the modified LLUNA in the current study ($\kappa_{qw} = .75 - .93$) were also on par or higher than those obtained by the older version of LLUNA used in the pilot study ($\kappa_{qw} = .74 - .89$). Coordinating conjunctions had the largest difference, where the scoring accuracy of the modified LLUNA was .15 points higher than the pilot LLUNA version. ENP had the smallest difference between the modified and pilot LLUNA, with accuracy increasing by .01 point in the current project.

The modified LLUNA also demonstrated greater levels of reliability than scores obtained by the trained hand-scorers in the pilot study. ENP was the only language measure that continued to be associated with lower levels of scoring reliability ($\kappa_{qw} = .75$) than the those obtained by trained scorers ($\kappa_{qw} = .78$), though the difference in reliability levels was minimal at .03 points.

Research question four addressed whether accuracy levels observed for the modified LLUNA generalized to the narratives of school-age (6;0-11;11) children with impaired language abilities (AR/DLD). Score accuracy is presented for both the validation and test sets in Table 4.3. In addition, score accuracy was evaluated on all ages (6;0-11;11), as well as on separate age-ranges (6;0-8;11 and 9;0-11;11) to examine whether there were any differences in accuracy by age-group amongst AR and DLD samples, see Table 4.3. The interrater reliability levels seen between the two raters in the current project were also included for reference, in order to compare LLUNA's reliability against that of trained hand-scorers.

Generally, the scoring accuracy of LLUNA on the AR/DLD test set was high, with kappas ranging .67-1. The accuracy of adverbs in the test set ($\kappa_{qw} = .67$) was lower than what was seen for the TD test set (point-by-point accuracy = 97.06%) or in the pilot study ($\kappa_{qw} = .79$), however, it still met acceptable levels of accuracy

Table 4.3: Test Set Accuracy of LLUNA on AR and DLD

Index	Valid: All ages ($n = 102$)	Test: All ages ($n = 68$)	...
CC	.88	1	...
SC	.78	.97	...
Mental Verbs	.86	.91	...
Linguistic Verbs	.85	.87	...
Adverbs	.77	.67	...
ENP	.82	.81	...
Index	Valid: 6;0-8;11	Test: 6;0-8;11	Hand: 6;0-8;11
CC	.87	1*	.91
SC	.76	.98*	.79
Mental Verbs	.78	.93*	.92
Linguistic Verbs	.79	.91*	.90
Adverbs	.88	.66*	.62
ENP	.74	.84*	.83
Index	Valid: 9;0-11;11	Test: 9;0-11;11	Hand: 9;0-11;11
CC	.91	1*	.95
SC	.78	.97*	.73
Mental Verbs	.95	.88	.92*
Linguistic Verbs	.89	.81	.87*
Adverbs	.66	.69*	.65
ENP	.90	.70	.76*

_Note._* indicates the highest kappa between the LLUNA test set and the hand-scores. Hand = interrater reliability (as measured by quadratic weighted kappa) between two trained hand-scorers.

(Dikli, 2006). An evaluation of scoring accuracy split on the different age-ranges (6;0-8;11 and 9;0-9;11) indicated the LLUNA test set scoring accuracy between the two age groups was comparable on coordinating conjunctions, subordinating conjunctions, mental verbs, and adverbs, only differing by a few points. Greater differences in accuracy between the 6;0-8;11 and 9;0-11;11 age-ranges were seen in the test sets for linguistic verbs and ENP. In both cases, the 6;0-8;11 age-range had higher score accuracy, with the kappa of linguistic verbs being .10 points higher and the kappa of ENP being .14 points higher for the younger age-range.

This finding was also reflected in the comparison of LLUNA kappa levels to hand-score kappa levels. LLUNA test set kappas for both age-ranges (6;0-8;11 and 9;0-11;11) were consistently on par or higher than those observed between hand-scorers for coordinating conjunctions, subordinating conjunctions, and adverbs. For the 6;0-8;11 age-range, LLUNA also had kappas that were on par with or higher than the hand-scorers for mental verbs, linguistic verbs and ENP. The LLUNA kappas for the 9;0-11;11 age-range on linguistic verbs and ENPs were lower than hand-scorers, however, falling 0.4-.06 points below hand-scorers on these indices. It is worth noting the difference in sample size between the 9;0-11;11 ($n = 23$) test set and the hand-scores ($n = 70$), which may have contributed to both the variance in kappa scores between the different age groups and the lower kappa values seen for the 9;0-11;11 test set.

Research Questions 5 & 6: Accuracy of ASR + LLUNA

Research questions five and six addressed the degree to which LLUNA scoring accuracy was impacted by ASR transcription errors. Research question five evaluated how LLUNA score accuracy varied across ASR transcription error levels. ASR transcripts were not split into a validation and test set since they were not used to-

Table 4.4: Accuracy of LLUNA on All ASR Transcripts

Index	ASR-Reference Score (κ_{qw})
Coordinating Conjunctions	.52
Subordinating Conjunctions	.52
Mental Verbs	.62
Linguistic Verbs	.59
Adverbs	.27
Elaborated Noun Phrase	.48

Note. $n = 170$. κ_{qw} = quadratic weighted kappa between LLUNA generated scores on ASR transcripts and the reference (ground truth) MISL scores.

wards modifications of LLUNA, and only the final version of LLUNA was used in determining the accuracy of scores on ASR produced transcripts. Further, no ASR transcripts were 100% accurate in their transcription, meaning that all transcripts in the ASR corpus were at least somewhat different (and often times quite different) from transcripts in the reference corpus validation and test sets.

Table 4.4, shows the LLUNA scoring accuracy for each literate language index across all 170 ASR transcripts, as measured by quadratic weighted kappa (κ_{qw}).

Scoring accuracy was generally poor across all literate language indices, with kappas ranging between .27-.62. The highest level of interrater reliability between LLUNA scores generated from ASR and reference transcripts was for mental verbs ($\kappa_{qw} = .62$), which surpassed the threshold of acceptable accuracy by .02 points. The lowest level of interrater reliability was observed for adverbs ($\kappa_{qw} = .27$). As can be seen in Figure 4.7, accurate LLUNA scores (i.e., difference score of 0) were associated with lower levels of transcription error. LLUNA scores were underestimates of the reference score (i.e., negative difference score) were associated with higher levels of transcription error by comparison. Interestingly, LLUNA scores that were overestimates (i.e., positive difference score) were associated with similar levels of transcription error as those seen for accurate LLUNA scores.

Research question six aimed to determine whether there was an acceptable range of transcription error associated with accurate LLUNA scoring. To address this question, the level of transcription error was plotted against LLUNA score error levels (i.e., exact difference scores) for each literate language index in six separate box-plot visualizations (see Figure 4.7), such that the median and interquartile ranges of transcription error associated with each difference score could be observed.

For each literate language index, the median transcription error associated with a difference score of 0 fell within the 25-50% WER range. The interquartile ranges (i.e., the middle 50% of the distribution) of transcription error associated with a difference score of 0 were more variable. For coordinating conjunctions and adverbs the interquartile ranges of transcription error were small, with a WER between about 25-50% being associated accurate LLUNA scores. The remaining indices had wider interquartile ranges associated with accurate LLUNA scores, including WERs between about 25-60%.

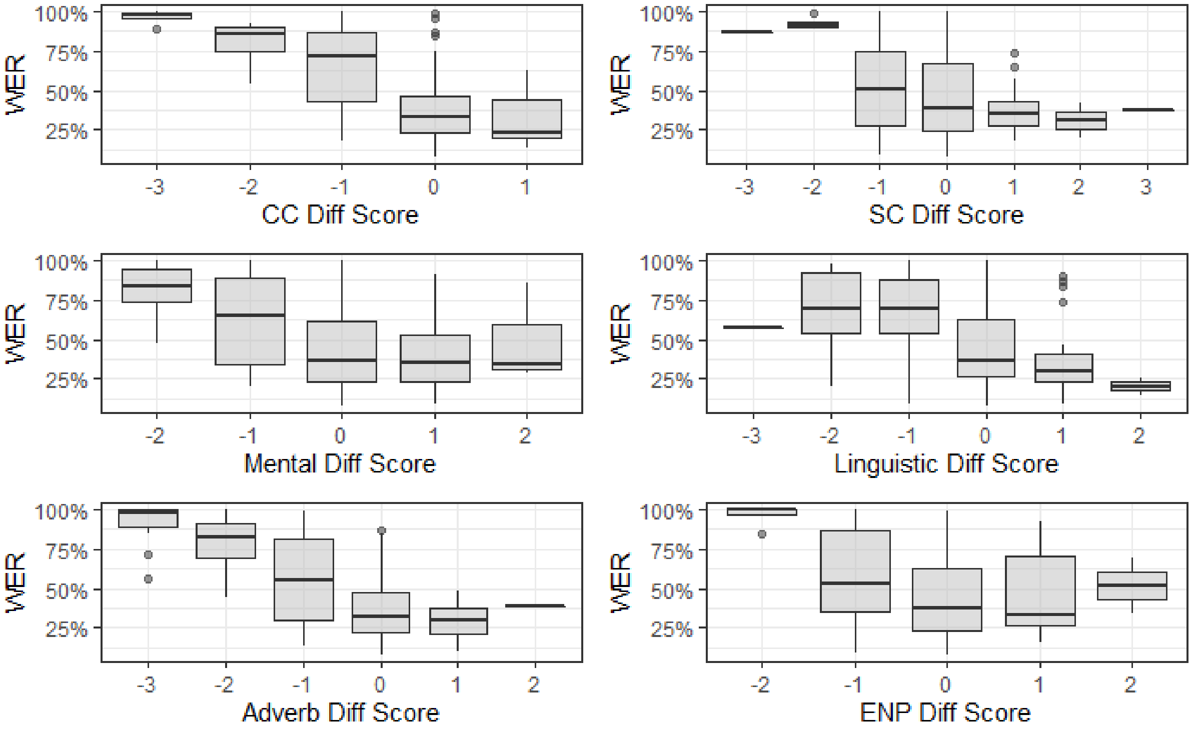


Figure 4.7: $n = 170$. CC = Coordinating conjunctions, SC = Subordinating conjunctions, Mental = mental verbs, Linguistic = linguistic verbs, ENP = elaborated noun phrase. Diff Score represents the difference between the LLUNA generated score from an ASR transcript and its corresponding ground truth score.

CHAPTER 5

Discussion

The purpose of the current study was to build upon the findings from two prior pilot studies where separate pieces of the language sample analysis process were automated, specifically 1) transcription and 2) scoring of selected literate language indices, with the primary goal of assessing the feasibility of streamlining the entire language sample analysis process by combining the two. Several research questions were posed to address these aims. The discussion is divided into sections and subsections by aims and research questions below.

Aim 1: To Assess the Accuracy of Automated Transcription

Research Q1

The first research question assessed the degree of error present within narrative language samples of school-age children (6;0-11;11) classified as either at-risk for language and literacy difficulties (AR) or having developmental language disorder (DLD) transcribed by Google Cloud Speech, a cloud-based automatic speech recognition (ASR) system. To investigate the accuracy of ASR produced transcripts for this population, a total of 170 oral narrative samples were transcribed with Google Cloud Speech, with an equal split across language abilities (AR = 85 and DLD = 85); 100 of which were sourced from children in a younger age-range (6;0-8;11) and 70 sourced from children in an older age-range (9;0-11;11).

The average transcription error of Google Cloud Speech, as measured by WER (word error rate), across all 170 narrative samples was 48%, meaning close to half of the words in the language sample were incorrectly transcribed. The median WER was lower at 40%, but was still higher than might be considered acceptable. Sources of transcription error included deletions , substitutions, and insertions. Post hoc error analyses revealed deletions to be the most common type of transcription error, followed by substitutions, and then insertions. This meant that Google Cloud Speech often either failed to generate words present in the language sample recording or misclassified them. Insertions (i.e., adding in a word not present in the recording) were not common. This finding differed from the pilot study of Google Cloud Speech (Fox et al., 2021) on narrative samples, where substitutions were found to be the most common type of word-error. This difference can likely be attributed to the inclusion of recordings of variable quality in the current project.

There was minimal difference in the average WER of transcripts produced from the language samples of AR children and children with DLD. The potential for difference amongst these groups was of interest, since the model underlying Google Cloud Speech partially relies on a language model to determine likely adjacent words in an utterance. Presumably, children with DLD have poorer language abilities than children that are AR, therefore it was possible that language samples elicited from the DLD group would have higher transcription error, based on input from the Google Cloud Speech language model (e.g., it might perform word substitutions when non-grammatical utterances occur). While the WER was slightly lower on average for the AR group (lower word error rates), both groups had WERs ranging from around 8%-100%, meaning language ability status didn't appear to impact transcription error.

Research Q2

Research question two asked whether subjective ratings of audio recording quality (as measured by mean opinion score), background noise ratings, or audio encoding/decoding type (i.e., codec) were associated with transcription error. Findings revealed that audio encoding (i.e., codec) was the factor that was most related to transcription error rates (WERs); not subjective ratings of audio quality or the ratings of the presence of background noise. Samples originally recorded with a WAV codec had the lowest average WERs, followed by MPEG, and then Windows Media . This meant that it was not sufficient to transcode samples to WAV (or FLAC) codecs after the fact, as was conducted in the audio preprocessing stage; instead, samples needed to be originally recorded with a WAV codec to obtain lower levels of transcription error. ASR transcriptions of samples recorded with a Windows Media codec were consistently poor, with only two such samples having WERs below 50%. This was due to the amount of compression (i.e., audio information lost) during encoding. WAV is a lossless codec, meaning it retained all audio information during encoding, while both MPEG and Windows media applied compression to reduce file size. Google Cloud Speech recommends against the usage of lossy codecs such as these, and this recommendation was supported in the current project. There are a number of means through which WAV recordings can be obtained, including using high-quality digital recorders (e.g., Zoom H2-4, Tascam DR-1), voice recorder apps (e.g., ASR Voice Recorder), or computer programs paired with external microphones (e.g., Audacity).

Interestingly, subjective ratings related to perceived audio quality (MOS) were not highly associated with transcription error, meaning that listener judgement of audio quality had little to do with how well Google Cloud Speech performed. The reason for this lack of a relationship became clearer when stratifying by codec. The relationship between subjective audio quality and transcription error only appeared to hold when an MPEG codec was used. When there was either no compression (WAV) or high compression (Windows Media), there was no clear association between these factors. This suggests that even when an audio recording “sounded good” to a listener, it was not a reliable indicator of how accurate Google Cloud Speech transcriptions would be. This was an important finding because it suggests that the appropriateness of using Google Cloud Speech to automatically transcribe language samples should be based on the codec used in recording, and not on the basis of perceived audio quality. Of note, MPEG codecs applied lower levels of compression, and for these samples ratings of audio quality were associated with transcription error (i.e., higher perceived quality was associated with lower transcription error). While it is recommended that users seek to utilize lossless codecs for Google Cloud Speech, it is possible that some degree of judgement on audio quality may be useful if a low compression codec like MPEG (e.g., .mp3 file format) must be used.

Similarly, the association between background noise rating and transcription error differed by audio codec. On average, audio samples recorded with both MPEG and Windows Media had lower transcription error when there was no background noise present, however, this relationship was stronger for samples recorded with MPEG. Further, the clearest association between age and transcription error was seen for MPEG recordings containing background noise, where recordings containing background noise had higher average transcription error when they were elicited from younger children. Collectively, this indicates that age and background noise may have an additive impact when codecs with compression are used. If recording audio with a lossy codec, users should seek to record in a quiet environment to prevent higher transcription error than necessary; if background noise is present, then the usage of Google Cloud Speech with younger school-age children should likely be avoided. The same comparisons for audio samples recorded in WAV could not be made since only a few were rated as containing background noise.

Based on these findings, in combination with findings from the previous research question, it was apparent that language sample recordings characteristics were more important to transcription error than the language status or age of the child it was elicited from.

Aim 2: To Increase the Scoring Accuracy of LLUNA for a more Diverse Population

Research Q3

The third research question assessed the accuracy of scores produced by modified version of LLUNA, a hard-coded function designed to automatically score the use of literate language indices in narrative language samples. In the pilot study, the scoring accuracy of LLUNA was found to be on par, or higher than the interrater reliability seen between trained MISL scorers on most of the literate language indices. LLUNA was most accurate on mental and linguistic verbs, and least accurate on ENPs. While LLUNA met acceptable levels of accuracy on all six indices, its kappas for coordinating conjunctions and ENPs were lower than those of trained MISL scorers. In the current study, modifications to LLUNA were made in an attempt to increase scoring accuracy on these two indices, and potentially to the remaining indices as well. In addition, LLUNA was modified to better suit an older (6;0-11;11) and more linguistically diverse (AR/DLD) population.

Gains in scoring accuracy were observed for all six indices when evaluated on a test set of narratives elicited from school-age (6;0-11;11) TD children (n = 85). These improvements were small for subordinating conjunctions, mental verbs, linguistic verbs, and elaborated noun phrase, and large for coordinating conjunctions and adverbs. In addition, LLUNA was observed to achieve higher levels of inter-rater reliability than scores obtained by trained MISL scorers in Fox et al., (under review) for all six measures; meaning that when evaluated on narrative samples of school-age (6;0-11;11) children with typical language abilities, LLUNA showed evidence of scoring accuracy that was higher than trained hand-scorers.

Of note, the increased scoring accuracy observed in coordinating conjunctions can likely be attributed to a simplification in its scoring ruleset, whereby habitual openers were no longer removed before calculating the total score. Habitual openers (e.g., beginning every utterance with *and* or *so*) were removed from narrative transcripts in the pilot study due to segmentation rules used in the dataset (i.e., mazes). In the current dataset, this was not the case because habitual openers are no longer mazed out during transcription. This resulted in less ambiguity of the scoring of coordinating conjunctions (i.e., the use of *and*) by LLUNA, which likely led to its improved accuracy, not any modifications made to LLUNA itself. The comparison between LLUNA and hand-scores on AR/DLD transcripts, all of which were scored for the current study, thus served as a better comparison of whether LLUNA was more accurate than a hand-scorers for coordinating conjunctions. Other gains in accuracy seen for the remaining indices can more confidently be attributed to modifications in LLUNA's scoring.

Research Q4

To address the fourth research question, it was of interest to examine whether LLUNA scoring accuracy generalized to narratives elicited from school-age (6;0-11;11) children with impaired language abilities (AR/DLD). Scoring accuracy for the AR/DLD test set was higher than the TD test set on coordinating conjunctions, subordinating conjunctions, and ENP, while mental and linguistic verbs had kappas which were slightly lower, but on par with the TD samples. These findings indicated that five these literate language indices could be consistently and accurately scored by LLUNA from plain text, traditionally transcribed transcripts of school-age children (6;0-11;11) across varying language abilities (TD, AR, DLD).

Test sets were also split by age-group (6;0-8;11 and 9;0-11;11) to determine whether LLUNA scoring accuracy differed by age-range. For most of the literate language indices, scoring accuracy was comparable across age-groups, with slightly higher scoring accuracy for the younger age-range. The higher accuracies seen for the younger age-range may have been sample specific, particularly due to the larger test set used for the 6;0-8;11 age-range ($n = 45$) as compared to the 9;0-11;11 age-range ($n = 23$). It is also possible that LLUNA was still missing some higher level vocabulary used by the older-age range, which may have accounted for the larger discrepancy seen between the scoring accuracy on linguistic verbs for the younger and older age-ranges.

Similarly, for ENP, children in the older age-range (9;0-11;11) may have used more complex combinations of parts-of-speech to elaborate on nouns that were not included in the ENP scoring ruleset. It may therefore be necessary to make additional modifications to LLUNA to achieve higher scoring accuracy for this older age-range (9;0-11;11). The test set accuracies for ENP still fell within an acceptable range (above .60 kappa level) for clinical usage, however.

The conclusions around LLUNA's ability to score adverbs across all ages (6;0-11;11) were less clear, as compared to the other five literate language indices. The scoring accuracy for adverbs was variable across the different test sets in the current project and the pilot LLUNA study. While the scoring accuracy of adverbs in the TD test set was near perfect in the current study, it was lower in the pilot study of LLUNA, and even lower on the AR/DLD test set. It is important to note however, that adverbs have also been difficult to consistently score by hand, as indicated by the lower levels of interrater reliability observed between trained scorers in both the current and pilot studies. Adverbs are challenging to score reliably due to their high level of semantic ambiguity (i.e., their meaning/part-of-speech tag is context dependent). It was this difficulty with reliably identifying adverbs in context that may have led to the variable performance of LLUNA's part-of-speech tagger. It is possible that the AR/DLD corpus contained more instances of ambiguous adverbs than the TD corpus or the corpus used in the pilot study. Having a larger number of ambiguous adverbs may have then led to greater levels of misclassification by LLUNA's part-of-speech tagger, even though it is a state-of-the-art version. As part-of-speech tagging technology continues to improve, the accuracy of LLUNA generated adverb scores will likely become more consistent. Even with the inconsistency of LLUNA accuracy in scoring adverbs however, it still surpassed the initial level of interrater reliability achieved between hand-scorers.

Aim 3: Assess how ASR Transcription Error Impacts LLUNA Scoring Accuracy

Research Q5

The fifth research question assessed how LLUNA scoring accuracies varied by level of transcription error. This was investigated in order to determine the impact of ASR transcription error on the clinical utility of LLUNA, as the ultimate goal of the current project was to address the feasibility of automating both the transcription and scoring portions of language sample analysis. Scoring accuracy (as measured by quadratic weighted kappa) was first calculated on LLUNA scores generated from the ASR transcripts. Recall that transcription error ranged between 8 (i.e., nearly perfect) to 100% (i.e., completely misclassified). Overall, analyses indicated that LLUNA scoring accuracy on ASR transcripts was poor, with only the kappa for mental verbs falling above the acceptable threshold of .60. While all scoring accuracies were poor, some literate language indices were more negatively impacted by transcription error than others.

Mental and linguistic verbs appeared to be the most robust to transcription error, followed by coordinating and subordinating conjunctions. This meant that conjunctions were more likely to be incorrectly transcribed (likely deleted or substituted) by Google Cloud Speech than either verb type. Scoring accuracy for ENP was a few points lower than either type of conjunction. This low scoring accuracy was expected given that ENP scoring relied on the combination of several words preceding a noun in a noun phrase. If any words within the noun phrase were to be deleted (or inserted) then the ENP score generated by LLUNA would be altered. An aspect of ENP that made it potentially more robust to transcription error was that it could tolerate some errors of substitution, as long as an acceptable part-of-speech tag was maintained in the substituted word (i.e., some valid part-of-speech combination present in the ENP syntactic scoring scheme). For example, if the original sentence the *big black dog* was substituted with *a very big frog* the ENP score would still be 3. LLUNA scoring for adverbs was no better than chance on ASR transcripts, indicating that adverbs were disproportionately impacted by transcription error. Adverbs cover a wide variety of potential words, so it is possible that the wide variety of adverbs used in this sample had a collectively higher probability of deletion or substitution by Google Cloud Speech.

Box plots were generated to show the relationship between transcription error (WER) and LLUNA scoring accuracy for each literate language index, as measured by the exact difference score. The difference scores were evaluated to show the exact difference between LLUNA generated scores and the reference (ground truth) MISL scores. Across all six literate language indices, in cases where LLUNA under-estimated the index score by one or more points, the associated transcription error was high. In most cases, the greater the difference score in the negative direction (e.g., -3, -2, -1) the higher the transcription error. Conversely, difference scores that were positive, meaning LLUNA overestimated the index score, tended to be associated with similar levels of transcription error as seen for correct LLUNA scores.

This can likely be attributed to the finding that the most common types of word-errors were deletions and substitutions, with insertions being relatively uncommon. When a large amount of deletions and substitutions occurred in ASR transcription, it led to high WERs and likely removed/replaced instances of literate language indices that would have otherwise been scored by LLUNA. In rarer cases where insertions occurred, there were likely not enough to significantly alter the WER, but may have caused LLUNA to count instances of literate language indices not present in the reference transcript. An additional likely cause for LLUNA overestimating scores were cases where it overgeneralized its scoring rules due semantic ambiguity, such as for misclassifying adverbs.

Research Q6

The final research question aimed to determine whether there was an acceptable range of transcription error that could maintain acceptable levels of LLUNA scoring accuracy. Across all six literate language indices, accurate LLUNA scores (i.e., difference score of 0) were associated with a median transcription error ranging between 25-50%. Meaning that it was common for transcripts with to of their words incorrectly transcribed to still have accurate LLUNA scores. LLUNA scoring for subordinating conjunctions, mental verbs, linguistic verbs, and ENPs were most tolerant of ASR transcription error, while coordinating conjunctions and adverbs were still robust, but more likely to be scored inaccurately as transcription error increased. These findings provided evidence that LLUNA was robust to transcription error, or alternatively, that literate language devices, especially subordinating conjunctions, linguistic, and mental verbs, and the words within an ENP, were less likely to be deleted or substituted in ASR transcripts as compared to other word varieties. Of note, the intended conclusion is not for users of ASR to strive for a particular range of transcription error when using LLUNA. Such a statement would be impractical given that those utilizing automated transcription for clinical purposes would not be calculating transcription error. Instead, these findings illustrate that even with a relatively large amount of transcription error, LLUNA could still reliably score all six literate language conventions.

Take for example the following ASR transcription of a narrative language sample that had a WER of 53%:

dolphin bed. ate some breakfast he didn't even to get brushes teeth to your perception. God rest on the bus he got dressed. what's the name of the store I'm trying out of bed brush your teeth. shoelace broke news tennis shoes. I don't know what you're going to do. I guess he put a NADA and tigers again. can you not let the other half of it. they took that forever to get that on the bus so I had to walk to school. how much does a principal I'm guessing it's a. prison why are you late to relax my shoelace broke and I had to catch the bus. Dance.

Upon inspection, it is immediately clear that this was a poor transcription, regardless of the exact word-error rate (53%). However, even given the high transcription error, the only difference between the LLUNA generated scores and the reference scores was an underestimation of subordinating conjunctions by one point. The remaining five literate language indices were scored accurately by LLUNA (i.e., difference score of 0). To show what LLUNA identified within this ASR transcript, the same text is presented below with the different literate language devices bolded.

*dolphin bed. ate some breakfast he did**n't even** to get brushes teeth to* *your perception. God rest on the bus he got dressed. what's the name of the store I'm trying out of bed brush your teeth. shoelace broke news tennis shoes. I don't* ***know*** *what you're going to do. I* ***guess*** *he put a NADA and tigers* ***again****. can you* ***not*** *let* ***the other half*** *of it. they took that forever to get that on the bus so I had to walk to school. how much does a principal I'm* ***guess****ing it's a. prison why are you late to relax my shoelace broke* ***and*** *I had to catch the bus. Dance.*

- Coordinating Conjunctions: and, so = 2
- Subordinating Conjunctions: 0
- Mental Verbs: know, guess(ing) = 2
- Linguistic Verbs: 0
- Adverbs: even, not/n't, again = 3
- ENP: The other half = 2

*this is alex. he got up from bed. he ate some breakfast. he did****n't even*** *get brush his teeth. he were in like. he got dressed on the bus. well* ***not*** *on the bus. he got dressed. this is* ***not even*** *a story. I'm* ***just*** *like. okay. he got out of bed brushed his. he ate breakfast brushed his teeth. I guess. shoelace broke* ***when*** *he was tying his shoes. I don't* ***know*** *what you're gonna do. I guess he put a knot in it and tied it* ***again****. he knotted it to* ***the other half*** *of it. x took that forever to get that on.* ***so then*** *he missed the bus.* ***so*** *he had to walk to school. I'm not sure if that's the principal. I'm guessing that's a principal. principal be like why are you late. he be like my shoelace my shoelace broke.* ***and*** *I had to catch the bus. okay. sounds good I* ***guess****.*

- Coordinating Conjunctions: and, so = 2
- Subordinating Conjunctions: when = 1
- Mental Verbs: know, guess(ing) = 2

- Linguistic Verbs: 0
- Adverbs: even, not/n't, again, just, then = 3
- ENP: The other half = 2

What this example illustrates is that even when faced with an obviously subpar transcription, LLUNA could often still produce accurate scores. While the occurrence of deletions and substitutions in the example ASR transcript resulted in a high WER, they only led to the incorrect scoring of subordinating conjunctions ("when" was deleted), which was off by one point. This indicates that other word types were likely disproportionately impacted by transcription error, as compared to literate language indices. It is unclear why this was case and may require further investigation that was beyond the scope of the current project.

Summary and Clinical Implications

Survey research has indicated time and time again that SLPs underutilize language sample analysis (LSA), even though it is a critical evidence-based practice that should incorporated as a part of a child's language assessment profile. To date, efforts to increase the usage of LSA have included expediting individual pieces of the process, such as the usage of real-time transcription (Klee et al., 1991), language analysis software (Miller & Chapman, 2021; MacWhinney), or simplified analysis procedures (Castilla-Earls & Fulcher-Rood, 2018; Pavelko et al., 2017). Unfortunately, even with the availability of these resources, patterns in the usage of LSA have not appeared to change much over time. The most frequently reported barrier to the implementation of LSA is a lack of time, so while each of these resources helps to reduce the time it takes to conduct LSA, a significant amount of time has been left up to the SLP to perform transcription, coding, and/or analysis (Pavelko et al., 2016). In the current project it was proposed that this barrier could be addressed by automating the most time-consuming aspects of LSA: transcription and scoring. The goal of the current project was to evaluate the feasibility of combining Google Cloud Speech ASR with LLUNA to streamline the LSA process, significantly reducing the time spent to complete this assessment protocol.

The current study first built upon prior work (Fox et al., 2021; Gonzalez Villasanti et al., 2020) by providing additional support for the use of automated transcription and scoring to expedite individual components of the language sample analysis process. In addition, it provided preliminary support for combining the two into a single streamlined LSA system under specific circumstances. These circumstances are discussed in further detail below.

In the current project, the average transcription error across all language samples was high, but also variable. Importantly, the variance in transcription error was not associated with characteristics of the narrator (e.g., age, language impairment), but with characteristics of the recording. This finding is supported by prior work, given that the average transcription error in the current study fell in between what was reported in Gonzalez Villasanti et al. (2020) and the Fox et al. (2021) pilot of Google Cloud Speech for LSA. One of the primary differences between Gonzalez Villasanti et al. (2020), which had higher overall transcription error (all over 60% WER) and Fox et al., (2021), which had lower overall transcription error (mean of 30%), was the recording characteristics of the evaluated language samples. Gonzalez Villasanti et al. (2020) used LENA recording units to capture child speech in a natural preschool setting where the speech of conversational partners (i.e., adults and peers) were recorded alongside the primary speech signal. LENA devices utilize a DVI4 codec, which applies a large amount of audio compression, likely to cut down on file size since its intended for long-form recording of child speech. Conversely, Fox et al. (2021) used language samples recorded in E-Prime, which uses a lossless WAV codec, in a quiet one-on-one environment. In addition, language samples rated as “poor quality” were excluded. The current project served as a middle-ground between these two prior works, by using samples that varied in recording characteristics, including variable levels of subjective audio quality, background noise, and audio compression (i.e., specified by the codec). The inclusion of language samples of variable recording qualities allowed for the evaluation the association between transcription error and recording characteristics. Information gained from this evaluation allowed for more precise clinical recommendations regarding the optimal implementation of ASR, as well as ASR with LLUNA.

According to the current study, the most important recording characteristic was the original codec used to encode the language sample audio. Recordings encoded in WAV had by far the lowest levels of transcription error on average. Fox et al. (2021), who also used WAV recordings, reported similar levels of average transcription error. Conversely, language samples encoded with codecs like MPEG and Windows Media, which apply compression during recording, had higher levels of average transcription error. The range of WERs seen for these language samples was similar to Gonzalez Villasanti et al. (2020) who also recorded language samples with a high compression codec. While the role of background noise was less salient, findings indicated that recordings encoded in MPEG and Windows Media had lower transcription error when background noise was not present. Collectively these findings support the recommendation that individuals intending to use ASR for automated transcription use a lossless codec while minimizing background noise during recording. In addition, findings related to subjective audio quality ratings indicated that users should not rely on their own judgements of “audio quality” when deciding whether or not it is appropriate to utilize automated transcription.

Of note, certain procedures that were included in this study, such as preprocessing, may be more difficult for clinicians to implement themselves without additional guidance or resources. All preprocessing in the current project was done through Adobe Audition, which is a proprietary software with a subscription fee. Free software such as VLC Media Player (VideoLan, 2020) can be used to perform preprocessing (e.g., converting file formats, modifying sampling rate, etc.) and to trim audio files, but these processes are more time-consuming in VLC than Adobe. It may therefore be necessary to offer preprocessing services to clinicians in conjunction with automated transcription service, and/or to separately evaluate the potential negative impact that failing to preprocess audio has on transcription error.

In terms of automated scoring, LLUNA scoring accuracy was relatively consistent across language samples, regardless of age-range (6;0-8;11) or language status (TD and AR/DLD). The least consistent LLUNA scoring was seen for adverbs, which was highly accurate on samples, while on others was only just above the threshold of acceptable. In all cases however, adverb scores generated by LLUNA had higher levels of interrater reliability than was observed for trained hand-scorers. This indicated that when used upon plain-text narrative language samples, LLUNA could consistently score the six literate language indices with accuracy that was on par with (if not higher in some cases than) trained MISL hand-scorers. Even with this finding, scores generated in LLUNA were not perfect. Because of this, it will be recommended that users spot-check LLUNA generated scores to ensure the highest level of accuracy.

When ASR and LLUNA were combined to automate transcription and scoring, scoring accuracy dropped considerably for each of the six literate language indices; though they were differentially impacted. Analysis revealed that LLUNA scoring of mental and linguistic verbs were most robust to transcription error, while adverb scoring was the least. The accuracy of LLUNA scoring was also found to vary by transcription error. When transcription error levels ranged between 25-50%, LLUNA often produced accurate scores across all six literate language indices, while higher levels of transcription error often led LLUNA to underestimate scores.

These findings provide preliminary support for the usage of ASR in combination with LLUNA, when transcription error is not exceedingly high. While users are not expected to calculate WER on their ASR transcripts, findings from the current project and prior work suggest that transcription error below 50% can often be obtained when recording language samples from school-age children (6;0-11;11), as long as a lossless codec is utilized in an environment with minimal background noise. Highest accuracy can be ensured by spot-checking the ASR transcription (i.e., listen once through the original recording) and LLUNA scores, however.

At this point in time, the streamlined ASR and LLUNA system is not ready for clinical implementation, mainly because it is not in an accessible format yet. However, once the combined ASR + LLUNA system is ready for clinical implementation, the following procedure will be recommended to ensure optimal results: 1) find a quiet place to record the language sample where there are minimal/no background speakers and noises (if possible), 2) record the language sample with a digital recorder, voice recorder app, or external mic and computer software pairing that can transcode to WAV (or another lossless codec), 3) transcribe the audio file with Google Cloud Speech, 4) automatically score the ASR transcript with LLUNA, 5) review the words identified by LLUNA, 6) listen once through the original audio to spot check the identified words within the child's language sample, 7) adjust scores if necessary. Even with the recommendations of spot-checking this streamlined system has the potential to save clinicians considerable time and effort in conducting LSA.

Currently a web-based application is under works in order to streamline steps 3-4, such that clinicians would not have to separately access Google Cloud Speech or LLUNA, but instead upload their language samples to one location. This is discussed further under future directions.

Limitations

The current project had a number of limitations. The first was that only limited usage of randomization was possible in the selection of language samples for each age-range (6;0-8;11 and 9;0-11;11) and language ability (TD, AR, DLD). This was due to having a number of inclusionary criteria (monolingual English speaker, age, and duration) that limited the potential pool. Requiring that all included samples be at least one minute long led to many potential narrative samples being excluded, particularly amongst AR and DLD children. This was not surprising given that children with impaired language abilities tend to tell shorter stories on average, as compared to their TD peers. However, setting a minimum duration as an inclusionary criterion was necessary to ensure that samples were representative of the included children's language abilities (Heilman et al., 2010; Tilstra & McMaster, 2007). Fortunately, the lack of randomization should not have altered the results, since the inclusion criteria were unrelated to audio recording quality (i.e., transcription error) or scoring accuracy; also, no statistical testing was utilized to establish causal relationships or test hypotheses. Instead, descriptive statistics and visualizations were to evaluate associations present in the data.

Another limitation was the lack of experimental control over audio recording quality/environment that was due to the use of existing data. While it was intended to include audio samples of variable recording qualities, the lack of experimental manipulation of these characteristics made it impossible to determine the exact singular or combined impact subjective audio quality ratings, background noise, and codec had on transcription error. Visual inspection of scatter plots and descriptive statistics were utilized to help parse transcription error variance by recording characteristics. This provided useful insights, particularly on codec, but no relationships were causally conclusive. What the current project may have lacked in internal validity, it made up for in external validity. In the real-world clinicians are likely not collecting language samples under highly controlled conditions. They are more likely collecting language samples when they get a chance to do so, in environments where there may or may not be other people/children in the room or background noise, without an awareness of how much audio compression their recording is applying. It was therefore beneficial to include recordings of variable quality in the current study to provide insight into the associations between recording characteristics and Google Cloud Speech's transcription error. In future work, a more tightly controlled comparison of factors that impact transcription error could be used to establish causality.

A final limitation was the lack of cultural and linguistic diversity amongst the selected language samples. This applied both in terms of the representation for language differences and speech disorders. While there was some representation of minority cultural and linguistic groups among the included language samples, the majority of language samples were still elicited from children utilizing General American English dialects. There was also no attempt to include language samples representative of children with speech disorders, as this was beyond the scope of the current project. Machine learning has been widely criticized for its inherent bias against minoritized populations including women, people of color, and individuals with disabilities. It is not the models themselves that are problematic, but the data that they are trained upon. Google Cloud Speech has been trained primarily on the speech of neurotypical white adult males, meaning that the further a language sample is from this demographic, the higher the transcription error will likely be. The issue of underrepresentation of minoritized populations in training data is important and will have a direct impact on the ability of clinicians to utilize Google Cloud Speech or other ASR systems for clients of culturally and linguistically diverse backgrounds.

Future Directions

There are several future directions for this work that are necessary to make this technology accessible and clinically useful to SLPs. The first is in the creation of a usable app or website for clinicians to easily access integrated automated transcription and scoring. This project is already underway. Currently, Google Cloud Speech and LLUNA have been integrated into a basic web applet, along with two other automated scorers (one for macrostructure and one for basic quantitative metrics), called *MISL Launch*. At this time MISL Launch is functional, but still requires some work in terms of web development and user accessibility before it can be made publicly available. It is author's intent to keep improving upon LLUNA by undergoing further fine-tuning on additional language samples, as well as to one day incorporate an open-source ASR model in place of Google Cloud Speech.

The creation of an open-source replacement for Google Cloud Speech is an additional future direction, which will first require the collection of a large corpus of language samples elicited from children of diverse age-ranges, language/speech abilities, and cultural and linguistic backgrounds. This corpus could be utilized for the purpose of training one or several child-specific ASR model(s) that may be more appropriate than Google Cloud Speech for this population. In the child ASR space, transfer-learning has been used to successfully adapt models trained on adult speech in order to increase the transcription accuracy of children's speech (Yadav & Prahdan, 2021; Gale et al., 2019). This research is still in its early stages however, and limited work has been put towards training models for children with diverse linguistic backgrounds, whether that be based on language differences (e.g., lingual status, dialect) or language/speech disorders. Another potential avenue is in the creation of speaker-dependent models for individual children. Past work has seen success in training speaker-dependent ASR models for adults with different degrees of speech impairment due to dysarthria with only limited training data from the individual (Marini et al., 2021; Mulfari et al., 2018). Such an approach may be useful for children with more severe language or speech impairments. This future work has the potential to provide accessible and accurate automatic transcription to clinicians serving children of diverse backgrounds, which is the long-term goal for this research.

Conclusions

The primary goal of the current project was to determine the feasibility of streamlining the LSA process by combining automated transcription (ASR) and scoring (LLUNA) for the evaluation of school-age oral narrative samples. This goal was broken down into several aims, including 1) examining the transcription of accuracy of Google Cloud Speech ASR on the narrative language samples of elementary school-age children (6;0-11;11) with impaired language abilities (AR and DLD), 2) examining the accuracy a modified version of LLUNA on school-age language samples when they're traditionally transcribed, and then finally 3) examining the accuracy of LLUNA when its used on ASR transcripts, essentially automating the LSA process. While there was room for improvement in both transcription and scoring accuracy, this study provided preliminary evidence that under the right contexts, both transcription and scoring could be automated while maintaining accuracy. These contexts included using a recording device that encoded audio with lossless compression (e.g., WAV codec), minimized background noise, and generally followed Google Cloud Speech recommendations. In addition, when this technology becomes available for clinical usage, it will be recommended that users still spot check LLUNA generated scores to ensure the highest level of accuracy. Even with the recommendation of spot-checking, the usage of combined computer automated transcription and scoring has the potential to save clinicians a considerable amount of time and effort, making LSA a more accessible evidence-based practice.

REFERENCES

Asgari, M., Sliter, A., & Van Santen, J. (2016). Automatic scoring of a sentence repetition task from voice recordings. *International Conference on Text, Speech, and Dialogue*, 470–477.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv Preprint arXiv:1409.0473*.

Ben-David, A. (2008). Comparison of classification accuracy using cohen's weighted kappa. *Expert Systems with Applications*, *34*(2), 825–832.

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157–166.

Berman, R. A., & Nir-Sagiv, B. (2007). Comparing narrative and expository text construction across adolescence: A developmental paradox. *Discourse Processes*, *43*(2), 79–120.

Bird, S. (2006). NLTK: The natural language toolkit. *Proceedings of the Coling/Acl 2006 Interactive Presentation Sessions*, 69–72.

Bloom, L., & Lahey, M. (1978). *Language development and language disorders.*

Botting, N. (2002). Narrative as a tool for the assessment of linguistic and pragmatic impairments. *Child Language Teaching and Therapy*, *18*(1), 1–21.

Bottou, L., & others. (1991). Stochastic gradient learning in neural networks. *Proceedings of Neuro-Nımes*, *91*(8), 12.

Calvo, I., Tropea, P., Viganò, M., Scialla, M., Cavalcante, A. B., Grajzer, M., Gilardone, M., & Corbo, M. (2021). Evaluation of an automatic speech recognition platform for dysarthric speech. *Folia Phoniatrica et Logopaedica*, *73*(5), 432–441.

Casby, M. W. (2011). An examination of the relationship of sample size and mean length of utterance for children with developmental language impairment. *Child Language Teaching and Therapy*, *27*(3), 286–293.

Castilla-Earls, A., & Fulcher-Rood, K. (2018). Convergent and divergent validity of the grammaticality and utterance length instrument. *Journal of Speech, Language, and Hearing Research*, *61*(1), 120–129.

Chan, W., Jaitly, N., Le, Q. V., & Vinyals, O. (2015). Listen, attend and spell. *arXiv Preprint arXiv:1508.01211*.

Chollet, F. (2017). *Deep learning with python*. Simon; Schuster.

Christ, T. J., & Silberglitt, B. (2007). Estimates of the standard error of measurement for curriculum-based measures of oral reading fluency. *School Psychology Review*, *36*(1), 130–146.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213.

Cummings, K. D., Biancarosa, G., Schaper, A., & Reed, D. K. (2014). Examiner error in curriculum-based measurement of oral reading. *Journal of School Psychology*, *52*(4), 361–375.

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, *5*(1).

Eisenberg, S. L., Ukrainetz, T. A., Hsu, J. R., Kaderavek, J. N., Justice, L. M., & Gillam, R. B. (2008). *Noun phrase elaboration in children's spoken stories.*

Espana-Bonet, C., & Fonollosa, J. A. (2016). Automatic speech recognition with deep neural networks for impaired speech. *International Conference on Advances in Speech and Language Technologies for Iberian Languages*, 97–107.

Evans, J. (1996). Plotting the complexities of language sample analysis: Linear and non-linear dynamical models of assessment. *Assessment of Communication and Language*, *6*, 207–256.

Evans, J. L., & Craig, H. K. (1992). Language sample collection and analysis: Interview compared to freeplay assessment contexts. *Journal of Speech, Language, and Hearing Research*, *35*(2), 343–353.

Finestack, L. H., Rohwer, B., Hilliard, L., & Abbeduto, L. (2020). Using computerized language analysis to evaluate grammatical skills. *Language, Speech, and Hearing Services in Schools*, *51*(2), 184–204.

Fox, C. B., Israelsen-Augenstein, M., Jones, S., & Gillam, S. L. (2021). An evaluation of expedited transcription methods for school-age children's narrative language: Automatic speech recognition and real-time transcription. *Journal of Speech, Language, and Hearing Research*, *64*(9), 3533–3548.

Fox, C., Jones, Sharad, Schwartz, S., Gillam, S., & Gillam, R. B. (2021). *Removing barriers to language sample analysis: Literate language use in narrative analysis (lluna).* Under review.

Fulcher-Rood, K., Castilla-Earls, A. P., & Higginbotham, J. (2018). School-based speech-language pathologists' perspectives on diagnostic decision making. *American Journal of Speech-Language Pathology*, *27*(2), 796–812.

Gabani, K., Sherman, M., Solorio, T., Liu, Y., Bedore, L., & Pena, E. (2009). A corpus-based approach for the prediction of language impairment in monolingual english and spanish-english bilingual children. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American*

Chapter of the Association for Computational Linguistics, 46–55.

Gale, R., Chen, L., Dolata, J., Van Santen, J., & Asgari, M. (2019). Improving asr systems for children with autism and language impairment using domain-focused dnn transfer techniques. *Interspeech*, *2019*, 11.

Gillam, R. B., & Pearson, N. A. (2004). *TNL: Test of narrative language*. Pro-ed Austin, TX.

Gillam, R., & Pearson, N. (2017). Test of narrative language–second edition (tnl-2). *Austin, TX: Pro-Ed*.

Gillam, S. L., Gillam, R. B., Fargo, J. D., Olszewski, A., & Segura, H. (2017). Monitoring indicators of scholarly language: A progress-monitoring instrument for measuring narrative discourse skills. *Communication Disorders Quarterly*, *38*(2), 96–106.

Gonzalez Villasanti, H., Justice, L. M., Chaparro-Moreno, L. J., Lin, T.-J., & Purtell, K. (2020). Automatized analysis of children's exposure to child-directed speech in reschool settings: Validation and application. *PloS One*, *15*(11), e0242511.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Gutiérrez-Clellen, V. F., & Simon-Cereijido, G. (2007). *The discriminant accuracy of a grammatical measure with latino english-speaking children*.

Hassanali, K.-n., Liu, Y., Iglesias, A., Solorio, T., & Dollaghan, C. (2014). Automatic generation of the index of productive syntax for child language transcripts. *Behavior Research Methods*, *46*(1), 254–262.

Hassanali, K.-n., Liu, Y., & Solorio, T. (2012). Evaluating nlp features for automatic prediction of language impairment using child speech transcripts. *Thirteenth Annual Conference of the International Speech Communication Association*.

Heilmann, J. J. (2010). Myths and realities of language sample analysis. *Perspectives* *on Language Learning and Education*, *17*(1), 4–8.

Heilmann, J., Nockerts, A., & Miller, J. F. (2010). *Language sampling: Does the length of the transcript matter?*

Hochreiter, S. (1991). Untersuchungen zu dynamischen neuronalen netzen. *Diploma, Technische Universität München*, *91*(1).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hornik, K. (2019). *OpenNLP: Apache opennlp tools interface* (Version 0.2-7). CRAN.

Hui, J. (n.d.). *Speech recognition - gmm, hmm.* `https://jonathan-hui.medium.com/speech-recognition-gmm-hmm-8bb5eff8b196`

Jurafsky, D., & Martin, J. H. (2020). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.*

Kemp, K., & Klee, T. (1997). Clinical language sampling practices: Results of a survey of speech-language pathologists in the united states. *Child Language Teaching and Therapy*, *13*(2), 161–176.

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

Laing, S. P., & Kamhi, A. (2003). *Alternative assessment of language and literacy in culturally and linguistically diverse populations.*

MacWhinney, B. (2000). CLAN [computer software]. *Pittsburgh, PA: Carnegie Mellon University.*

Miller, J. F., Andriacchi, K., & Nockerts, A. (2016). Using language sample analysis to assess spoken language production in adolescents. *Language, Speech, and*

Hearing Services in Schools, *47*(2), 99–112.

Miller, J., & Iglesias, A. (2010). Systematic analysis of language transcripts (salt) research version. *Madison, WI: SALT Software*.

Nese, J. F., & Kamata, A. (2020). Evidence for automated scoring and shorter passages of cbm-r in early elementary school. *School Psychology*.

Nielsen, M. A. (2015). *Neural networks and deep learning* (Vol. 25). Determination press San Francisco, CA.

Park, Y., Patwardhan, S., Visweswariah, K., & Gates, S. C. (2008). An empirical analysis of word error rate and keyword error rate. *INTERSPEECH*, 2070–2073.

Pavelko, S. L., & Owens Jr, R. E. (2017). Sampling utterances and grammatical analysis revised (sugar): New normative values for language sample analysis measures. *Language, Speech, and Hearing Services in Schools*, *48*(3), 197–215.

Pavelko, S. L., Owens Jr, 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.

Pezold, M. J., Imgrund, C. M., & Storkel, H. L. (2020). Using computer programs for language sample analysis. *Language, Speech, and Hearing Services in Schools*, *51*(1), 103–114.

Potamianos, A., Narayanan, S., & Lee, S. (1997). Automatic speech recognition for children. *Fifth European Conference on Speech Communication and Technology*.

Ratner, N. B., & MacWhinney, B. (2016). Your laptop to the rescue: Using the child language data exchange system archive and clan utilities to improve child language sample analysis. *Seminars in Speech and Language*, *37*, 074–084.

RCore-Team. (2021). *R: A language and environment for statistical computing.*

Reiter, C. (2016). Python multimedia tagging library: Mutagen. *GNU General Public License V2 or Later.*

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536.

Sagae, K., Lavie, A., & MacWhinney, B. (2005). Automatic measurement of syntactic development in child language. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (Acl'05)*, 197–204.

Sanger, T. D. (1989). Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Networks*, *2*(6), 459–473.

Scarborough, H. S. (1990). Very early language deficits in dyslexic children. *Child Development*, *61*(6), 1728–1743.

Schalkwyk, J. (n.d.). *An all neural on-device speech recognizer.* `https://ai.googleblog.com/2019/03/an-all-neural-on-device-speech.html`

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673–2681.

Shahnawazuddin, S., Bandarupalli, T. S., & Chakravarthy, R. (2020). Improving automatic speech recognition by classifying adult and child speakers into separate groups using speech rate rhythmicity parameter. *2020 International Conference on Signal Processing and Communications (Spcom)*, 1–5.

Shivakumar, P. G., & Georgiou, P. (2020). Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations. *Computer Speech & Language*, *63*, 101077.

Tilstra, J., & McMaster, K. (2007). Productivity, fluency, and grammaticality measures from narratives: Potential indicators of language proficiency? *Communication Disorders Quarterly*, *29*(1), 43–53.

Torrey, L., & Shavlik, J. (2010). Transfer learning. In *Handbook of research on machine learning applications and trends: Algorithms, methods, and techniques* (pp. 242–264). IGI global.

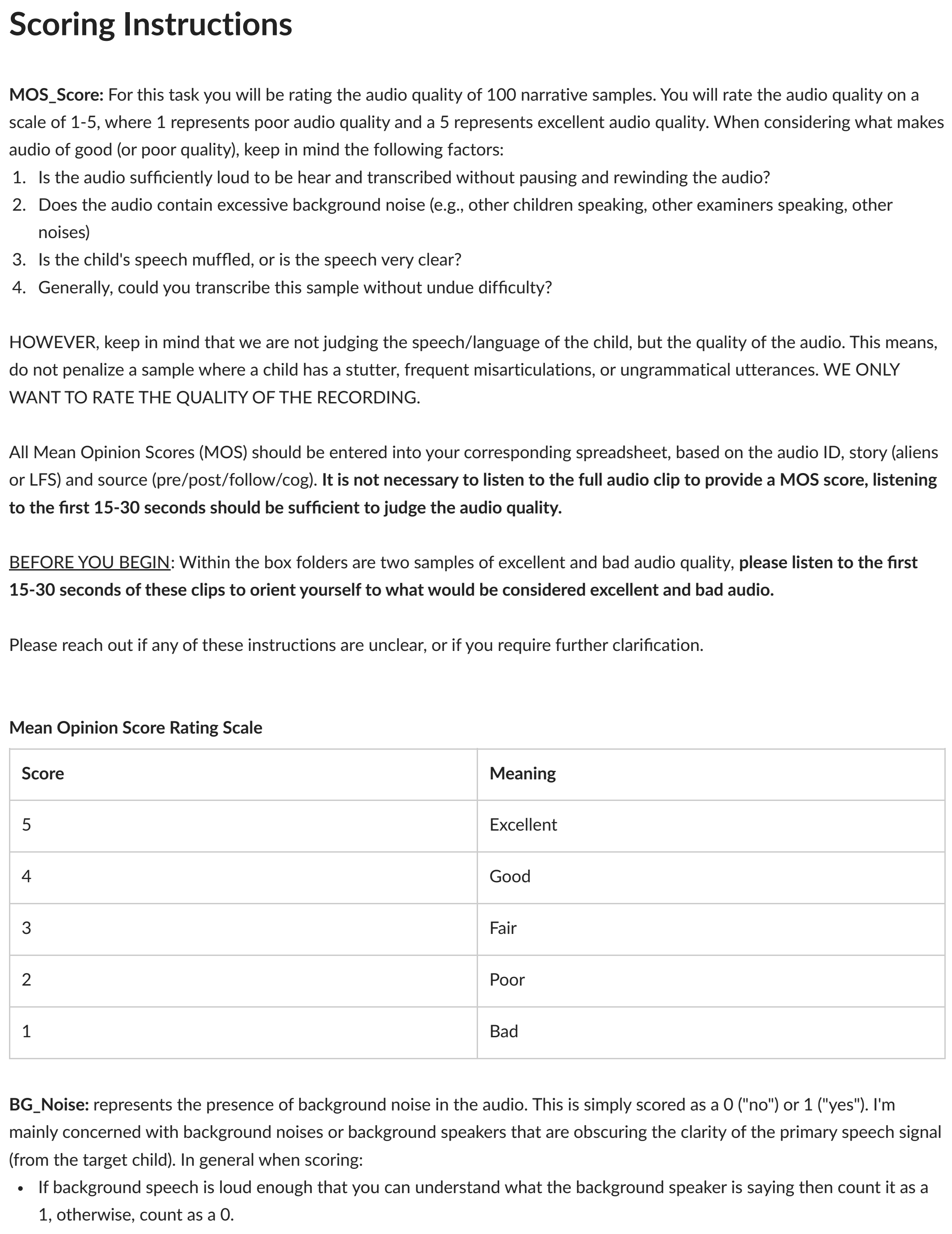
Van Rossum, G., & Drake. (2007). Python programming language. *USENIX Annual Technical Conference*, *41*, 36.

Westerveld, M. F., & Claessen, M. (2014). Clinician survey of language sampling practices in australia. *International Journal of Speech-Language Pathology*, *16*(3), 242–249.

Yadav, I. C., & Pradhan, G. (2021). Pitch and noise normalized acoustic feature for children's asr. *Digital Signal Processing*, *109*, 102922.

Young, V., & Mihailidis, A. (2010). Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review. *Assistive Technology*, *22*(2), 99–112.

APPENDIX



Scoring Instructions

MOS_Score: For this task you will be rating the audio quality of 100 narrative samples. You will rate the audio quality on a scale of 1-5, where 1 represents poor audio quality and a 5 represents excellent audio quality. When considering what makes audio of good (or poor quality), keep in mind the following factors:

1. Is the audio sufficiently loud to be hear and transcribed without pausing and rewinding the audio?
2. Does the audio contain excessive background noise (e.g., other children speaking, other examiners speaking, other noises)
3. Is the child's speech muffled, or is the speech very clear?
4. Generally, could you transcribe this sample without undue difficulty?

HOWEVER, keep in mind that we are not judging the speech/language of the child, but the quality of the audio. This means, do not penalize a sample where a child has a stutter, frequent misarticulations, or ungrammatical utterances. WE ONLY WANT TO RATE THE QUALITY OF THE RECORDING.

All Mean Opinion Scores (MOS) should be entered into your corresponding spreadsheet, based on the audio ID, story (aliens or LFS) and source (pre/post/follow/cog). **It is not necessary to listen to the full audio clip to provide a MOS score, listening to the first 15-30 seconds should be sufficient to judge the audio quality.**

BEFORE YOU BEGIN: Within the box folders are two samples of excellent and bad audio quality, **please listen to the first 15-30 seconds of these clips to orient yourself to what would be considered excellent and bad audio.**

Please reach out if any of these instructions are unclear, or if you require further clarification.

Mean Opinion Score Rating Scale

Score	**Meaning**
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

BG_Noise: represents the presence of background noise in the audio. This is simply scored as a 0 ("no") or 1 ("yes"). I'm mainly concerned with background noises or background speakers that are obscuring the clarity of the primary speech signal (from the target child). In general when scoring:

- If background speech is loud enough that you can understand what the background speaker is saying then count it as a 1, otherwise, count as a 0.

- If the primary speech signal (from the target child) is quiet/difficult to understand, and you can hear someone speaking in the background (even if you can't understand what they're saying), also score as a 1.
- If there is background noise (e.g., scratching, machine whirring, etc.) that is present through a substantial portion of the audio, also count as a 1.
  - If there is a sudden background noise that only occurs infrequently, you can count that as a 0.
- Otherwise, the audio should receive a score of 0, indicating "no background noise"

CURRICULUM VITAE

Carly Beth Fox

Curriculum Vitae
December 6, 2021

Utah State University
800 N 900 E
Logan, UT
84322
(860) 933-3161

476 W 1460 N
apt 104
Logan, UT
84341
carlyb.fox@gmail.com

EDUCATION

Ph.D. Disabilities Disciplines – Speech Language Pathology
Special Education and Rehabilitation
Utah State University
2021

M.S. Communication Sciences
Communication Disorders and Deaf Education
Utah State University
2020

B.S. Psychological Sciences
Psychology
Arizona State University
2017

PUBLICATIONS

Refereed Journal Articles

Fox, C., Gillam, S., Israelsen-Augenstein, M., & Jones, S (2021). An evaluation of expedited transcription methods for school-age children's narrative language: automated speech recognition & real-time transcription. *Journal of Speech, Language, and Hearing Research*, 1-16.

Peterson, A., Fox, C., Israelsen, M. (2020). A systematic review of academic discourse interventions for school-aged children with language-related learning difficulties. *Language, Speech, and Hearing Services in Schools*.

Jones, S., Fox, C., Gillam, S., & Gillam, R. B. (2019). An exploration of automated narrative analysis via machine learning. *PloS one*, *14*(10).

Koebele, S. V., Palmer, J. M., Hadder, B., Melikian, R., Fox, C., Strouse, I., Denardo, D., George, C., Daunis, E., Nimer, A., Mayer, L., Dyer, C. & Bimonte-Nelson, H. A.

(2018). Hysterectomy uniquely impacts spatial memory performance in a rodent model: A role for the nonpregnant uterus in cognitive processes. *Endocrinology*.

Prakapenka, A. V., Hiroi, R., Quihuis, A. M., Carson, C., Patel, S., Berns-Leone, C., Fox, C., Sirianni, R. W. & Bimonte-Nelson, H. A. (2018). Contrasting effects of individual versus combined estrogen and progestogen regimens as working memory load increases in middle-aged ovariectomized rats: one plus one does not equal two. *Neurobiology of Aging* (64) 1-14.

Book Chapters

Gillam, S., Mecham, J., & Fox, C. (2019). Special Education. In Jack S. Damico & Martin J. Ball (Eds.), The SAGE Encyclopedia of Human Communication Sciences. SAGE Publications. Thousand Oaks, California.

FELLOWSHIPS & AWARDS

2021 College of Education & Human Services Graduate Opportunity Award
Utah State University, 2021-2022

2020 Leadership Grant, USU
Utah State University, 2020-present

2020 Graduate Research & Creative Opportunities Grant
Utah State University, 2020-2021

2019 Graduate Poster Award for Social Sciences and Education Research, USU

2018 Presidential Doctoral Research Fellowship
Utah State University, 2018-present

CONFERENCE ACTIVITY

2021 The Future of Narrative Language Sample Analysis for Progress Monitoring, American Speech & Hearing Association National conference (November 20)

2021 An evaluation of expedited transcription methods for school-age children’s narrative language: Automated speech recognition & real-time transcription, Symposium on Research in Child Language Disorders (June 3-4)

2020 Measuring Similar but Distinct Aspects of the Memory System: A Proof-of -Concept Study, American Speech & Hearing Association National Conference (cancelled due to COVID-19)

2020 Chunking Abilities of Children with and without Developmental Language Disorder in Recall Tasks, American Speech & Hearing Association National Conference (cancelled due to COVID-19)

2020 The Attitudes of Speech-Language Pathologists and Speech-Language Pathologist-Assistants Toward Evidence-Based Practice, American Speech & Hearing Association National Conference (cancelled due to COVID-19)

2020 A Systematic Review of Academic Discourse Interventions for School-Aged Children with Language-Related Learning Disabilities, Symposium on Research in Child Language Disorders (cancelled due to COVID-19)

2019 Moving Forward in LSA: Computer Automated Microstructure Scoring (CAMS), American Speech & Hearing Association National Conference (November 21-23)

2019 Characteristics of Macrostructure & Microstructure used in Stories in Preschool & Early Elementary Years, American Speech & Hearing Association National Conference (November 21-23)

2019 Exploration of Automated Narrative Analysis via Machine Learning, American Speech & Hearing Association National Conference (November 21-23)

TEACHING EXPERIENCE

Regression Analysis, TA & guest-lecturer [online synchronous] (Summer 2020, Spring 2021; as a TA I designed in-class coding assignments, designed and updated lecture materials, and co-taught course material)

GUEST LECTURES

2021	Categorical Predictors
2021	Missing Data & Other Real-World Problems
2021	Introduction to Generalized Linear Models
2021	Predictive Modeling & Machine Learning
2020	Literature-Based Language Interventions

RESEARCH EXPERIENCE

2018-	Child-Language Laboratory, Communication Disorders & Deaf Education
2016-2017	Autism and Aging Brain Laboratory, Speech & Hearing Science
2016-2017	Human Brain Imaging Laboratory, Clinical Psychology

SERVICE TO PROFESSION

2021	Frontiers in Human Neuroscience, Ad Hoc Reviewer
2021	California State University Student Research Competition Judge
2018-2019	PLOS One, Ad Hoc Reviewer
2018	Journal of Behavioral Education, Ad Hoc Reviewer

DEPARTMENTAL/UNIVERSITY SERVICE

Undergraduate Research Grant Reviewer (2019-2021)

Volunteer at Utah Conference of Undergraduate Research (2020)

Master’s Thesis Mentor (2018-2020)

Presenter at Undergraduate Workshop on *How to be a Better Mentee* (2019)

Fall Undergraduate Student Research Symposium Judge (2018)

COMMUNITY OUTREACH

2020	Co-Organizer and Host for *Science on Tap* at the Cache Venue

RELATED PROFESSIONAL SKILLS

R (6+ years experience)

Python (1+ years experience)

COURSES DEVELOPED

Regression Analysis

Language Development

PROFESSIONAL MEMBERSHIPS/AFFILIATIONS

National Student Speech Language Hearing Association (NSSLHA)

R-Ladies