2023

Does Accuracy Matter?: Methodological Considerations When Using Automated Speech-to-Text for Social Science Research

Steven J. Pentland
Boise State University

Christie M. Fuller
Boise State University

Lee A. Spitzley
University at Albany

Douglas P. Twitchell
Boise State University

This is an Accepted Manuscript of an article published by Taylor & Francis in *International Journal of Social Research Methodology* on June 17, 2022, available online: https://doi.org/10.1080/13645579.2022.2087849

Does Accuracy Matter?: Methodological Considerations When Using Automated Speech-to-Text for Social Science Research

Steven J. Pentland*
Boise State University
stevenpentland@boisestate.edu

Christie M. Fuller
Boise State University

Lee A. Spitzley
University at Albany

Douglas P. Twitchell
Boise State University

Abstract

The analysis of spoken language has been integral to a breadth of research in social science and beyond. However, for analyses to occur with efficiency, language must be in the form of computer-readable text. Historically, the speech-to-text process has occurred manually using human transcriptionists. Automated speech recognition (ASR) is advertised as an efficient and inexpensive alternative, but research shows this method of speech-to-text is prone to error. This paper investigates the viability of using error prone ASR transcriptions as part of the methodological process of language analysis. Results show that at the individual feature level, analysis of ASR transcriptions differ dramatically from human transcriptions. However, when the same features are used for classification, a common machine learning task, performance results between ASR and human transcriptions are similar. We present these findings and conclude with a discussion on the methodological considerations for researchers who opt to use automated speech recognition for social science research.

Keywords: linguistic analysis, text-to-speech, LIWC, machine learning, automated speech recognition

Introduction

The systematic analysis of written or spoken language has been integral to a breadth of research in social science and beyond. A wide variety of methodologies are available that allow researchers to investigate connections between language usage and constructs such as personality (Holtzman et al., 2019), emotion (Abe, 2020), persuasion (Humă et al., 2019), culture (Spencer-Oatey & Wang, 2019), and deception (Burgoon, 2018). For these investigations to occur with efficiency, language must be in the form of computer-readable text (Crowston et al., 2012). Software packages such as Linguistic Inquiry and Word Count (LIWC, Pennebaker, J.W. et al., 2015) have been developed to process this text and then output psychological metrics. In many situations, language is already in an electronic format (e.g. essays, Tweets, comments, and emails), which allows for efficient processing. In other situations, spoken content must first be transcribed before analysis can occur. The transcription process has historically occurred manually. However, for social scientists with limited resources and/or those interested in exploring large volumes of speech recordings (e.g., large archival records, social media video recordings), manual transcription is slow, time consuming, and expensive. For example, one hour of recorded speech takes approximately four hours to transcribe (Dempster & Woods, 2011; Rev.com, n.d.). Automated speech recognition (ASR) is advertised as an efficient and inexpensive alternative, but research shows this method of transcription is prone to error. This paper investigates the viability of using error prone ASR transcriptions as part of the methodological process of speech analysis.

In the sections that follow, we will: 1) overview some recent social science research that uses language as the primary data source, 2) describe the general methodological process for computerized linguistic analysis, 3) and highlight the advantages and pitfalls of ASR transcriptions. We then seek answers to the question: Do text analysis results substantively differ when conducted using low-quality automated transcriptions versus high-quality manual transcriptions? To answer this question, we perform a series of analyses on a job-interviewing dataset where question responses are transcribed manually and automatically. We conclude with a discussion on the methodological considerations for researchers that opt to use automated speech recognition for social science research.

Background

Text Analysis in Social Science Research

Prior studies have illustrated the great potential of using text to investigate human behavior. Personality (Holtzman et al., 2019), affect (Cheng et al., 2017), consumer sentiment (Chatterjee et al., 2020), impression management (Pan et al., 2018), and health (Ziemer & Korkmaz, 2017) are just a few constructs studied in relation to language. Slight differences in language usage such as the use of first-person plural versus first-person singular can suggest characteristics like narcissism (Holtzman et al., 2019) or deception (DePaulo et al., 2003). Ho et al. (2017) also found that deceptive actors use more negative language and language associated with cognitive processes during group interactions. Studies looking at less nefarious behaviors have correlated language behaviors with playfulness (Proyer & Brauer, 2018) and personality traits (Park et al., 2015).

Most studies to date rely on text produced by the subject(s) of study. This text may come in the form of essays, notes, Tweets, etc. that have been typed by subjects. For instance, Okdie & Rempala (2019) looked at student writing prompts, Supreme Court justice's written statements, and celebrity Tweets to investigate textual indicators of political affiliation. In another study, Prati et al. (2019) used archival employee performance appraisals to investigate potential gender bias in review committees. The breadth and volume of text that can be created, shared, and archived outside of a laboratory setting offers researchers great insight into social behaviors without the need for controlled experiments. The creation and sharing of text has been further amplified by the internet. Social media posts, product reviews, forums, etc. all offer rich sources of data for researchers from a wide range of specialties.

A less predominate, yet still influential source of text comes from transcribed speech. Transcribed speech refers to spoken language recordings that have been converted to text. For instance, Humă, Stokoe, & Sikveland (2019) transcribed telemarketing phone calls to investigate persuasive discourse. Foley et al. (2020) transcribed provider-patient interactions in a healthcare setting to study patients' perception of advice quality. Given advancements in technology, speech is easier to record and share than any time in history. Technology offers many sources of speech online such as YouTube, social media posts, and personal video blogs. It is also easier than ever before to conduct dyadic interviews or focus groups online where interactions can be recorded.

Text Analysis Overview

At its core, text analysis relies on structuring text so that common analytical techniques can be applied. Arguably, bag-of-word (BoW) and lexicon-based approaches are most commonly used to structure text. In BoW, unique terms are counted in each document and the results are organized in a matrix with columns representing unique terms and rows representing documents. The content of the matrix is generally the count/frequency of each unique term in each document. These are referred to as term-document matrices.

A lexicon-based approach attempts to assign a label to words in a document. One of the simpler examples of this is polarity. For this method, a polarity lexicon maps unique words to either positive or negative polarity. For instance, the word, "happy" would be assigned a positive label, and the word, "sad" would be assigned a negative label. The frequency of positive words versus negative words can then be used for analysis.

Linguistic Inquiry and Word Count (LIWC) is a popular lexicon-based tool used by social science researchers. Beyond summary dimension such as word and part-of-speech count, LIWC also uses extensive lexicons to measure constructs such as affect, time orientation, and cognitive processes in text. Each category may include multiple sub-categories. For instance, the LIWC category of Affect includes Positive Emotion, Negative Emotion, Anxiety, Anger and Sadness. LIWC has been used to study the relationship between linguistic styles and success in online learning (Abe, 2020), how a movie synopsis may impact the movie's financial performance (Hung & Guan, 2020), the linguistic markers of grandiose narcissism (Holtzman et al., 2019), and much more.

Whether using a bag-of-words or lexicon-based approach, past researchers have used the structured text that results from these methods as inputs to machine learning analyses in a wide range of applications. For example, multiple researchers have used machine learning techniques, such as decision trees, neural networks or support vector machines to classify text as deceptive or fraudulent (Fuller et al., 2009). Another stream of research using this methodology is suicide risk (Bayram & Benhiba, 2021). Others have researched personality characteristics using machine learning

(Souri et al., 2018; Wang et al., 2021). These studies go beyond just converting unstructured to structured text and analyzing differences at the individual feature level. This research uses advanced techniques to uncover complex patterns in the text to classify varying groups of interest.

Automated Speech Recognition

The general time estimate for manual speech transcription is 4 hours for every 1 hour of recorded audio (Dempster & Woods, 2011; Rev.com, n.d.). However, Bazillion et al. (2008) found that variables such as having multiple speakers, spontaneous speech (versus prepared speech), and correcting transcriptionist spelling errors can significantly increase transcription time. For example, the transcriptionists in their study needed approximately 90 minutes to transcribe, tag speakers, and correct spelling errors for every 10 minutes of spontaneous speech. Bazillion et al. (2008) also evaluated a hybrid approach where transcriptionists corrected an ASR generated transcription instead of manually producing the full transcript. Even with this approach, the study found that transcriptionists needed about 40 minutes to correct and tag 10 minutes of prepared speech. Spontaneous speech took even longer.

In general, humans produce the highest quality transcriptions with error rates of about 5% and are used as the benchmark for ASR technologies (Glaser, 2017; Saon, 2017). However, human transcription is slow, expensive, and in many situations, impractical. The desire for accurate, real-time transcriptions is realized by the many automated transcription services currently on the market. These include Siri, Amazon Transcribe, IBM Watson, Google Speech-to-Text, CMU Sphinx, Dragon and Microsoft Azure.

Arguably, there are cases where manual transcription is preferable. For instance, researchers conducting conversational analysis are likely also interested in paralinguistic features (e.g., vocal pitch, speech rate, word emphasis) associated with spoken words (Paulus et al., 2014). Transcribing these types of features is likely better done using manual human interpretation. Also, ASR may not accurately transcribe slang, an important feature of some linguistic investigations (e.g., Zhou & Fan, 2013), since the systems rely on established dictionaries to match speech to text. However, for social scientists strictly interested in assessing linguistic content, ASR potentially offers increased efficiency. This efficiency allows transcripts to be produced faster and cheaper, and in turn, allows for data to be analyzed that perhaps previously was out of the resource bounds of researchers. Araújo et al. (2017), for instance, collected data from 12,000 YouTube videos to evaluate video content targeted towards children. Their methodology used tags by video uploaders to categorize video content. Likely, richer insights could have been derived about video content if the video transcriptions were evaluated, but given the volume of data, manual transcription was likely prohibitive. Even laboratory experiments can produce sample sizes large enough to make manual transcription problematic. Dorn et al. (2021) used a laboratory-based group interaction scenario to study deception. The experiment involved 96 group interactions with 693 participants. Each interaction took approximately 1 hour. A low estimate of the time it would take to manually transcribe these interactions is 384 hours (96 group-hours × 4 transcription-hours). Though, given the multiple speakers per interaction, total transcription time is likely to be much higher than this.

In order for social scientists to efficiently leverage large datasets of speech recordings, automated transcription will likely need to be a component of the methodological process. Otherwise, most researchers will be resource constrained when attempting to transcribe speech. If shown to have sufficient accuracy, automated transcriptions could prove advantageous. Overall, automated transcription offers efficiency and scalability allowing greater access to linguistic data.

Although the efficacy of ASR has improved substantially over time, accuracy rates can still vary dramatically (Dernoncourt et al., 2018; Këpuska & Bohouta, 2017). Background noise, speaker accent, poor recording quality are a few variables that may affect quality. While some sources report ASR transcription accuracy of up to 95%, rivaling that of human transcription (Glaser, 2017; Saon, 2017), accuracy appears to vary widely across tools and samples. One recent study reported Word Error Rates (WER) ranging from 7.3% to 54.2% depending on the corpus and tool (Dernoncourt et al., 2018). Another study reported WERs ranging from 0 to 83% (Këpuska & Bohouta, 2017).

The effects of transcription quality on downstream analysis, to date, has predominately focused on topic modeling and information retrieval tasks. Agarwal et al. (2007) discovered that document noise as high as 40% had very little impact on supervised topic modeling. On the other hand, Walker et al.'s (2010) evaluation of non-supervised (clustering and LDA) methods did not yield the same robustness showing a negative correlation between noise and model

performance. Research has yet to show the impact of transcript errors on text analysis processes beyond topic modeling and information retrieval. For those seeking to gain social and behavioral insights using text produced by ASR, it is crucial to understand the extent to which errors influence results.

Case Study

To evaluate the effects of human and automated transcription accuracy on subsequent linguistic feature generation and downstream analysis, we needed to identify a dataset that had three components: 1) original audio recordings to be processed using ASR tools; 2) manually generated transcriptions for assessment of ASR accuracy; 3) and a dependent variable to be assessed using the human and automated transcriptions. A job interviewing dataset collected by [REDACTED] met these criteria and was used for the current study. The dataset included audio-video recordings of subjects responding to job interview questions as well as a hireability score generated by third-party judges. Audio from interview-question responses was transcribed by humans and three different ASR tools. Human transcriptions acted as ground-truth.

Subjects (n = 89; Male = 35, Female= 54) participated in a laboratory-based experiment where they responded to job interview questions while being video recorded. The interview replicated automated one-way interviewing systems commonly used to screen job candidates. Interview questions were presented one at a time as text on a computer screen. Subjects responded to questions using the same computer and camera setup in a controlled laboratory setting. Responses were captured in a compressed FLV video format using a Logitech C920 webcam. Each subject had 30 seconds to read each question and to consider their response. Following the 30 second prep-period, the camera would automatically start recording. Subjects were given 60 seconds to respond to each question before the camera stopped recording. One subject (female) was removed from the dataset due to a recording malfunction that affected audio.

Each subject responded to 15 questions. However, only three questions were reviewed by third-party judges – the following three questions are the focus of the current study: 1) Tell me about a time when you successfully balanced several tasks at one time. How did you decide what to do first? In hindsight, was there a better way to have approached these tasks? 2) On a scale from 0 to 5 with 0 being none and 5 being a great deal, rate your level of experience with the following: Microsoft Excel. Give a brief example to back your rating. 3) Why should we hire you?

Seven judges watched each interview-response and then completed a hireability assessment derived from Cable & Judge (1997). The assessments of the seven judges were averaged to create a single hireability score for each participant. Audio from the three question responses were converted to text transcription using a professional transcription service and three popular ASR services: Watson Speech-to-Text (IBM), Google Speech-to-Text, and Amazon Transcribe. The professional service relied on human transcription. All transcriptions were then converted to term-document matrices and LIWC feature-sets.

Transcription Quality Analysis

Dernoncourt et al.'s (2018) method and published code was used to determine ASR transcript quality with the professional, human transcription acting as the baseline. Table 1 highlights the percentage of substitutions, insertions, deletions and overall word-error-rate (WER) for each ASR service compared to human-transcription. The transcriptions contain errors at a far higher rate than would be expected from human transcription (~3% to 5%; Glaser 2017), and is much higher than rates reported by ASR service providers – e.g., IBM's reported 5.5% WER (Saon, 2017). For all tools in our trial, the largest category of errors was substitutions (transcribing a spoken word as a different word), while the least common type of error was insertions (adding a word to a transcription that was not spoken). Error rates are within the wide range found in other papers (e.g., Dernoncourt et al., 2018; Këpuska & Bohouta, 2017), and support previous findings which suggest that the error rate in the transcription varies with the audio source. Appendix A provides extended details on common errors made by the ASR services.

<Table 1 here>

Text and Linguistic Features

To further understand the key differences between transcription sources, text transcriptions were converted to term-document matrices. Transcriptions were also processed with LIWC to generate a linguistic feature-set. The features generated using these methods were then compared between human-transcriptions and ASR-transcriptions.

Term-Document Matrices

Term-document matrices or bag-of-word models are commonly used to represent text documents in a structured format for various text-mining and information retrieval tasks. Unique terms are counted in each document and the results are organized in a matrix with columns representing unique terms and rows representing documents. The content of the matrix is generally the count of each unique term in each document. Throughout this paper, these unique words are referred to as features. All transcriptions underwent standard natural language preprocessing including conversion to lower-case, removal of contractions, punctuation, stop-words (e.g. common words such as "the", "but", "is"), and finally, stemming was performed. Stemming is the process of converting words to a root form. For instance, the words "plays", "played", and "playing" are convert to the root form, "play". Finally, a term-document matrix was created for each transcription source.

Table 2 outlines the number of unique features from each transcription source after preprocessing, as well as the number of common features found in ASR transcriptions compared to human transcriptions. Overall, about 85% of ASR terms were also in the human transcriptions. Interestingly, for all the ASR transcriptions, there were more unique features compared to the human transcriptions. In the Google transcriptions, there were over 600 additional features that were not in the human transcriptions.

<Table 2 here>

When comparing error rates (Table 1) and bag-of-word feature counts (Table 2), we see that substitutions outweigh insertions, but each ASR service has more unique features compared to human transcriptions. This finding suggests that ASR services are inconsistently substituting words. For instance, ASR services are misrepresenting the repeated occurrence of the same word in several different ways.

LIWC Generated Linguistic Features

Next, the human and ASR transcripts from the interviewing study were each processed using LIWC. LIWC output includes summary statistics such as word count and part-of-speech usage. LIWC features also include a mapping between text and psychological dimensions such as perceptions, affect, and personal concerns. LIWC output includes 14 categories with approximately 90 individual features.

An initial distribution evaluation revealed that most LIWC output variables violated normality, making a paired t-test inappropriate for comparing differences between human and ASR transcriptions. Instead, the nonparametric Wilcoxon Signed Rank Test was used, which allows for paired comparison, but does not assume normality. For the test, LIWC features for each transcript (interview) were compared between human transcriptions and ASR transcriptions. For instance, LIWC calculated the word count for each interview for each transcript type (human, IBM, Google, Amazon). A comparison was then made between the word count from human transcriptions and each ASR transcription. The Wilcoxon Signed Rank Test relies on the ranked difference between samples to test the hypothesis that the median values differ between samples.

The Wilcoxon Signed Rank Test excludes tied pairs, and the n value is adjusted based on these exclusions. Given that certain LIWC features are rare in text and lead to high levels of sparseness (feature values of zero) causing frequent tied pairs, certain features were excluded from the analysis to maintain an appropriate sample size. A large contributor of sparseness was related to LIWC features measuring punctuation, which may be sparse in the case of exclamation marks and are somewhat subjective when converting speech to text. All LIWC measures of punctuation were removed as well as any measure that appeared in fewer than 15 transcripts for at least one of the ASR services analyzed. The LIWC measure of Words Per Sentence was also removed from the analysis since it relies on punctuation for the calculation. In total, the differences between 75 LIWC measures were assessed. Table 3 describes each category of measure and the number of assessed features in that category.

<Table 3 here>

For the three ASR transcription services, the total percentage of significantly different ($p < .05$) features compared to the human transcription were: IBM = 62.67%; Google = 64%, Amazon = 72%. For example, out for the 75 LIWC measures assessed, 47 (62.67%) of these features statistically differed when comparing human transcript generated features to IBM Watson transcript generated features. For all three ASR transcription services, LIWC largely

overestimated measure values compared to measures generated from human transcriptions; 41 of the 47 significantly different measures were higher for IBM Watson transcriptions vs Human transcriptions; 43 of the 48 significantly different measures were higher for Google vs Human; and 44 of the 59 significantly different measures were higher for Amazon vs. Human. This is consistent with our finding above that the number of features generated by ASR transcripts is higher than the number of features found in human transcripts.

Machine Learning Comparison

Our initial comparison of bag-of-word features and LIWC generated features revealed widespread feature differences between human transcriptions and ASR transcriptions. However, it is unknown if these differences greatly impact downstream analyses. To evaluate the impact of ASR errors on analyses we used a machine-learning task.

Machine learning was selected as the sample analytical approach since a large number of variables can be assessed simultaneously for overall predictability. Further, in Big Data research where ASR technology could prove most advantageous, relying on the traditional statistical techniques to gage variable significance can be fraught with misrepresentation as sample sizes increase (Lin et al., 2013).

Separate models were trained to predict ratings of hireability using human and ASR text features. The hireability score was transformed into a binary value for classification. Scores above the median were coded as *High* and those less than or equal to the median as *Low*. Classification models using Random Forest, Naïve Bayes, and Support Vector Machine were trained to predict High and Low levels of hireability using bag-of-word and LIWC generated features. These algorithms were selected for their popularity in supervised classification literature and as a general representation of machine learning models for the purpose of seeing differences between ASR and Human transcriptions. Google Speech-to-Text was the focus of our investigation between the ASR transcriptions. Google Speech-to-Text had the highest word-error rate (WER) during our initial evaluation and represents the worst-case scenario for the current dataset.

Three text feature sets were selected for machine learning evaluation: LIWC only features, Bag-of-Word only features, and combined LIWC-BoW features. For each dataset, classification performance was compared between human transcriptions and Google Speech-to-Text transcriptions. Because of the relatively small sample size, Leave-One-Out Cross-Validation was selected to train and evaluate each classifier. For this performance validation strategy, $(n-1)$ data points are used to train a classifier. The model is then used to predict the class of the hold-out data point. The train-predict process is repeated n-times until all data points have been predicted. Overall performance measures are calculated across all train-predict iterations.

Classification Performance Results

Figure 1 below displays overall classification accuracy for each text feature set and machine learning model. Appendix B displays detailed performance statistics for each feature set and model.

<Figure 1 here>

LIWC Features

The first set of models were trained using 75 LIWC features described above. Although performance results between human and ASR transcriptions were similar, across the three models, the average accuracy over machine learning algorithms for human transcription was 70.45% (SD = 4.95%) versus 65.15% (SD = 8.68%) for ASR. Given the differences between human and ASR LIWC features found in the previous analysis, it is somewhat surprising that performance was not further apart. This initial assessment suggests WER may not have a significant impact on machine learning performance when analyzing psycholinguistic datasets.

Bag-of-Words Features

Next, Bag-of-Word features sets were used to predict ratings of hireability. Prior to training, word counts in each dataset were weighted for importance using term frequency—inverse document frequency (TF-IDF). TF-IDF is a popular method used to deemphasize the frequency of words that are common throughout a particular corpus.

To further refine the models due to the large number of features relative to sample size, three levels of word sparseness were selected for evaluation. Sparseness refers to the absence of a unique terms across documents. Sparseness levels were set to 0%, where all terms in the document matrix were included in the models; 5%, where terms appearing in less than 5% of the transcripts were removed; and 15%, where words appearing in less than 15% of the transcripts were removed. When the 5% and 15% sparseness thresholds were applied to human transcriptions, 478 and 201 features remained respectively. When applied to Google ASR transcriptions 499 and 190 features remained respectively. These sparseness levels combined with three models (Naïve Bayes, Random Forest, SVM) and two data sources (human and ASR transcripts) yield 18 classification models.

On average, model accuracy rates were highest when sparseness was set to 15%. At this level of sparseness, the average accuracy rate for all three models trained on human transcriptions was 64.02% and 61.74% when trained using Google ASR transcriptions. Similar to models generated using LIWC features, performance was higher for human transcriptions.

Combined LIWC and Bag-of-Word Features

Previous research (Ott, 2011) has shown that classification results may be improved by combining inputs from LIWC and bag-of-words feature sets. To see if accuracy can be further improved and to determine if such a strategy reduces or increases the differences between human and ASR results, we combined LIWC derived features and TF-IDF features with sparseness set to 15%.

The average accuracy rate across models was 72.27% when trained using human transcriptions and 69.70% when trained using ASR transcriptions. When considering all models that were evaluated (LIWC only, BoW only, & LIWC + BoW), performance was on average higher when trained on a combination of LIWC and BoW features.

Standardizing Variables

Our evaluation of machine learning algorithms trained using ASR and human transcription revealed that performance did not vastly differ between transcription types. This finding is consistent with findings from previous research in topic modeling and information retrieval. In total, 15 machine learning models were created for each type of transcription (ASR and human). Across these 15 models, the average accuracy at predicting hireability for human transcriptions was 65.76% and the average accuracy for ASR was 60.61%. In some cases, ASR models outperformed models created using human transcriptions.

The similarity between human-transcription and machine-transcription learning performance suggested that ASR error consistency was leading to a standardization effect in the data. Because of this standardization, machine learning models were not greatly affected by errors since classifications were based on relative values within the dataset. For example, word count for Google transcriptions is consistently lower than the word count of human generated transcriptions, but the word count ranking (highest to lowest) between transcription types was relatively unchanged between the different types of transcripts. Meaning that the interview with the highest word count was the same for human generated transcription and Google ASR generated transcriptions.

To further investigate the standardization phenomenon, Spearman Rank Correlation was performed between LIWC features generated using human transcription and LIWC features generated using Google ASR. Correlation results indicate that LIWC feature ranks remain largely consistent between transcription types. The correlation was greater than .80 for the majority of the LIWC features (see Figure 2 for histogram of correlations).

<Figure 2 here>

Next, the z-score of each LIWC feature was calculated for each transcript type. The Wilcoxon Signed Rank test was then rerun to determine if LIWC features statistically differed after formal standardization occurred. Table 4 displays the number of significant differences for each LIWC Category when comparing human transcription versus machine transcription feature values. Prior to scaling, 64% of LIWC features significantly ($p < .05$) differed between human and Google ASR. However, these differences dropped to 14.67% of LIWC features after scaling. For the majority of features that differed after scaling, the features were sparse (many occurrences of zero). For instance, when counting

the occurrence of zero values produced for each LIWC feature for ASR generated transcriptions, 11 out of the top 21 features were found to significantly differ between human and ASR generated transcripts. This indicates that the sparser a feature is, the less likely scaling can correct the data.

<Table 4 here>

Discussion

The objective of this study was to determine if error prone ASR transcriptions can be used in place of human-generated transcriptions when conducting linguistic analyses on speech. This question was motivated by the widespread use of linguistic analysis in social science research and the potential efficiency of using automated transcription as part of the methodological process where manual transcription is prohibitive. If the efficacy of analysis using automated transcriptions is comparable to that of analysis performed with manually transcribed speech, ASR would significantly increase the efficiency of research using transcribed speech.

Our evaluation offers several key insights. First, our analyses indicate that at the individual feature level, significant differences exist between human and machine transcription. For word count, ASR services consistently underestimate the amount of spoken content. On average, human transcribed speech included a significantly higher word count. However, the ASR services contained more unique terms and often produced higher LIWC measures. These results demonstrate that when the same word is spoken multiple times, ASR is transcribing the word in multiple ways.

The most striking difference between transcripts was found when comparing LIWC output. Across the three ASR services and 75 LIWC features analyzed, on average 66.22% of features significantly differed compared to features generated using human transcriptions. In most cases, LIWC measures were inflated for ASR transcriptions. The difference in feature measures from ASR transcripts may greatly impact outcomes for social scientists who attempt to understand human behavior through word-level analysis – that is, when explanation of the model is important along with prediction (R. Agarwal & Dhar, 2014). Transcripts produced using ASR are likely to inflate linguistic measures, which may artificially boost effect sizes of subsequent analyses.

Second, our evaluation of machine learning algorithms trained using Google ASR and human transcription revealed that performance did not vastly differ, despite the imperfect transcriptions. This is consistent with the findings of previous research in topic modeling and information retrieval. Given that a large proportion of text features significantly differed between human and ASR transcriptions, the overall competitiveness of machine learning performance between the transcript types was somewhat surprising and suggested that ASR consistently produces errors between transcripts. Error consistency appears to produce a standardization effect, which leads to similar machine learning performance between transcript types.

We see the implications of this research as two-fold. First, our results suggest that researchers considering using ASR as part of their methodological process should do so with caution. ASR largely overestimate measures which may affect downstream analyses. Our results also indicate that standardizing LIWC features will closely align measurements between human and ASR transcriptions. However, at higher levels of sparseness this standardization may not be reliable. We suggest that investigators carefully determine if sparse features should be included in analyses when using ASR. Very likely, these features will add more noise than signal. Second, our findings indicate that error prone ASR transcriptions can be used for machine learning tasks without greatly deteriorating performance. We attribute this to error consistency between transcriptions and a standardization effect where the rank of metrics does not substantially change between human and ASR feature sets.

We suggest that researchers who are considering using ASR establish error rates and determine the consistency of these errors for a subset of their audio-sources prior to relying on ASR transcriptions. This method involves transcribing a subset of audio recordings manually and automatically, and then evaluating the consistency of feature differences before selecting ASR as a viable approach. We view this recommendation as a balanced approach where the efficiencies of ASR can be applied, but with an initial investment in manual transcription.

Limitations and Future Work

There are several limitations in this study that future work should attempt to address. First, although samples with fewer than 100 participants are common in behavioral research, this may be too small a sample to identify significant differences in machine learning methods. Our use of machine learning on the current dataset was intended to provide general insights about the effects of transcription errors on prediction tasks. Future work should investigate error effects on a larger corpus where parameter optimization is more reliable. This suggestion is however resource intensive since manual transcription must be applied to create a ground-truth corpus.

Second, each participant in the current study used the same audio-visual recording equipment when responding to interview questions. The standardization effect that we propose relies on relatively consistent errors between transcriptions. The audio recordings in our study have a consistent quality since the same recording device was used for all participants in a controlled laboratory setting. It is unknown whether errors will remain consistent when ASR is applied to recordings using different equipment, environments, or in situations where multiple speakers are present. Future research should investigate ASR error consistency when recording quality is varied. If errors are not consistent between recordings, then downstream analyses are likely to suffer.

Finally, this study only considered one context (hireability). When exploring potential data for this study, the researchers found it difficult to locate a dataset that provided audio recordings, ground truth text transcriptions, and a reliable target variable. The publicly available datasets that we explored generally would not include all three components necessary for the study of ASR errors. Generally, audio recordings would not be made available along with text transcriptions and/or a measure to investigate. Future work should attempt to evaluate transcription error effects on well-studied psycholinguistic results to see if the models still hold when using inaccurate data.

Conclusion

This study reinforces previous findings that ASR accuracy rates are not as high as advertised. The speech recordings in the current dataset were captured in a controlled laboratory setting with no background noise and a consistent, high-quality equipment setup. The quality and consistency of the audio recordings used in this study underpins the importance of understanding the impact of errors on linguistic analysis. Likely, the quality of *in-the-wild* recordings will be much lower than what we used in this study. We cannot assume that ASR will meet the accuracy rates of humans in the short-term. Speculatively, long-term ASR performance will meet or surpass that of human transcription, but our evaluations demonstrates that ASR currently struggle even with relatively high-quality recordings. In-the-wild or non-optimal audio recordings are likely to further reduce ASR accuracy rates and ASR technologies are not likely to overcome variabilities in recordings anytime soon. Researchers should not blindly apply automated speech technology and expect the same outcomes as manually transcribed speech. Although ASR affords significant increases in efficiency, researchers should build in reliability checks, such as evaluating the accuracy of a subset of ASR transcripts, into their methodological process when considering the use of ASR technology. However, despite the imperfections of ASR, this technology shows great promise for researchers who look to large audio datasets to draw inferences about human behavior.

References

Abe, J. A. A. (2020). Big five, linguistic styles, and successful online learning. *The Internet and Higher Education*, *45*, 100724. https://doi.org/10.1016/j.iheduc.2019.100724

Agarwal, R., & Dhar, V. (2014). Editorial—Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research. *Information Systems Research*, *25*(3), 443–448. https://doi.org/10.1287/isre.2014.0546

Agarwal, S., Godbole, S., Punjani, D., & Roy, S. (2007). How Much Noise Is Too Much: A Study in Automatic Text Classification. *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, 3–12. https://doi.org/10.1109/ICDM.2007.21

Araújo, C. S., Magno, G., Meira, W., Almeida, V., Hartung, P., & Doneda, D. (2017). Characterizing Videos, Audience and Advertising in Youtube Channels for Kids. In G. L. Ciampaglia, A. Mashhadi, & T. Yasseri (Eds.), *Social Informatics* (pp. 341–359). Springer International Publishing.

Bayram, U., & Benhiba, L. (2021). Determining a Person's Suicide Risk by Voting on the Short-Term History of Tweets for the CLPsych 2021 Shared Task. *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, 81–86. https://doi.org/10.18653/v1/2021.clpsych-1.8

Bazillon, T., Esteve, Y., & Luzzati, D. (2008). Manual vs Assisted Transcription of Prepared and Spontaneous Speech. *LREC*.

Burgoon, J. K. (2018). Predicting Veracity From Linguistic Indicators. *Journal of Language and Social Psychology*, *37*(6), 603–631. https://doi.org/10.1177/0261927X18784119

Cable, D. M., & Judge, T. A. (1997). Interviewers' perceptions of person–organization fit and organizational selection decisions. *Journal of Applied Psychology*, *82*(4), 546–561. https://doi.org/10.1037/0021-9010.82.4.546

Chatterjee, S., Goyal, D., Prakash, A., & Sharma, J. (2020). Exploring healthcare/health-product ecommerce satisfaction: A text mining and machine learning application. *Journal of Business Research*. https://doi.org/10.1016/j.jbusres.2020.10.043

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2017). Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1217–1230. https://doi.org/10.1145/2998181.2998213

Crowston, K., Allen, E. E., & Heckman, R. (2012). Using natural language processing technology for qualitative data analysis. *International Journal of Social Research Methodology*, *15*(6), 523–543. https://doi.org/10.1080/13645579.2011.625764

Dempster, P. G., & Woods, D. K. (2011). The Economic Crisis Though the Eyes of Transana. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *12*(1). https://doi.org/10.17169/FQS-12.1.1515

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, *129*(1), 74–118. https://doi.org/10.1037/0033-2909.129.1.74

Dernoncourt, F., Bui, T., & Chang, W. (2018). A Framework for Speech Recognition Benchmarking. *Proceedings of Interspeech 2018*, 169–170.

Dorn, B., Dunbar, N. E., Burgoon, J. K., Nunamaker, J. F., Giles, M., Walls, B., Chen, X., Wang, X. (Rebecca), Ge, S. (Tina), & Subrahmanian, V. S. (2021). A System for Multi-person, Multi-modal Data Collection in Behavioral Information Systems. In V. S. Subrahmanian, J. K. Burgoon, & N. E. Dunbar (Eds.), *Detecting Trust and Deception in Group Interaction* (pp. 57–73). Springer International Publishing. https://doi.org/10.1007/978-3-030-54383-9_4

Foley, K. A., MacGeorge, E. L., Brinker, D. L., Li, Y., & Zhou, Y. (2020). Health Providers' Advising on Symptom Management for Upper Respiratory Tract Infections: Does Elaboration of Reasoning Influence Outcomes Relevant to Antibiotic Stewardship? *Journal of Language and Social Psychology*, *39*(3), 349–374. https://doi.org/10.1177/0261927X20912460

Fuller, C. M., Biros, D. P., & Wilson, R. L. (2009). Decision support for determining veracity via linguistic-based cues. *Decision Support Systems*, *46*(3), 695–703. https://doi.org/10.1016/j.dss.2008.11.001

Glaser, A. (2017). *Google's ability to understand language is nearly equivalent to humans*. https://www.recode.net/2017/5/31/15720118/google-understand-language-speech-equivalent-humans-code-conference-mary-meeker

Ho, S. M., Hancock, J. T., & Booth, C. (2017). Ethical dilemma: Deception dynamics in computer-mediated group communication. *Journal of the Association for Information Science and Technology*, *68*(12), 2729–2742. https://doi.org/10.1002/asi.23849

Holtzman, N. S., Tackman, A. M., Carey, A. L., Brucks, M. S., Küfner, A. C. P., Deters, F. G., Back, M. D., Donnellan, M. B., Pennebaker, J. W., Sherman, R. A., & Mehl, M. R. (2019). Linguistic Markers of Grandiose Narcissism: A LIWC Analysis of 15 Samples. *Journal of Language and Social Psychology*, *38*(5–6), 773–786. https://doi.org/10.1177/0261927X19871084

Humă, B., Stokoe, E., & Sikveland, R. O. (2019). Persuasive Conduct: Alignment and Resistance in Prospecting "Cold" Calls. *Journal of Language and Social Psychology*, *38*(1), 33–60. https://doi.org/10.1177/0261927X18783474

Hung, Y.-C., & Guan, C. (2020). Winning box office with the right movie synopsis. *European Journal of Marketing*, *54*(3), 594–614. https://doi.org/10.1108/EJM-01-2019-0096

Këpuska, V., & Bohouta, G. (2017). Comparing speech recognition systems (Microsoft API, Google API and CMU Sphinx). *Int. J. Eng. Res. Appl*, *7*(03), 20–24.

Lin, M., Lucas, H. C., & Shmueli, G. (2013). Research Commentary—Too Big to Fail: Large Samples and the p-Value Problem. *Information Systems Research*, *24*(4), 906–917. https://doi.org/10.1287/isre.2013.0480

Okdie, B. M., & Rempala, D. M. (2019). Brief Textual Indicators of Political Orientation. *Journal of Language and Social Psychology*, *38*(1), 106–125. https://doi.org/10.1177/0261927X18762973

Pan, L., McNamara, G., Lee, J. J., Haleblian, J. (John), & Devers, C. E. (2018). Give it to us straight (most of the time): Top managers' use of concrete language and its effect on investor reactions. *Strategic Management Journal*, *39*(8), 2204–2225. https://doi.org/10.1002/smj.2733

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., Ungar, L. H., & Seligman, M. E. P. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, *108*(6), 934–952. https://doi.org/10.1037/pspp0000020

Paulus, T., Lester, J., & Dempster, P. (2014). *Digital Tools for Qualitative Research* (By pages 93-113; pp. 93–113). SAGE Publications Ltd. https://doi.org/10.4135/9781473957671

Pennebaker, J.W., Booth, R. J., Boyd, R. L., & Francis, M. E. (2015). *Linguistic Inquiry and Word Count: LIWC2015*.

Prati, F., Menegatti, M., Moscatelli, S., Kana Kenfack, C. S., Pireddu, S., Crocetti, E., Mariani, M. G., & Rubini, M. (2019). Are Mixed-Gender Committees Less Biased Toward Female and Male Candidates? An Investigation of Competence-, Morality-, and Sociability-Related Terms in Performance Appraisal. *Journal of Language and Social Psychology*, *38*(5–6), 586–605. https://doi.org/10.1177/0261927X19844808

Proyer, R. T., & Brauer, K. (2018). Exploring adult Playfulness: Examining the accuracy of personality judgments at zero-acquaintance and an LIWC analysis of textual information. *Journal of Research in Personality*, *73*, 12–20. https://doi.org/10.1016/j.jrp.2017.10.002

Rev.com. (n.d.). *How Long Does It Take to Transcribe One Hour of Audio or Video?* https://www.rev.com/blog/resources/how-long-does-it-take-to-transcribe-audio-video

Saon, G. (2017). Reaching New Records in Speech Recognition. *IBM*. https://www.ibm.com/blogs/watson/2017/03/reaching-new-records-in-speech-recognition/

Souri, A., Hosseinpour, S., & Rahmani, A. M. (2018). Personality classification based on profiles of social networks' users and the five-factor model of personality. *Human-Centric Computing and Information Sciences*, *8*(1), 24. https://doi.org/10.1186/s13673-018-0147-4

Spencer-Oatey, H., & Wang, J. (2019). Culture, Context, and Concerns About Face: Synergistic Insights From Pragmatics and Social Psychology. *Journal of Language and Social Psychology*, *38*(4), 423–440. https://doi.org/10.1177/0261927X19865293

Walker, D., Lund, W. B., & Ringger, E. (2010). Evaluating models of latent document semantics in the presence of OCR errors. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 240–250.

Wang, P., Yan, M., Zhan, X., Tian, M., Si, Y., Sun, Y., Jiao, L., & Wu, X. (2021). Predicting Self-Reported Proactive Personality Classification With Weibo Text and Short Answer Text. *IEEE Access*, *9*, 77203–77211. https://doi.org/10.1109/ACCESS.2021.3078052

Zhou, Y., & Fan, Y. (2013). A sociolinguistic study of American slang. *Theory and Practice in Language Studies*, *3*(12), 2209.

Ziemer, K. S., & Korkmaz, G. (2017). Using text to predict psychological and physical health: A comparison of human raters and computerized text analysis. *Computers in Human Behavior*, *76*, 122–127. https://doi.org/10.1016/j.chb.2017.06.038

Tables

Table 1: Transcription Quality

ASR Service	***Weighted Averages***			
	% Substitutions	**% Insertions**	**% Deletions**	**WER**
Watson Speech-to-Text (IBM)	11.38%	1.87%	6.65%	19.90%
Google Speech-to-Text	11.52%	1.53%	7.29%	20.33%
Amazon Transcript	9.02%	2.15%	5.69%	16.85%

Table 2: Bag-of-Word Feature Count

		Shared Features (compared to human)	
Transcription Type	**# features**	**# shared**	**% shared**
Human	1696	-	-
Google	2076	1442	85.02%
Amazon	1884	1461	86.14%
Watson	1736	1399	82.49%

Table 3: LIWC Category Comparison Between Human and ASR Transcriptions

		# sig. differences in category		
LIWC Category	**# sub-categories**	**IBM**	**Google**	**Amazon**
Affect	6	5 (83.33%)	5 (83.33%)	5 (83.33%)
Biological Processes	5	1 (20%)	5 (100%)	5 (100%)
Cognitive Processing	7	5 (71.43%)	3 (42.86%)	5 (71.43%)
Drive	6	3 (50%)	2 (33.33%)	3 (50%)
Function Words	8	5 (62.5%)	6 (75%)	7 (87.5%)
Function Words - Impersonal Pronoun	1	0 (0%)	0 (0%)	1 (100%)
Function Words - Personal Pronoun	6	4 (66.67%)	4 (66.67%)	3 (50%)
Informal Language	4	3 (75%)	2 (50%)	3 (75%)
Other Grammar	6	4 (66.67%)	3 (50%)	3 (50%)
Perceptual Processes	4	2 (50%)	1 (25%)	3 (75%)
Personal Concerns	4	4 (100%)	4 (100%)	4 (100%)
Relativity	4	3 (75%)	4 (100%)	3 (75%)
Social	4	2 (50%)	3 (75%)	2 (50%)
Summary Dimensions	7	4 (57.14%)	4 (57.14%)	6 (85.71%)
Time Orientation	3	2 (66.67%)	2 (66.67%)	1 (33.33%)
Total	75	47 (62.67%)	48 (64%)	54 (72%)

Table 4: LIWC Category Comparison After Standardization

		# sig. differences in category	
LIWC Category	**# sub-categories**	**Non-Scaled**	**Scaled Data (Google)**
Affect	6	5 (83.33%)	3 (50%)
Biological Processes	5	5 (100%)	2 (40%)
Cognitive Processing	7	3 (42.86%)	0 (0%)
Drive	6	2 (33.33%)	1 (16.67%)
Function Words	8	6 (75%)	0 (0%)
Function Words - Impersonal Pronoun	1	0 (0%)	0 (0%)
Function Words - Personal Pronoun	6	4 (66.67%)	0 (0%)
Informal Language	4	2 (50%)	2 (50%)
Other Grammar	6	3 (50%)	0 (0%)
Perceptual Processes	4	1 (25%)	1 (25%)
Personal Concerns	4	4 (100%)	1 (25%)
Relativity	4	4 (100%)	0 (0%)
Social	4	3 (75%)	1 (25%)
Summary Dimensions	7	4 (57.14%)	0 (0%)
Time Orientation	3	2 (66.67%)	0 (0%)
Total	75	48 (64%)	11 (14.67%)

Figures

Figure 1: Machine Learning Accuracy

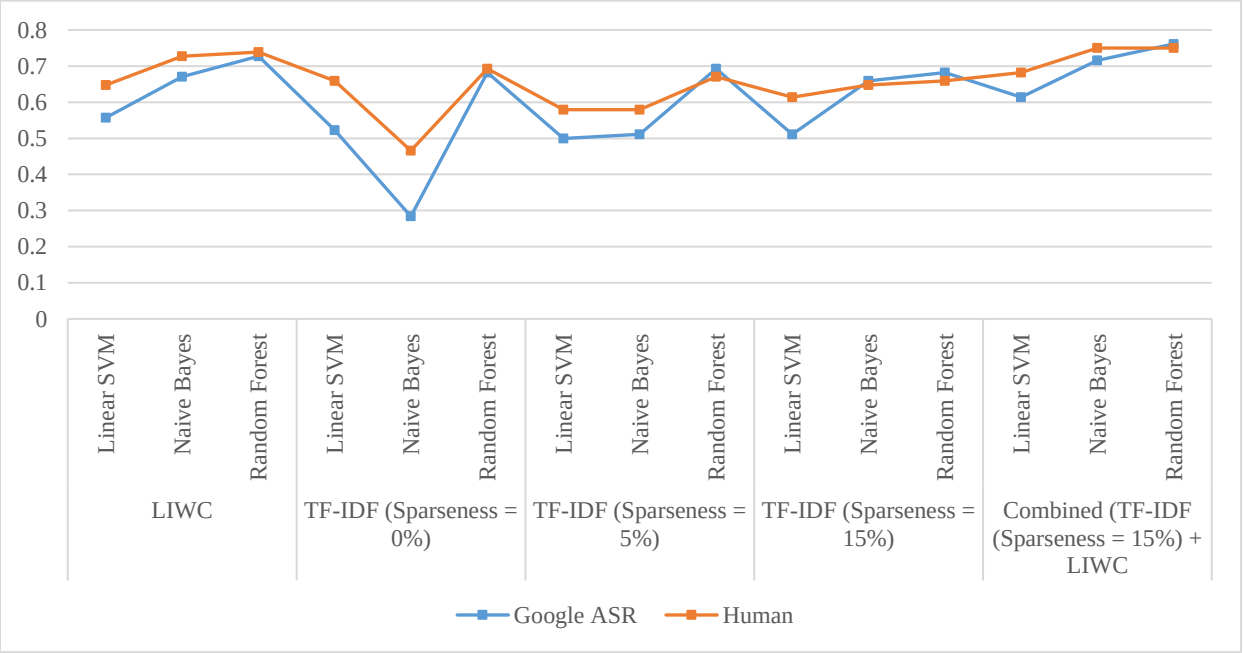

Figure 2: Frequency of Rank Correlations of LIWC Features between Google and Human Transcripts