ASR and Human Recognition Errors: Predictability and Lexical Factors

Courtney S Mansfield

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

Doctor of Philosophy

University of Washington

2021

Reading Committee:

Gina-Anne Levow, Chair

Richard A. Wright

Mari Ostendorf

Program Authorized to Offer Degree:
Linguistics

University of Washington

Abstract

ASR and Human Recognition Errors:
Predictability and Lexical Factors

Courtney S Mansfield

Chair of the Supervisory Committee:
Professor Gina-Anne Levow
Linguistics

Considering the complexity of speech communication, it is unsurprising that a listener occasionally misrecognizes an utterance. However, by examining patterns across many recognition errors, researchers can better understand the mechanisms of speech perception. Through a systematic study of automatic speech recognition (ASR) errors, it is likewise possible to better understand the state of speech processing including its strengths and limitations.

This dissertation considers lexical factors involved in speech misperception and focuses on the role of predictability, which has been less studied in previous work. It uses a set of alignments from the corrected Switchboard corpus produced by Zayats et al. (2019). By taking statistics over many instances of transcription errors, this work considers the role of predictability in speech perception.

The findings indicate that hallucinations, where a speaker identifies a word not present in the utterance (i.e. transcription insertions), and misses, where a speaker fails to hear a word (i.e. transcription deletions) tend to be relatively high in predictability. To measure this, a metric called the surprisal difference is introduced, based on linguistic surprisal (Hale, 2001; Levy, 2008) between the hypothesis and reference text. In a sentence choice task, it is found that predictability affects the sentence a listener chooses, regardless of whether the sentence

accurately matches the audio.

The second part of this dissertation considers differences between human transcription errors and ASR errors from a state-of-the-art ASR system (Xiong et al., 2016, 2017). Although the total number of errors made by humans and ASR on the same evaluation set is similar, there are differences in the constitution of these errors. The distribution of token frequency, predictability, and the surprisal difference is found to vary significantly. These distributional differences are accounted for in part by the failure of ASR to accurately recognize very low-frequency words and differences in human and ASR recognition of fillers such as filled pauses and backchannels.

This work also supports the effectiveness of speech transcription to study speech misperception, relying on a corpus of 32K errors, the largest of its kind. It features a crowd-sourcing study which demonstrates the replicability of the transcription errors under different task conditions. By considering errors in conversational transcription, this work provides a better understanding of the relationship of predictability and other lexical features to misrecognition by humans and machines.

TABLE OF CONTENTS

Page

List of Figures iii

List of Tables iv

Chapter 1: Introduction 1
1.1 Speech misperceptions 2
1.2 Predictibility in human recognition 2
1.3 Human recognition vs. machine recognition 3
1.4 Summary of key findings 4
1.5 Structure of the thesis 5

Chapter 2: Background 6
2.1 Speech misperception 6
2.2 Transcription errors 14
2.3 Summary 17

Chapter 3: Corpora 19
3.1 Introduction 19
3.2 The Switchboard corpus 20
3.3 The NIST 2000 Hub5 evaluation dataset 23
3.4 Conclusion 24

Chapter 4: Lexical predictability in speech transcription errors 25
4.1 Introduction 25
4.2 Methodology 26
4.3 Results 36
4.4 Conclusion 42

Chapter 5: Crowd-sourcing misperceptions from conversational speech 47
5.1 Introduction 47
5.2 Methodology 50
5.3 Statistical model 56
5.4 Results 60
5.5 Discussion 66
5.6 Conclusion 69

Chapter 6: Comparison of ASR and human transcription errors 76
6.1 Introduction 76
6.2 Analysis of features 81
6.3 Results 85
6.4 Discussion 93
6.5 Conclusion 99

Chapter 7: Conclusion 110
7.1 Misperception and predictability 110
7.2 Transcription alignments and misperception 111
7.3 Recognition in humans and machines 112
7.4 Applications and future work 113
7.5 Concluding remarks 115

Bibliography 116

Appendix A: Supplemental stop words list 124

Appendix B: Switchboard - most frequent words 125

LIST OF FIGURES

Figure Number		Page
4.1	Insertion and substitution errors with PTB lexical class	31
4.2	Deletion and substitution errors with MsState lexical class	31
4.3	Count of errors by sequence length	32
4.4	Model predictions using error type linear models	44
4.5	Model fit for linear models with lexical class interactions.	45
4.6	Top twenty errors by count and average surprisal difference.	46
5.1	Participant instructions.	54
5.2	Experiment user interface.	55
5.3	Count of response accuracy scores	57
5.4	Categorical model effects	71
5.5	Surprisal difference model effects	72
5.6	Diagnostic plots	73
5.7	Density of the surprisal difference means.	74
5.8	Count of participant accuracy	75
5.9	Participant agreement on stimuli	75
6.1	Proportion of errors of each type as a total of the reference tokens.	87
6.2	Counts of errors in lexical class	89
6.3	Proportion of errors in lexical class	101
6.4	Unigram logprob density	102
6.5	Surprisal density	103
6.6	Surprisal difference density	104
6.7	Surprisal difference density in insertions	105
6.8	Surprisal difference density in deletions	106
6.9	Surprisal difference density in substitutions	107
6.10	Utterance position density	108
6.11	Utterance length density	109

LIST OF TABLES

Table Number		Page
2.1	Examples of misperceptions in speech	7
4.1	Transcription error examples from PTB and MsState transcriptions	34
4.2	Description of differences between linear-mixed effects models	35
4.3	Stepwise regression of model factors.	37
4.4	R-squared for PTB and MsState models	38
4.5	Distribution of lexical types for substitutions	41
4.6	A random sample of content-content substitutions.	42
5.1	Speech error datasets	48
5.2	Mistranscription and control sentence pairs	53
5.3	Participant languages spoken	56
5.4	Predictor variables	58
5.5	Full model coefficients	61
5.6	Stepwise regression of full model	62
5.7	Comparison of reduced and full model	62
5.8	Reduced model coefficients	63
5.9	Brant test	64
5.10	Coefficients for stimuli subgroup models	65
6.1	WER for the Callhome and Switchboard portions of the transcriptions.	80
6.2	Wilcoxon test of error counts at the utterance level.	86
6.3	Error sequences - count of error types.	88
6.4	Wilcoxon test of discourse words	91
6.5	Most frequent discourse word deletions.	97
6.6	Most frequent discourse word substitutions	98

ACKNOWLEDGMENTS

It has been a privilege to work on this dissertation while receiving immense support from my mentors, colleagues, family, and friends. First, a generous thanks to my advisor Dr. Gina-Anne Levow. Her expertise and wisdom were tantamount to my success in graduate school. I looked forward to our advising meetings and always walked away with some interesting insights into research and academic life. Thanks also to Penelope for making our online meetings all the more lively. A generous thanks is due to my committee members. Dr. Richard Wright was an amazing mentor, stepping in to be my surrogate advisor when needed and generously connecting me with new opportunities. Dr. Mari Ostendorf provided indispensable feedback and a valuable perspective along the way.

Without the work of my fellow colleagues and RAs, this dissertation would not have been possible. Thank you to Drs. Trang Tran, Vicky Zayats, and Esther Le Grézause for their work and guidance related to the Switchboard alignments and language models. Thank you to Sara Ng for her crucial contribution to the analysis.

It has been a delight to work in the Linguistics department in a friendly and collegiate atmosphere. Thanks to Mike Furr, Joyce Parvi, Misha Burgess and Monica Cohn for helping me navigate the university system, without whom I would have been utterly lost. Thank you to my colleagues and generous friends for well-needed distraction and coffee trips, including but in no way limited to Dr. Kristen Howell, Dr. Leanne Rolston, Dr. Marina Oganyan, Anna Moroz, Agatha Downey, and Amandalynne Paullada.

The generosity of my family was essential. Thank you to Stan and Cindy Mansfield

and Kai and Julia Büchsenmann for providing support and shelter. There is not space enough to thank Paul Büchsenmann as he deserves for helping me to maintain my well-being and providing caregiving for our toddler. Fynn Büchsenmann, you are my joy.

Finally, there are a number of people who encouraged me to pursue the path of Computational Linguistics. Thanks especially to Edward Un and my former colleagues in Beijing. Thank you to Drs. Lisa Stifelman, Margaret Mitchell and Leslie Carmichael for each meeting me over coffee and inspiring me to start this journey.

DEDICATION

To my wonderful parents, Stan and Cindy Mansfield.

Chapter 1
INTRODUCTION

Successful communication requires many ingredients. First, a speaker must produce their intended message properly. This involves the precise coordination of different muscles across time, which must often be controlled within a dozen or so milliseconds. Of course, the listener must decode the speech. Herein lie other difficulties. There are rarely pauses to deliminate words in the speech stream. Speakers are known to repeat themselves, restart sentences mid-stream, and make any number of speech production errors. Speakers have unique dialects, accents, and other idiosyncrasies. And in addition to these complications, most conversations don't occur in a quiet, sound-proof booth, but instead listeners must contend with the forces of background music, crying babies, and other forms of noise.

It is unsurprising that listeners occasionally misrecognise speech. This thesis examines factors related to misrecognition, and considers both human listeners and machines (i.e. automatic speech recognition applications or ASR). It uses a novel method of speech transcription analysis to examine these errors. By analysing perception errors, it is possible to detect patterns which can provide valuable information. In the case of human listeners, a better understanding of these patterns can be used to inform theories about how speech is processed. Examining the kinds of errors that ASR produces can help to gauge the performance of ASR and guide its development.

In recent years, ASR has improved greatly, with some researchers claiming that their systems perform on par with human listeners (Xiong et al., 2016). And yet modern voice assistants have been known to confuse phrases like 'feed the baby' with 'defeat the baby'.[1]

[1]Credit to @yipe on Twitter for sharing this recognition error.

How is this possible? Chapter 6 provides some explanation for this discrepancy.

The studies within this dissertation focus on lexical and contextual factors involved in misperception. These features can be extracted from text with the aid of language models and taggers. Lexical predictability is the primary focus of this work. Predictability has been less explored than other factors like token frequency, but is known to play a role in misrecognition (Rubenstein and Pollack, 1963; Kalikow et al., 1977; Dubno et al., 1984). A major motivation of the current work is to extend the limited research on predictability and perception.

The following sections discuss the primary goals of the thesis in more depth. First, this thesis seeks to establish a relationship between predictability and misperception using conversational speech transcription. It also considers the validity of transcription alignments as a benchmark of human perception. Finally, human and machine recognition errors are compared, taking features such as predictability into account.

1.1 Speech misperceptions

Misperception is defined here as a discrepancy between the speech perceived by a listener and the actual utterance produced by a speaker. Studies of misperception are an important window into the mechanisms of speech perception. They are especially important in understanding lexical retrieval, or how the speaker identifies a word from their mental lexicon based on the auditory input they receive. Previous studies of misperception use laboratory data or curated misperceptions. By using conversational speech transcription, the sample size for such studies can be greatly increased. While curating misperceptions requires recognizing and remembering such instances spontaneously, transcriptions do not have such constraints and are arguably less biased. More advantages and disadvantages are discussed in Chapter 3.

1.2 Predictibility in human recognition

Predictability is the measure of the expectedness of a word given its sentence context. Several studies have confirmed that predictable words are easier to recognize in word recognition tasks

(Rubenstein and Pollack, 1963; Kalikow et al., 1977; Dubno et al., 1984). However, other studies note that frequent and predictable words are more likely to be misunderstood and have shorter durations (Zayats et al., 2019; Bell et al., 2009). The Smooth Signal Hypothesis describes a direct relationship with the predictability of a word and a reduction in prosodic prominence (Aylett and Turk, 2004). Words with less prominance are more likely to be misrecognized. This dissertation aims to better understand how predictability interacts with recognition.

Previous work has suggested that listeners will 'fill in' incomplete or erroneous speech, which is likely to lead to misperception in a case of 'graceful degradation' (Vitevitch, 2002). Because of this, it is hypothesized that mistranscribed words (cases where a transcription differs from the 'gold standard'), will be more predictable than the actual transcription. To measure this, a novel metric called the surprisal difference is proposed. It quantifies the difference in normalized surprisal of two sequences, building upon previous research on linguistic surprisal (Hale, 2001; Levy, 2008).

Crowd-sourcing is also employed in Chapter 5 to consider this relationship between predictability and misperception. Furthermore, the crowd-sourcing study works to validate the use of transcription alignments as a means to study human misperception.

1.3 Human recognition vs. machine recognition

In addition to modeling human recognition, Chapter 6 analyzes errors made by an ASR system. Machine recognition directly employs predictive language models to identify speech. Therefore, it is likely that a relationship between predictability and errors is likely, and it may pattern similarly with human transcription errors. However, differences in patterns of human and machine errors may be particularly revealing.

While previous studies have looked at the linguistic features of ASR errors (Greenberg and Chang, 2000; Goldwater et al., 2010), this thesis contributes a side-by-side analysis of human and machine recognition performance. In 2016, researchers alleged to have achieved human parity, in other words, claiming that speech recognition reached human levels of performance

(Xiong et al., 2016). However, this claim raises more questions than answers. First of all, what is the state of human performance? And, how can we effectively judge ASR against human performance? While these questions do not have immediate answers, the final study in this thesis points out clear similarities and differences between ASR and human recognition errors. These differences identify both limitations and advantages in machine recognition compared to that of its human counterparts.

1.4 Summary of key findings

This dissertation consists of three primary studies which examine the above research questions. A short summary of the related findings are presented here. In the analysis of human transcriptions, there are several key findings:

- Hallucinated and missed words tend to be highly predictable compared to their respective reference tokens (see Chapter 4).
- Listeners are more likely to misrecognize sentences when the wrong alternative is also more highly predictable (see Chapter 5).
- Function word errors are most numerous by absolute count, but filler words, like hesitations and backchannels, have the highest error rate. Professional transcribers missed 10% of these words (see Chapters 4 and 6).

Chapter 6 compares ASR and human transcriptions, and uncovers significant insights regarding ASR:

- While ASR is better able to recall tokens, it is more likely to insert or substitute erroneous words.
- Differences in the frequency distributions of ASR errors show that low-frequency content words are misrecognized much more often by ASR.
- Although ASR is better able to disambiguate filled pauses, it often confuses acoustically similar words that function as hesitations and backchannels.

1.5 Structure of the thesis

This thesis consists of 7 chapters, including the introduction. The following, Chapter 2, presents a review of literature regarding misperceptions and their related lexical features. It also showcases research on human and machine transcription performance and error analysis.

The next three chapters consist of three related research studies. Chapter 4 models predictability and the surprisal difference in human transcription alignments. Chapter 5 uses crowd-sourcing to replicate transcription errors, and examines the relationship between predictability and error rate. Chapter 6 provides an analysis of linguistic features in human and machine recognition errors.

Finally, Chapter 7 presents the conclusion. It summarizes the findings of the three aforementioned studies and closes with a discussion of possible future directions.

Chapter 2
BACKGROUND

This chapter begins with an overview of work focused on speech perception errors. Studies of perception errors and characteristics which relate to these errors have been used to inform theories of speech perception in the field of psycholinguistics. While many studies examine the relationship between acoustic characteristics and misperception (see for example Miller et al. (1951); Wang and Bilger (1973); Bond and Garnes (1980); Browman (1980)), this thesis focuses on lexical properties which are associated with speech errors. A number of studies have found that properties of the word and its surrounding context can influence the intelligibility of the word.

The next part of this chapter focuses on studies describing errors made by human transcribers and ASR systems from conversational telephone speech (CTS) transcription. Recognition errors in transcription will be the focus of the current thesis. There is a small body of work examining various linguistic features associated with human or machine recognition errors.

2.1 Speech misperception

Speech misperceptions are defined here as departures by the listener from the intended utterance of the speaker. Misperceptions differ from production errors in that the speaker's intended and actual utterance are identical. However, the listener's perception of the utterance is different; the listener is unable to accurately recognize the utterance. There are several different types of misperceptions which will be referred to in this thesis. 'Substitutions' occur when a listener mishears the speaker's intended word, and perceives another (often characteristically similar) word in its place. Substitutions are also commonly referred to as

	Hypothesis	Reference
Hallucination	**a** couple months	couple months
Miss	i didn't know they had bands playing at the main	i didn't know **that** they had bands playing at the main
Substitution	it's kind of **like for me** it's really crowded here in town now	it's kind of **nice i mean** it's really crowded here in town now

Table 2.1: Examples of misperceptions including hallucinations, misses, and substitutions. Differences between the hypothesis and referenced are emphasized in red.

'slips-of-the-ear' in the literature (Garnes and Bond, 1980). In laboratory studies, listeners often hear words in isolation, and therefore such studies tend to focus on the substitution of the target word. However, in the context of an utterance, other types of misperceptions are possible. Listeners may fail to perceive a word or a string of words within an utterance, which will be referred to as a 'miss' error. In other cases, a listener may erroneously perceive an additional word that was not a part of the speaker's intended utterance. This type of error will be referred to as a 'hallucination'. Examples of misperception errors from transcriptions of conversational telephone speech are provided in Table 2.1.

Examinations of speech misperceptions make a significant contribution to theoretical models of word recognition. Of significant interest to linguists are questions of how the mental lexicon is organized, and how words are accessed in the process of recognition. For instance, early studies of word recognition demonstrated that words that appear more frequently in a language are misheard less often than rare words (Howes, 1957; Savin, 1963; Broadbent, 1967). Because of this, models of speech perception typically include some mechanism in which frequent words have 'privileged access' in the lexicon. For example, later versions

of the TRACE word recognition model added a bias component related to word frequency (Dahan et al., 2001). The following sections will consider word frequency and several other characteristics that are notable in the speech misperception literature.

2.1.1 Misperception and Frequency

One of the earliest findings from studies of perception in noise is that the frequency of a word, judged by some measure of a word's prevalence, affects its recognition. Laboratories studies suggest that higher frequency words are easier to recognize (Howes, 1957; Broadbent, 1967). The speed of word recognition in reading tasks shows similar effects (Solomon and Postman, 1952). This finding has been described as the 'word frequency effect' in the literature.

More recent studies of misperception in naturalistic environments complicate the word frequency effect. Vitevitch (2002) presented 241 curated misperceptions. These misperceptions were collected by the authors and collaborators as they were noticed in the real world. They found that misperceptions in the corpus had generally high word frequencies according to a measure of English usage. A study of transcription errors in Switchboard showed similar results (Zayats et al., 2019). Transcribers were more likely to misperceive high frequency words.

Duration or word reduction may explain why more frequent words are often misunderstood. Vitevitch (2002) used word-recognition tasks where high- and low-frequency words were played with an artificially manipulated duration. Regardless of word frequency, a shorter duration resulted in more errors. The authors suggest that their set of curated misperceptions consisted of higher frequency words because of the tendency of speakers to produce high-frequency words with a shorter duration. However, predictability may be an overall better correlate of word duration. (This will be discussed more in the following section.)

The complex relationship between frequency and word perception is also captured in work which marries acoustic similarity and word frequency within the lexicon. Savin (1963) studied the confusability of isolated words in noise. Participants were played a variety of words at different signal-to-noise thresholds and wrote down the word they believed to have heard.

The results showed that infrequent words were more likely to be confused only if they were acoustically similar to more frequent words in the language. Frequency reflected both the recognition rate of certain words and the substitutes that participants choose to replace a misheard word. Participants also tended to hypothesize the same acoustically similar set of frequent words. The neighborhood activation model (NAM) for speech recognition takes into account both acoustics and frequency-related factors (Luce and Pisoni, 1998). One component of their model is the acoustic confusibility between a target word and acoustically similar competitors (determined using confusion matrices). These competitors are part of a target word's lexical 'neighborhood'. The frequency of the target word and the frequency of the competitors are also taken into account. The final model is correlated with tasks including identifying words-in-noise and repeating isolated words.

2.1.2 Misperception and Predictability

Although many studies use words in isolation, listeners in most everyday settings will hear words in the context of a sentence, and sentences in the context of a conversation. It has been shown that words can be identified more accurately if they are presented with sentential context (Miller et al., 1951). Of course, the target word and the content of the sentence is important. American English speakers are likely to predict that 'cookies' is the next word in the sentence 'Santa Claus drank the milk and ate the _____'. In the field of psycholinguistics, this type of word prediction exercise is known as a cloze task. In the sentence above, 'shoes' would be unexpected due to its semantic context, and 'tasty' due to its syntactic mismatch. Predictability is defined here as the expectedness of a word (both syntactic and semantic) given its context.

Predictability has been shown to be a meaningful factor in word recognition. Rubenstein and Pollack (1963) determined that predictability, like frequency, has a strong correlation with intelligibility in speech-in-noise tasks. When signal-to-noise ratios were fixed, predictable words were easier to recognize. Additional studies have confirmed this effect (Kalikow et al., 1977; Dubno et al., 1984). The idea that listeners actively predict words during the sentence

comprehension process is somewhat controversial. However, some event-related potential studies (ERP) have supported the idea that listeners are actively making predictions about future words during the comprehension process. Wicha et al. (2003) had participants listen to Spanish sentences with target nouns which were semantically predictable (e.g. 'Little Red Riding hood carried the food for her grandmother in a **basket**'). However, the authors manipulated the article which preceded the noun. Spanish has grammatical gender, where articles and nouns show grammatical agreement. Hearing articles that mismatched the predictable noun elicited an N400 response. The N400 response is known to associate with sentence processing difficulty. Notably, these responses occurred prior to hearing the target noun.

While the above studies link predictability to better recognition, there is reason to believe predictability may correlate with reduced performance in other circumstances. Aylett and Turk (2004) propose the Smooth Signal Hypothesis, which suggests that predictability has a direct relationship with prosodic prominence. They claim that speakers balance production in a noisy signal with the effort needed to produce speech. Efficient speakers will therefore reduce more predictable speech, and emphasize less predictable speech. Bell et al. (2009) support these findings, although they suggest that the relationship between predictability and prominence is complex, where word frequency and other factors interact. The Smooth Signal Hypothesis may help to explain the frequency-related findings in Vitevitch (2002).

2.1.3 Misperception and surprisal

In this thesis, 'predictability' is measured with linguistic surprisal. Presently, a more detailed description of surprisal is warranted. Surprisal is calculated from a probability distribution over the predicted next words in a sentence. Linguistic surprisal is simply the log-linear inverse of the probability. Surprisal is shown in Equation 2.1:

$$H(w_i) = -\log\ p(w_i|c(w_i)) \tag{2.1}$$

where $p(w_i|c(w_i))$ is the conditional probability of a word w_i given context from previous words $c(w_i)$. Surprisal theory aims to directly capture sentence processing difficulty in an incremental fashion (Hale, 2001; Levy, 2008). The surprisal metric is a complexity measure in the tradition of Information Theory (Shannon, 1948). Sentences are random events composed of symbols that carry a fixed amount of information. In other words, surprisal captures the information load of the word in a sentence. Surprisal has been especially successful in predicting reading time (Boston et al., 2008; Monsalve et al., 2012; Roark et al., 2009) and has a strong relationship with the N400 response in ERP studies (Frank et al., 2015; Michaelov and Bergen, 2020).

When quantifying surprisal, one important choice lies in how to quantify the conditional probability of a word. While early studies of predictability used cloze tasks (or their own judgements on what constitutes predictable or unpredictable words), recent studies use computational models of language. Language models are statistical models of word sequences. N-gram models are simple models that estimate sequences of words by conditioning their predictions on several previous tokens in a sentence. For instance, a bi-gram model conditions the prediction of w_i on previous words (w_{i-1}, w_{i-2}). Thus N-gram models carry the Markov assumption, where the probability of a word's occurrence depends only on a limited set of previous words in the sentence. More recently, recurrent neural networks (RNNs) have been used for language modeling applications. This type of neural network passes information in a temporal sequence from one hidden state to the next and is theoretically able to carry information about words across the entire sentence length. In practice, RNNs have difficulty making use of information across long sequences. To solve this problem, various subtypes of RNNs such as long short-term memory networks (LSTMs) and gated recurrent units (GRUs) have been developed (Cho et al., 2014). This thesis uses GRUs to calculate probability distributions. GRUs are a type of RNN which use gating, a mechanism to control the flow of information across the sequence. A more detailed description of the GRU with formulations can be found in Section 4.2.4.

Studies of surprisal have examined how effectively various types of language models

capture empirical data. Frank and Bod (2011) used three types of models to compute surprisal and relate these to reading times. The models included N-grams and RNNs, as discussed above, but also phrase-structure grammars (PSGs). PSGs consist of production rules and probabilities associated with a treebank, which can capture heirarchichal structure. However, they found that surprisal values from sequential models (RNNs) provided the best fit. Later studies showed that combining hierarchical and sequential information (Goodkind and Bicknell, 2018), or using an interpolated model of RNNs and N-grams, can boost performance. Still, the RNN models in Goodkind and Bicknell (2018) had a similar effect size to the best-performing models, suggesting that RNNs are a reliable method for calculating surprisal.

2.1.4 Misperception and Disfluencies

Spontaneous speech differs from written forms in several ways. Disfluencies are one notable feature of speech. Disfluencies are words or pauses that interrupt the flow of speech without providing propositional content. Disfluencies may consist of words like repetitions, filled pauses, and false starts. These words make up around 6% of spontaneous speech (Fox Tree, 1995). Most disfluencies follow a pattern of reparendum, editing phase, and repair (Shriberg, 1994). An example is shown 1:

(1) and $\underbrace{\text{we're}}_{\text{reparendum}}$ $\underbrace{\text{uh}}_{\text{edit}}$ $\underbrace{\text{they're}}_{\text{repair}}$ supposed to meet

Disfluencies have been shown to both aid and hinder the perception of an utterance, depending on their type. Speech processing does not seem to be hindered by repetitions, but false starts (where speakers must 'start over' following previous word(s) in an utterance) can slow processing (Fox Tree, 1995). Filler words in the editing phase (i.e. 'uh') help listeners to process a repair more quickly than pauses alone (Brennan and Schober, 2001). Disfluencies are also known to signal information about an utterance. For instance, they can be used to efficiently signal new words in the discourse (Arnold et al., 2003).

Although listeners may use disfluencies to comprehend speech, whether or not they are consciously aware of them is another matter. This study also examines the misperception of

a set of disfluency-related words such as filled pauses. These types of words are not regularly found in written speech or in captioning. They tend to be short in duration. Therefore, it is not surprising that transcribers may have some difficulty in recalling such words. This study considers disfluency and other filler words in a shared lexical category. Based on the POS-tagging conventions of Taylor et al. (2003), it includes filled pauses ('uh', 'um'), interjections ('oh', 'wow'), discourse markers ('i mean', 'like'), acknowledgements ('yeah', 'okay'), and backchannels ('uh-huh', 'mm-hm'). Previous work has shown that these kinds of words are more likely to be misperceived in transcription (Grezause, 2017; Zayats et al., 2019). Grezause (2017) finds that filler words are especially likely to be the subject of speech errors, and that the filled pause 'uh' is more likely to result in an error than its counterpart 'um'. Based on this result, she suggests that 'um' has a stronger role in marking discourse.

2.1.5 Misperception and Grammaticality

While this study does not set out to understand the contribution of grammaticality to misperception directly, it will be discussed briefly here due to the possible interaction between the predictability of a word and its grammatically. Briefly stated, when a speaker produces a grammatically nonstandard sentence, it is likely to be less predictable. Previous work suggests that listeners' hypotheses preserve grammatical structure, and they will hallucinate words in order to do so (Bond and Garnes, 1980). In shadowing studies, listeners often fail to notice erroneous structural or semantic input (Lackner, 1980). Graceful degradation is a strategy which allows a listener to 'fill in' incomplete or erroneous input in the speech signal (Vitevitch, 2002). This strategy may lead to transcriber error in cases where listeners change input to conform to their mental grammar.

2.1.6 Misperception and Other Contextual Factors

This thesis also explores several other context-related factors in misperception. One feature is sentence length. The idea that longer sentences will be more difficult to recall and therefore lead to a greater number of errors is intuitive. Miller and Selfridge (1950) look at sentence

recall using ideas from Information Theory (Shannon, 1948). Their study shows an advantage to recalling a sentence over a list of random words. Miller (1956) later suggests that sentences are broken into chunks of predictable words. While the length of a list may have a linear relationship with recall, the predictability of various words affects chunking and will interact with length and recall.

The word position of an utterance (in relation to the beginning, middle, and end of an utterance) is also studied in this thesis, but there is limited research on the topic. A related factor is the recency effect, which is well-known regarding certain tasks like list recall (Glanzer, 1972). Items later in the sequence are easier to remember. Where utterances are chunked and stored in memory, the recency effect would suggest that the most recent chunks will be easier to recall and therefore less prone to error.

2.2 Transcription errors

While most of the previous work discussed here considers misperceptions from laboratory studies or curated corpora, the current thesis uses alignments of corrected transcriptions to examine misperception. For a discussion of some of the advantages and disadvantages of this method, see Chapter 3. This section discusses how transcriber errors are measured and reports notable findings related to human and machine recognition errors in CTS.

2.2.1 Measuring transcriber error

Human and machine recognition errors are typically quantified using Word Error Rate (WER), which is a measure of the insertions, deletions, and substitutions needed to transform the hypothesis to reference, normalized by the number of words in the reference. WER is derived from the Levenshtein Edit Distance. The formula for WER is shown in Equation 2.2.

$$WER = \frac{S + D + I}{N_r} \tag{2.2}$$

While WER is widely used, it is also subject to criticism. It is difficult to interpret;

because of potentially large numbers of insertions, WER can surpass unity (i.e. exceed 1). It may not necessarily correspond with improvements in downstream NLP applications such as language understanding (Wang et al., 2003). Some work suggests weighting errors based on the saliency of the token or weighting the importance of insertions, deletions, and substitutions (Mishra et al., 2011). However, it is difficult to weight word errors by token because substitutions do not map to a single token. McCowan et al. (2004) proposes treating the recognition task as an information retrieval problem to address the issues of weighting and interpretibility. Chapter 6 attempts to understand some of the differences between human and machine transcriptions which may be obfuscated by the WER metric.

2.2.2 Human recognition

The WER for human transcription varies depending on the domain, speaker, experience and training of the transcriber, and so on. There are various estimates for WER on conversational speech. The RT-03 evaluation set contains subsets of the Fisher corpus and Switchboard, totaling 76K words. This work showed a WER of 4.1-4.5% for careful transcription, which can be produced at a rate of 50x the duration of audio (Glenn et al., 2010). Quick transcription has a WER of 9.63% at 5x the duration of the audio. The NIST 2000 Hub5 CTS dataset (Przybocki and Martin, 2001) has been used to benchmark ASR systems (Stolcke and Droppo, 2017; Xiong et al., 2017; Saon et al., 2017). It contains 21K words of Switchboard and 22K words of Callhome. Stolcke and Droppo (2017) note a WER of 5.9% and 11.3% on Switchboard and Callhome respectively. However, more careful transcriptions produced by IBM cite a WER of 5.1% and 6.8% (Saon et al., 2017).

Glenn et al. (2010) considered the disagreement rate between transcribers using SCLITE. In conversational telephone speech, they note that 95% of disagreements in careful transcription are judgement calls related to contractions, areas of difficult speech, and disfluencies. There is little additional work on the quality and features of transcription errors. The current thesis is motivated to expand this gap in the literature.

2.2.3 Machine recognition

Advancements in deep learning have led to significant gains in performance of ASR systems and associated reductions in the WER in the last years. Microsoft's system, using convolution and LSTM acoustic modeling and recurrent neural network language modeling, achieved state of the art performance in 2016 (Xiong et al., 2016, 2017). Their reported error rate was 5.8% and 11.0% on the Switchboard and Callhome portions of NIST 2000. With several improvements to the architecture, IBM later reported error rates of 5.5% and 10.5% (Saon et al., 2017). These are major improvements over the best WER reported on the same dataset in 2000, at 19.3% and 31.4% (Fiscus et al., 2000).

Some work has been done looking at the linguistic features of ASR errors. Goldwater et al. (2010) did an extensive study of ASR errors using a then state-of-the-art ASR system. They used mixed-effect modeling to model various lexical and prosodic features related to ASR errors. They found that the unigram log probability of a word was inversely related to the IWER (the individual WER attributed to each word type). Moreover, closed-class words, words with shorter duration, and words at the beginning of a turn had a high rate of error. Disfluencies also affected WER in various ways. In repetitions, the reparendum was more likely to result in error, but the repair less likely. Backchannels were correlated with high error rates. Other studies have examined linguistic features but have focused primarily on pronunciation factors. Greenberg and Chang (2000) used the phonetic-segment output and various components of eight ASR systems to investigate the association of factors like pronunciation discrepancies with insertion, deletion, and substitution errors.

A limited number of studies have directly compared ASR and human recognition errors. Vasilescu et al. (2011) looked at near homophones in English and French. They found that humans were able to disambiguate near homophones 5-6 times better than ASR. However, listeners performed worse on near-homophones that were misrecognized by ASR. Stolcke and Droppo (2017) looked at the difference in human and machine performance in Microsoft's ASR system, which claimed to reach human parity in 2016 (Xiong et al., 2016). Their study

highlighted similarities between human and machine recognition performance on the NIST 2000 set. They found that human and ASR performance was correlated at the level of the speaker. They also provided the 10 most frequent errors across the ASR and human transcriptions, which were found to be similar in support of their claims. However, the top token counts were not very high, ranging between 4 and 45 instances, calling into question their statistical reliability. Many of these tokens were not surprising given the literature on speech errors; for instance, 'a' and 'the' were a common substitution pair for both humans and machines. However, focusing on this small set of tokens may obscure differences in other parts of the distribution. Chapter 6 looks in more detail at the differences in features between these human and machine recognition errors in the Stolcke and Droppo (2017) transcriptions.

2.3 Summary

This chapter provided an overview of misperception errors and transcription errors. It identified three main types of misperceptions under study: hallucinations, misses, and substitutions. Frequency is shown to have a complicated relationship with misperceptions, although naturalistic data showed that high-frequency words are more easily confusable. Predictability is also shown to have some effect on word recognition. Disfluencies can affect speech perception, either helping or hindering recognition depending on the circumstances. Disfluent and filler words are notably more likely to be misperceived than other kinds of words. Other factors related to the error context were less well studied, such as utterance length and utterance position.

Previous work has also focused more specifically on transcription errors. Using WER, human performance is quantified as a benchmark to assess ASR systems. Human disagreement in CTS was shown to relate to judgement calls regarding ambiguous areas in the signal. A greater number of studies have examined the transcriptions produced by ASR. Features such as duration, probability, lexical class, sentence position, and disfluency are shown to correlate with error rates. A few studies have compared human and machine recognition errors directly, but these were more limited in scope. Building off of the areas of research detailed in this

chapter, the current thesis attempts to provide additional insights into human misperception and ASR recognition in the space of lexical and probabilistic factors. The following chapter will introduce the corpora used for these investigations.

Chapter 3

CORPORA

3.1 Introduction

This thesis uses several datasets of conversational telephone speech (CTS) to examine speech recognition errors. Finding corpora that can address speech perception, and especially the perception of speech in context, is a challenge. There is a divide in how data is collected in investigations of speech perception between carefully controlled studies and 'naturalistic' environments. These corpora fall on the naturalistic side of the spectrum.

3.1.1 Motivation

There are several advantages to using CTS to study perception. In laboratory studies, listeners typically perceive isolated words (or sentences). In contrast to this, the transcribers of CTS perceive words within a context of utterances and within a conversation. This is similar to 'everyday' speech perception and allows for the study of how context can aid in lexical retrieval.

CTS error sets may be less biased than other naturalistic error datasets. An alternate way to collect naturalistic data is to use 'found' speech errors (see Bond and Garnes, 1980; Browman, 1980; Vitevitch, 2002). In these corpora, the authors and their collaborators document speech errors that they overhear in their day-to-day lives. However, these samples are biased towards words that are semantically important or words that have a high information density. For example, the authors are more likely to notice the misperception of content words (e.g. 'jazzercise' vs. 'exercise'). They would be less likely to notice the confusion of function words like 'a' and 'the', although this error is commonly seen in CTS.

Another advantage is that CTS data offers a magnitude of samples. The Switchboard

error alignment contains 32K errors, while Bond and Garnes's (1980) pioneering corpus of found errors contains 1K examples. A larger sample size allows for statistical modeling of more subtle relationships in the perception of language.

However, there are disadvantages to using CTS data. Transcription is a specialized task which has both perception and production components. Production errors such as typos may add noise to the data. Upon manual inspection, these errors appear to be quite infrequent. The Switchboard transcriptions have also had multiple passes to correct for such errors. Some transcription errors may also be the result of stylistic differences. Non-speech sounds (e.g. uh, eh, or ah) or compound words (e.g. doghouse, dog house) may have a variety of written forms. Great care was taken to reduce these stylistic differences where possible.

3.1.2 Overview

The first dataset, the Switchboard corpus (Godfrey et al., 1992), is a dataset of conversational English which has generated several versions of speech transcripts. Inconsistencies between the transcriptions allow for a fine-grained analysis of recognition errors. The University of Washington's Transformation, Interpretation and Analysis of Language lab (TIAL) produced an alignment of two versions of Switchboard with better and worse accuracy levels (Zayats et al., 2019). The alignments are used to analyze human recognition errors at a large scale.

A second dataset is used to compare recognition errors made by humans and automatic speech recognition. The NIST 2000 Hub5 conversational telephone speech evaluation set (Fiscus et al., 2000) includes Switchboard data and data from the CallHome corpus (Canavan et al., 1997; Kingsbury et al., 1997). Stolcke and Droppo (2017) generated additional transcriptions from the NIST 2000 subset with then state-of-the-art automatic speech recognition. The following sections describe the two datasets in detail.

3.2 The Switchboard corpus

The Switchboard corpus (Godfrey et al., 1992) represents one of the largest available collections of English conversational telephone speech. The corpus was collected by Texas Instruments

in 1990-1991 and released in 1992-1993. In 1997, the corpus was corrected and released as Switchboard I release 2, and the following text refers to this version of the corpus. The original motivation for the corpus was to elicit natural speech for the development of speech algorithms such as speaker authentication. The corpus consists of 260 hours of speech and approximately 3 million words (Calhoun et al., 2010).

3.2.1 Corpus design

Approximately 2,400 conversations were recorded from 543 male and female speakers from across the United States. Prior to participation, demographic information including age, sex, level of educational attainment was collected. Participants called into an automated system and were matched with a conversation partner (a stranger). They then chose from a pre-determined list of conversational topics. For the length of the conversation, speakers were recorded synchronously in separate channels. Switchboard conversations are between 1.5 and 10 minutes long, with an average duration of 6.5 minutes (Calhoun et al., 2010).

3.2.2 Transcriptions

The Switchboard corpus has prompted multiple sets of transcriptions and linguistic annotations. The following sections discuss the two transcription versions that are used in this thesis.

Penn Treebank

The Penn Treebank project was an effort to create a large-scale corpus of English with POS tags and syntactic annotations. The original Penn Treebank provides annotations for written text and transcribed broadcast speech, such as radio and newswire transcriptions (Marcus et al., 1993). The last phase of the project, Penn Treebank-3, produced POS tags and disfluency annotations for the Switchboard corpus (Taylor et al., 2003). The Treebank3 annotations include 1,126 conversations from Switchboard which have been slightly modified

from the original transcriptions.

The Penn Treebank3 (henceforth referred to as PTB) includes utterance segmentation, POS tags, and disfluency annotations. Conversations are broken into sentence-like units described as slash units due to the convention of using the character '/' for marking boundaries. Discourse markers, sentences that begin with conjunctions (e.g. and/or), and back-channels are often treated as a stand-alone slash unit. POS tags are based on the Brown corpus, where a down set of 36 tags is used. Sentences were first automatically tagged and then manually reannotated. Disfluencies have been annotated following the conventions of Shriberg (1994). The PTB transcripts contain approximately 1 million words of speech.

Mississippi State University

A second more accurate transcription of the Switchboard PTB subset was produced by the Institute for Signal and Information Processing at Mississippi State University (Deshmukh et al., 1998; Hamaker et al., 1998). In the Mississippi State (henceforth referred to as MsState) transcripts, a subset of PTB conversations was manually corrected to produce more accurate transcripts. The original work estimates an error rate of less than 2%, in contrast to the 10% error rate in the original PTB transcripts. Work by Zayats et al. (2019) indicates an error rate difference of around 5%, when excluding differences in transcription conventions.

While the original MsState transcripts do not include disfluency annotations, Zayats et al. (2019) use an automatic method to align the previous disfluency annotations from PTB. They use a BIO (beginning, inside, outside) tagging scheme to label potential areas of disfluency for inserted and substituted words, and run an automatic disfluency detection classifier on the labels. The disfluency labels were compared to hand-annotations and showed a high level of consistency (F=90.1). Their release includes an alignment of the PTB and MsState transcripts by PTB slash-units (the initial slash-units were not maintained by MsState). They also include an alignment of PTB/MsState tokens and a gloss which indicates PTB tokens that are hallucinated (inserted), missed (deleted), or substituted compared to the corrected MsState version.

3.3 The NIST 2000 Hub5 evaluation dataset

The NIST (National Institute of Speech Technology) Hub5 evaluations were introduced to encourage advancements in speech technology and to gauge the progress of these technologies (Fiscus et al., 2000). In this thesis, the NIST 2000 Hub5 conversational telephone speech evaluation data[1] (henceforth referred to as NIST 2000) is used to examine human and machine recognition errors.

3.3.1 Corpus Design

The NIST 2000 dataset contains 2 sets of 20 held-out conversations from the original Switchboard corpus (Godfrey et al., 1992) and the CallHome English corpus (Canavan et al., 1997; Kingsbury et al., 1997). While Switchboard features strangers discussing a pre-defined set of topics, the CallHome corpus consists of participants calling friends and family members and openly conversing about any topic(s) of their choosing. The data consists of approximately 5-minute segments from each conversation. A conversation consists of a pair of segmented, single-channel conversational turns for each participant in the dyad.

At the time of the NIST 2000 evaluation, the Cambridge group achieved the best performance with a WER of 19.3% for Switchboard and 41.4% for CallHome (Fiscus et al., 2000). The nature of the Switchboard data (more formal speech, set topics) likely resulted in its better ASR performance.

3.3.2 Human and ASR transcriptions

Stolcke and Droppo (2017) used manual and automatic processes to generate transcriptions of the NIST 2000 conversations, allowing for a direct comparison of the errors produced by humans and ASR systems. Human transcribers followed Microsoft's production transcription pipeline. First, a transcriber generates an initial transcription of the audio. In a second close pass, the listener corrects errors in the pre-exisiting transcription. There are several

[1]For more details see https://catalog.ldc.upenn.edu/LDC2002T43

key differences in the human transcription process compared to the LDC (e.g. Switchboard) transcription workflow. While LDC transcribers utilize the entire audio of the conversation between two speakers, the Microsoft transcribers listen to each individual speaker and their turn segments separately. Furthermore, Microsoft transcribers have a comparatively limited set of instructions with regards to transcription conventions. In order to account for this, Stolcke and Droppo (2017) uses text normalization to minimize differences between the human transcripts and scoring reference. The resulting WER on the human transcription is 5.9% on the Switchboard section, and 11.3% on the CallHome section.

The ASR transcription is generated by Stolcke and Droppo (2017). Their system uses an approach that relies on competing neural network acoustic models trained on 2000 hours of conversational speech data and various n-gram and LSTM language models for decoding and rescoring. The system achieves word error rates of 5.8% on the Switchboard section, and 11.0% on the CallHome section.

3.4 Conclusion

In summary, the two main datasets used in the study are the Switchboard corpus and the NIST 2000 speech recognition evaluation dataset. Chapters 4 and 5 are studies of human speech errors and investigate human transcriber errors related to Switchboard. Chapter 6 examines ASR and human errors and uses transcriptions from the NIST 2000 evaluation dataset.

The process of transcribing and re-transcribing such a magnitude of data requires a large amount of resources. Fortunately, there has been heavy investment in speech technology which allows us to repurpose CTS datasets for linguistic research. Notably, the corpora are centered on various dialects of American English. There is a need for additional focus on large-scale conversational speech corpora in other languages, which would aid in the advancement of both speech technology and linguistic research in other populations.

Chapter 4

LEXICAL PREDICTABILITY IN SPEECH TRANSCRIPTION ERRORS

4.1 Introduction

Studies of perceptual errors give insight into the mechanisms that drive speech recognition. Previous work in this area focuses on acoustic properties of speech errors (Miller and Nicely, 1955; Bond and Garnes, 1980; Browman, 1980) or on intrinsic properties of the error such as word frequency (Savin, 1963; Vitevitch, 2002). This study focuses on the less-studied issue of contextual predictability in speech errors.

Listeners fare better at spoken word recognition when words are provided in context (Miller et al., 1951). Models of speech perception account for 'top-down' information in the form of syntactic or semantic predictability in the speech recognition mechanism, although the nuances of this process vary across models. For instance, in the Logogen model (Morton, 1969), contextual information can select for candidates prior to acoustic input, while the Cohort model (Marslen-Wilson and Welsh, 1978; Marslen-Wilson and Tyler, 1980) incorporates such information after the acoustic input is processed. Perception models suggest that context has the ability to encourage the activation of lexical candidates, while some also express the ability of context to inhibit the selection of disharmonious candidates.

This study hypothesizes that contextual predictability will have an effect on speech errors found in speech transcriptions. Where listeners hallucinate words (i.e. insertion errors), contextual information may encourage 'phantom' words that are harmonious with the sentence. In these cases hallucinations may appear more predictable in context than the actual sequence of words that are produced by the speaker. In substitutions, the same processes would lead to more predictable substitutes in comparison to the corrected sequence of words.

While high contextual predictability may facilitate 'hallucinations' of words, it may also correlate with 'missed' words (i.e. deletions). According to information theory, unpredictable or unexpected word sequences have a high information density. For a listener, missing words which carry a high information load is particularly detrimental to speech comprehension. Where information load is high and the audio signal is noisy, transcribers may actively work to avoid missing or mishearing using strategies such as replaying a noisy audio segment. However, missing words which carry a small information load is less detrimental to efficient communication. Therefore, I hypothesize that missed word sequences will be exceptionally predictable.

The predictability of any word is naturally associated with its usage. Function words have a grammatical role in a language and a higher overall frequency than content words. Some speech perception models suggest that function words are accessed via different (and more efficient) channels than content words, although processing efficiency may be better explained by frequency of exposure (Segalowitz and Lane, 2000). In any case, function words are expected to have higher overall predictability than content words. I hypothesize that the lexical status of an error will also be associated with its relative predictability compared to the corrected form. For instance, function word insertions will be especially predictable compared to content word insertions. To summarize the hypotheses:

- Inserted words will have higher predictability than their correction.
- Deleted words will have higher predictability than the first-pass transcription (i.e. the same immediate context without deletions).
- Substituted words will have higher predictability than their correction.
- An error's lexical type will have an affect on its predictability relative to the correction.

4.2 Methodology

This study uses an alignment of two sets of transcriptions from the Switchboard project (Godfrey et al., 1992). The 'first-pass' transcription is taken from the Penn Treebank (PTB)

version of the corpus (Taylor et al., 2003) while the corrected or 'second-pass' transcription is part of the Mississippi State (MsState) version (Deshmukh et al., 1998; Hamaker et al., 1998). Additional details about the resources can be found in Section 3.

The following section addresses the extraction of error sequences, including the composition of errors and excluded datapoints. Then, a discussion of the classification of lexical categories and lexical class composition of the data follows. The process of extracting conditional probabilities and the metric of surprisal difference is introduced. Finally, the analysis of data will be discussed. Linear-mixed effects models will be used to examine the correlation between the surprisal difference and error type and lexical class to test the aforementioned hypotheses.

4.2.1 Extracting and classifying error sequences

Error sequences are extracted based on error annotation tags in Zayats et al. (2019). The annotations include insertion, deletion, substitution, and the label 'CONT'. Insertions are words which are hallucinated by the transcriber in the first-pass transcription, while deletions are tokens which are missed in the first-pass transcriptions. Substitutions include a mismatch of token(s) between the first- and second-pass transcriptions. The aligned version of the transcriptions includes a 'CONT' annotation tag which identifies stylistic differences between the PTB and MsState versions of the corpus (for example, tokenization differences such as numbers 'sixty-three' and 'sixty three' or abbreviations 'ABC' and 'A. B. C.')[1]. When stylistic differences appear in the alignment, the MsState version of the text is assumed in both the first- and second-pass datasets. This study focuses on error 'sequences' as the primary unit of interest. Sequences are groups of error tokens delimited by previous or following non-error tokens (which may include BOS or EOS). For instance, several tokens may be inserted in a transcription, or a substitution may follow a deleted token.

The final dataset consists of 31.6K error sequences. Due to the processes involved in

[1]See Zayats et al. (2019) for more discussion about the stylistic differences between transcriptions.

aligning the transcriptions, it was not possible to assign the sentence position to deletions at sentence boundaries with certainty. For this reason, 7,491 sentence boundary deletions were removed from the data. The initial alignments also mark the replacement of first-pass 'uh-huh' with second-pass 'um-hum' as a substitution type error. However, the 'um-hum' token is not considered a valid token in the PTB version of the transcription and this study considers the substitution to be a stylistic difference. 5,832 instances of this substitution have been removed from the data.

To analyze the error sequences, the full dataset is broken into two (overlapping) subsets of data: the PTB error dataset (20.4K sequences) and the MsState error dataset (28.0K sequences). The PTB dataset includes all combinations of insertions and substitutions and annotates features (such as lexical category) from the PTB version of the transcriptions. The MsState error dataset includes all combinations of deletions and substitutions but includes features from the MsState version of transcriptions. While it would be conceivable to analyze one dataset which examines the PTB features of inserted sequences and the MsState features of deleted sequences, the substituted sequences in such an analysis would pose a question. Should lexical features in substitutions be derived from the PTB or MsState transcription? In order to comprehensively examine the contribution of each part of the substitution, the dataset is split in two parts. Another way to conceive of this split is that each transcription is seperately scanned for 'mismatching' sequences, where the presence of a mismatched PTB token sequence generates a datapoint in the PTB error set, and the presence of a mismatched MsState token sequence generates a datapoint in the MsState error set.

4.2.2 Lexical category classification

Sequences are annotated with aggregate lexical class labels for function, content, 'discourse', 'other', and a mix of lexical types. The category for discourse words includes words that function as discourse markers, filled pauses, and backchannels. Many discourse tokens are included in the top 100 most frequent tokens of the transcriptions, including 'uh', 'um', 'well', 'uh-huh', 'yeah', and 'okay'. The 'other' label describes primarily incomplete words, where

transcribers were only able to recover a part of the word. The 'mix' label represents error sequences that have multiple tokens which do not conform to a single class (such as the discourse-function insertion 'uh and').

In order to determine the lexical class for the error sequence, a sequence-to-sequence classifier from the Neural Sequence Labeling Toolkit (Yang and Zhang, 2018) is trained on the full set of human-annotated POS tags from the PTB corpus. The tagger has an accuracy of 96.4% on held-out test data from PTB. The classifier combines a character sequence layer with a bi-directional LSTM word layer and a softmax output layer. The resulting POS tags are mapped onto respective lexical categories using manually-created mappings. POS tags can be ambiguous, such as the adverb, which could be considered a function word (as in 'sometime') or a content word (as in 'happily'). In order to account for this, a stop word list (see Appendix A) is used to first filter a number of function words, and the remaining tokens are mapped using their POS tags.

4.2.3 Count of error sequences

Counts of insertion and substitution errors with their PTB lexical class labels are shown in Figure 4.1. Substitution errors outnumber insertion errors by a factor of 4. Function word sequences make up the largest lexical class of errors in this dataset (n=9K), followed by discourse sequences (n=5.9K). Mixed lexical class sequences make up a comparatively small number of error sequences (n=1K), although they are the second-most common category in mixed insertion/substitution type errors. This is unsurprising as mixed error types are longer in average length ($\overline{x}$=1.6) than insertion ($\overline{x}$=1.1) and substitution ($\overline{x}$=1.1) type errors.

The counts of deletion and substitution errors with lexical class labels from the MsState transcription are presented in Figure 4.2. While the overall count of substitution and mixed type errors is identical between the PTB- and MS-labeled datasets, the distribution of lexical class varies due to differences between substituted words (e.g. the substitution of function word 'our' for content word 'air'). Function sequences (n=10.3K) are the most common category in the MsState dataset, followed by discourse sequences (n=7.5K). Substitutions again outnumber

other categories, with 75% more substitution than deletion errors. When examining mixed type errors in both datasets, the most common combination is substitution/deletion (n=1.4K), followed by substitution/insertion (n=868), and insertion/deletion combinations (n=23).

Figure 4.3 illustrates the token length of the error sequences. The majority of errors, at 87%, are single tokens. There is a tail of longer sequences, with deletion sequences in particular reaching up to 17 tokens in length. A manual inspection of the data shows that longer deletions often correspond to missed phrases at the end of the turn.

4.2.4 Language model probabilities

Neural language models are used to estimate the contextual predictability of sequences in the PTB and MsState transcription. Conditional probabilities are generated with a Gated Recurrent Unit Network (GRU) (Chung et al., 2014). The GRU has an advantage over n-gram language models, because it can capture long-term dependencies and encode information across the entire sentence. A GRU network is able to encode information from the previous context of the sentence and store it in the hidden state using an 'update gate' and 'reset gate' to determine which information will be maintained. A probability distribution for each word in the vocabulary at timestep i is generated by applying a sigmoid function to the model's hidden state h_i. Formulas for the GRU network are shown in Equation 4.1:

$$z_t = \sigma(W^{(z)}x_t + U^{(z)}h_{t-1}) \tag{4.1a}$$

$$r_t = \sigma(W^{(r)}x_t + U^{(r)}h_{t-1}) \tag{4.1b}$$

$$h'_i = \tanh(Wx_t + r_t \odot UH_{t-1}) \tag{4.1c}$$

$$h_i = z_t \odot h_{t-1} + (1 - z_t) \odot h'_t \tag{4.1d}$$

where z_t represents the update gate and r_t represents the reset gate.

The training set consists of 17M tokens from the Fisher Corpus (Cieri et al., 2004), which was chosen to match the conversational style of the Switchboard Corpus. The vocabulary size

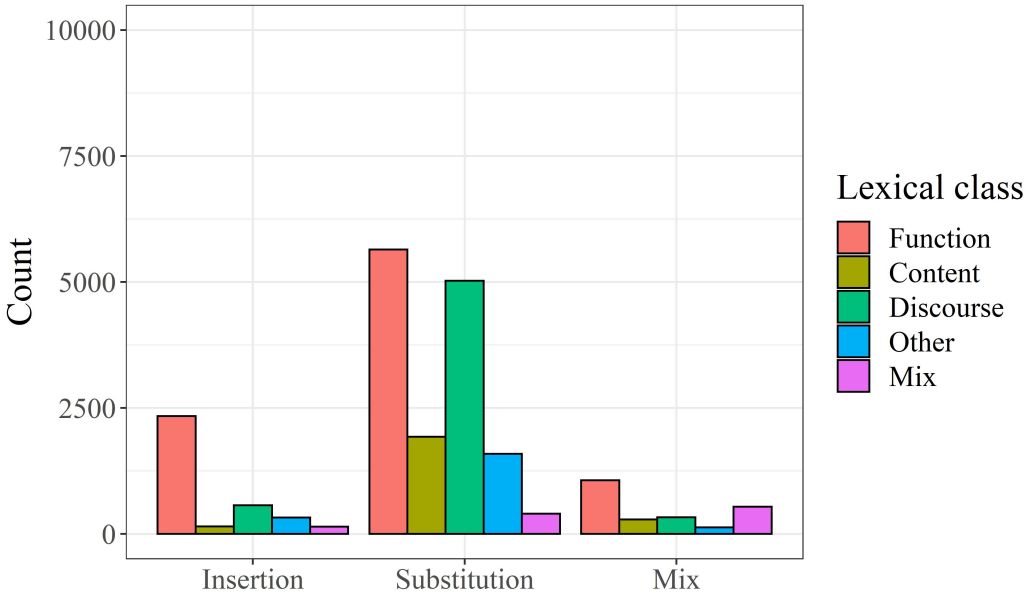


Figure 4.1: Insertion and substitution errors with PTB lexical class

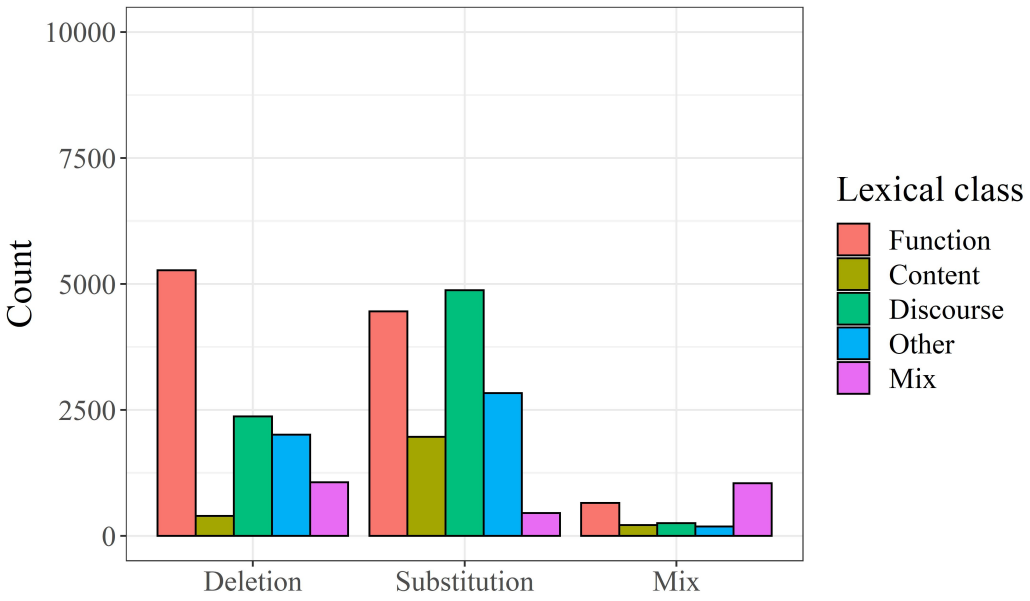


Figure 4.2: Deletion and substitution errors with MsState lexical class

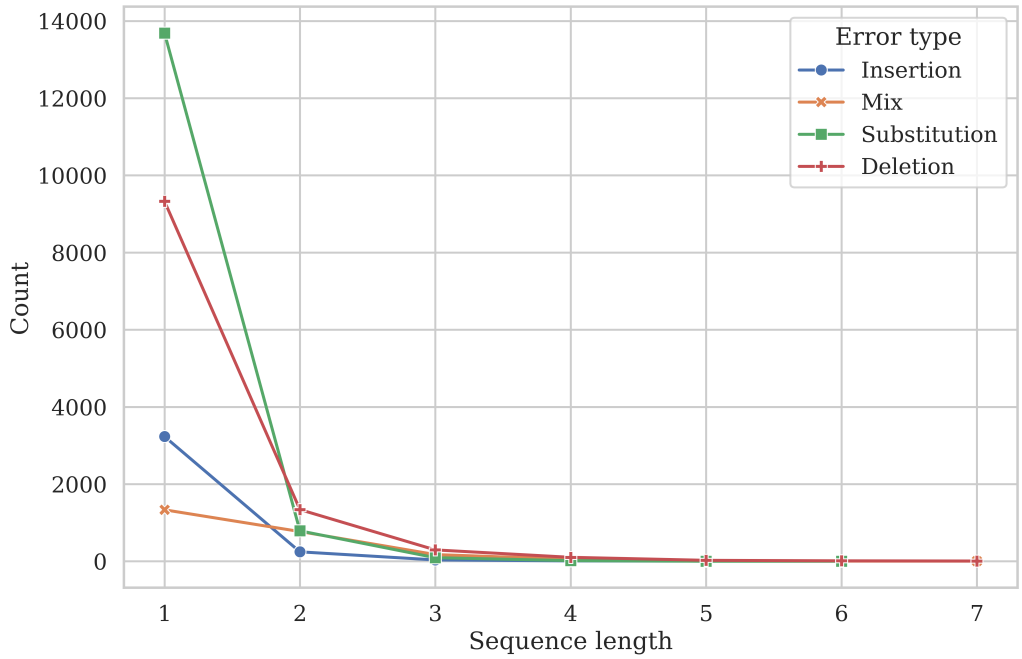


Figure 4.3: Count of errors by sequence length. Not shown are 10 deletion tokens that are greater than 7 tokens in length.

is 14K determined by a count threshold of 10. The training data includes disfluencies and the contractions are not tokenized. The bi-gram 'you know' is found to be frequently reduced and primarily used as a discourse marker, and is therefore conjoined as a single token in the training data and transcription sets.

The GRU is trained with PyTorch (Paszke et al., 2019). The model consists of 2 hidden layers, 256-dimensional word embeddings, 128-dimensional hidden layers and has a 0.2 dropout rate (Pham et al., 2014).

4.2.5 The surprisal difference

The aim of this study is to consider how the predictability of a sequence of words is distributed in a corpus of speech transcription errors. The hypotheses posed in Section 4.1 suggest that the predictibility of word sequences may be higher or lower in a first-pass set of transcription

errors compared to the second-pass transcription. The surprisal difference measure is used to capture these differences.

The formula for surprisal is shown in Equation 4.2. In this study, the conditional probability for the i^{th} word $p(w_i|c(w_i)$ is derived from the GRU language model.

$$H(w_i) = -\log_2\ p(w_i|c(w_i)) \tag{4.2}$$

This study considers sequences of errors which may be multiple tokens long. According to the chain rule, conditional probabilities in sequences are multiplied to produce a joint probability. In order to handle sequences which vary in total length, the probabilities are normalized by length as shown in Equation 4.3.

$$H(w_1 w_2 ... w_n) = -\frac{1}{n} \log_2 P(w_1 w_2 ... w_n) \tag{4.3a}$$

$$H(w_1 w_2 ... w_n) = -\frac{1}{n} \log_2 \prod_{i=1}^{n} p(w_i|c(w_i)) \tag{4.3b}$$

Next, surprisal of error sequences in the first-pass transcription is compared to surprisal of contextually matched sequences in the second-pass transcription. Contextually matched sequence pairs share a sentence position with the error. They are characterised by having a shared *prior context* $c(w_1)$ in $p(w_1|c(w_1))$ and shared *final token* w_n. Examples of these sequence pairs are found in Table 4.1.

The surprisal difference for sequences (s_{PTB}, s_{MS}) is shown in Equations 4.4. Equation 4.4a is used in the PTB insertion/substitution model. Positive values indicate a PTB transcription with lower surprisal and more predictability compared to its corrected MsState counterpart. For example, an insertion which is relatively predictable will result in a positive suprisal difference. Negative values indicate a PTB transcription with higher surprisal. Equation 4.4b is used in the MsState deletion/substitution model, where positive values indicate a PTB transcription with higher surprisal and lower predictability than its counterpart. Here, a relatively predictable deletion will result in a positive surprisal difference.

PTB transcription	MsState transcription
⟨BOS⟩ but **uh it's** so it's very very mixed ⟨EOS⟩	⟨BOS⟩ but **it's** so it's very very mixed ⟨EOS⟩
⟨BOS⟩ actual **a** ⟨EOS⟩	⟨BOS⟩ actual **actually it was a** ⟨EOS⟩
⟨BOS⟩ **fill them up** ⟨EOS⟩	⟨BOS⟩ **of foam up** ⟨EOS⟩
⟨BOS⟩ he was kind of **pushy too** ⟨EOS⟩	⟨BOS⟩ he was kind of **person who** ⟨EOS⟩

Table 4.1: Transcription error examples from PTB and corrected MsState transcriptions. The bold portions of text show contextually matched sequence pairs.

$$\Delta H(s_{ptb}, s_{ms}) = H(s_{ms}) - H(s_{ptb}) \tag{4.4a}$$

$$\Delta H(s_{ptb}, s_{ms}) = H(s_{ptb}) - H(s_{ms}) \tag{4.4b}$$

4.2.6 Linear Mixed-Effects Models

Two linear mixed-effects models are used to examine how various factors affect the difference in surprisal of the two transcriptions. Linear mixed effects models are regression models which take into account variability both from independent variables of interest and other sources. The first model examines characteristics of the PTB transcription and includes insertion, substitution, and mixed types of errors. The second model examines characteristics of the MS transcription and includes deletion, substitution, and mixed types of errors. For both models, the response variable is the surprisal difference measure. The choice of two models allows one to explore information sources in both transcription sets. For instance, when examining the effect of lexical class on surprisal in substitutions, one may want to consider both the class of the substituted word and the word that replaces it. Key differences between the two models are shown in Table 4.2.

Each model has two categorical predictor variables, error type (n=3) and lexical class

Description	Models	
	PTB model	MsState model
Surprisal difference	$H(s_{ms}) - H(s_{ptb})$	$H(s_{ptb}) - H(s_{ms})$
Error types	insertion	deletion
	substitution	substitution
	mix	mix
Lexical class labels		
Disfluency labels	from PTB transcription	from MsState transcription
Frequent word labels		

Table 4.2: A description which highlights key differences between the two linear-mixed effects models.

(n=5). There are also several control factors which are included in the model as random intercepts. To capture possible variability between transcribers, transcriber is included as a random factor (n=34). Characteristics of a word (such as frequency) will have an effect on the conditional probability of the sequence. Therefore, the word(s) in the error sequence are included as a random factor. In order to limit the complexity of the model, the most frequent 100 words in the corpus are controlled for, with a single category for all less frequent words (n=101). The most frequent 100 words can be found in Appendix B.

Transcriber errors are shown to associate with the reparandum of a disfluency (Zayats et al., 2019). Three possible control factors for disfluencies are examined. Disfluent labels are included targeting a) before the error in immediate context b) within the error itself or c) in the immediate context following the error.

The lme4 package in R (Bates et al., 2015) is used to generate mixed effect models with maximum likelihood estimation. An analysis of variance (ANOVA) using Satterthwaite

approximations for degrees of freedom (Satterthwaite, 1941) is used to confirm the significance of fixed and random effects. Stepwise regression can be misleading, particularly in the case of a large number of independent variables. However, this study uses a small number of explanatory factors which are carefully selected based on previous research. The lmerTest package (Kuznetsova et al., 2017) is used for testing. Factors are added to the null model and are expected to improve the model fit significantly ($p < 0.01$).

In the case of disfluencies, previous, current, and following disfluency labels are all highly correlated. Furthermore, an ANOVA of random factors shows that previous disfluencies and following disfluencies are not significant in either the PTB or MsState model. Therefore, random intercepts for previous and following disfluencies are removed from all models.

Two models are used to examine the hypotheses for each set of transcriptions. The first model considers information about type of error in the model, to address the general surprisal differences among different error types. The second model adds an additional consideration of lexical class and its interaction. The models are shown in Equation 4.5a:

$$\begin{aligned} \text{supDiff}_i =& \beta_1 + \beta_2 \text{errorType}_i + b_{i1} + b_{i2}\text{transcriber}_i + b_{i3}\text{frequentWord}_i \\ &+ b_{i4}\text{errorDisfluency}_i + \epsilon_i \end{aligned} \tag{4.5a}$$

$$\begin{aligned} \text{supDiff}_i =& \beta_1 + \beta_2 \text{errorType}_i + \beta_3 \text{lexicalType}_i + \beta_4 \text{errorType}_i\text{lexicalType}_i \\ &+ b_{i1} + b_{i2}\text{transcriber}_i + b_{i3}\text{frequentWord}_i + b_{i4}\text{errorDisfluency}_i + \epsilon_i \end{aligned} \tag{4.5b}$$

where $\beta_{(1...4)}$ are fixed effects, $b_{(1...4)}$ are random effects and ϵ_i is the error term for the i^{th} observation in the data.

4.3 Results

First, the results of the ANOVA are presented for the models. Model coefficients are then closely examined focusing first on the PTB insertion/substitution model and then on the MsState deletion/substitution model.

	PTB model			MsState model		
	ΔR^2adj (%)	F-value	P-value	ΔR^2adj (%)	F-value	P-value
Error type	0.55	117.44	<0.001	3.92	468.71	<0.001
Lexical type	6.18	141.41	<0.001	12.96	365.93	<0.001
Interaction:						
Error type × lexical type	1.98	59.69	<0.001	0.64	20.92	<0.001

Table 4.3: Stepwise regression of model factors.

4.3.1 Significance of model factors

To confirm the significance of the fixed effects, a forward stepwise regression is performed on error type, lexical class, and interactions for both models. Results are shown in Table 4.3. For the PTB surprisal model, the addition of either error type or lexical type to the null model is significant ($p<0.001$). The change in adjusted R-squared is also reported. Adjusted R-squared describes the proportion of variance accounted for by model predictors, adjusting for the number of predictors in the model. In the PTB model, the inclusion of lexical type explains an additional 6.18% of the variance, while error type results in a relatively small increase. The interaction term is also significant when compared to a model with error type and lexical type as main effects. The MsState model similarly found error type, lexical type, and interactions to significantly improve the model fit ($p<0.001$). Like the PTB model, lexical type captured the variance in the model more robustly than error type.

Table 4.4 shows the absolute R-squared value across models, where marginal R-squared considers the variance captured by fixed effects, and conditional R-squared includes fixed and random effects. Models with the interaction of lexical type and error type account for more variance than error type alone. The full PTB model accounts for 22.14% of variance, and MsState nearly a quarter (24.26%). The next section will discuss the model predictions in

	PTB model		MsState model	
	R^2M (%)	R^2C (%)	R^2M (%)	R^2C (%)
Error type	1.14	14.54	3.67	13.14
Error type × lexical type	8.63	22.14	12.27	24.26

Table 4.4: Adjusted marginal (M) and conditional (C) R-squared for the PTB and MsState models.

more detail.

4.3.2 Modeling the surprisal difference of insertions

The error type only model is first examined for the PTB insertion/substitution errors. Model predictions can be see in Figure 4.4a. Values above zero indicate that the PTB transcription version is less surprising than the corrected MsState version, while values below zero indicate that PTB is more surprising. Insertions and mixed insertion errors have positive means, showing that on average insertions and mixed type errors are less surprising than their MsState corrections.

Model predictions with interactions are shown in Figure 4.5a. In congruence with the high amount of variance captured by lexical effects, lexical type has a strong influence on the surprisal difference in each error category. Inserted function sequences are less surprising, which aligns with the fact that function words are high frequency words, which in many environments exhibit low surprisal. Mixed lexical type sequences are also less surprising compared to the corrections. Mixed sequences appear to behave similarly to function sequences. In fact, a function word appears in 92% of mixed sequences, where they often co-occur with a content word (69%) or discourse word (21.71%). There is not a strong positive/negative surprisal difference alignment of discourse and content sequences. 'Other' type sequences

are relatively more surprising. The negative relationship seems to be an artifact of the low probabilities of 'cutoff' words in the language model, where they exhibit high surprisal regardless of error type.

Figure 4.6a shows the top twenty insertion errors, which almost entirely consist of function and discourse words (the exception is 'just', which could arguably be included in the function word category). All top insertions have a positive surprisal difference. The most common tokens 'and' and 'uh' account for 19% of all insertions.

In the PTB model, mixed type errors are errors which include predominantly insertions and substitutions in a sequence (although 2% of mixes are some combination of insertion and deletion). Mixed errors pattern similarly to insertions, where function and mixed lexical class errors appear less surprising. In the mixed category, discourse errors average close to a zero average surprisal difference, while PTB content errors are relatively surprising compared to the corrected transcription.

4.3.3 Modeling the surprisal difference of deletions

The error type model predictions for deletions and substitutions with MsState lexical information can be see in Figure 4.4b. In this set of models, the surprisal difference scale is interpreted differently. Values above zero indicate that the corrected MsState version is less surprising than the PTB version, while values below zero indicate that MsState is more surprising. Deletions on average appear less surprising than the corresponding PTB sequences, while mixed deletion errors are closer to zero.

Lexical class and its interactions are included in the full model and predictions can be seen in Figure 4.5b. Here, the function and mixed lexical class deletions are relatively low in surprisal, patterning similarly to insertions. Content, discourse, and other words have lower means which are close to zero. In the mixed error condition, which is composed of mostly deletion-substitution errors and a small set of insertions and deletions (1%), the lexical class patterns very similarly to deletions, although means are lower overall. Discourse and 'other' classes are definitively more surprising.

The top deletions are shown in Figure 4.6b. These deletions cluster into two groups, a group of function and discourse deletions which have a positive surprisal difference (indicating the deletions themselves are not very surprising) and a group of 'other' subwords which are negative. The discourse token 'uh' makes up 13% of all deletions. Although the top discourse deletions have a high surprisal difference, the model predicts discourse deletions at around zero. There is a long tail of less frequent discourse deletions that average a low surprisal difference, including deletions of 'ah', 'yeah', 'huh', and 'see'.

4.3.4 Modeling the surprisal difference of substitutions

The substitution errors are considered from the perspective of the substituted word in the PTB models, and the corrected word in the MsState models. In the PTB error type model, substituted sequences are marked as significantly less surprising than their MsState counterparts. The MsState model generally corroborates these findings, showing that the corrected part of the substitution is slightly more surprising (although the confidence interval crosses zero). This provides some support for the hypothesis that the error part of a substitute will be on average less surprising than the correction.

Taking lexical class interactions into consideration, almost all lexical types of the substituted (PTB) sequences are in the less surprising range, with the exception of 'other' sequences. When looking at the lexical types of the substituted sequences in the MS model, there is a wider range of variance, with discourse corrections being typically more surprising but all other types varying around zero.

In the case of mix and content lexical types, both the substituted PTB sequences and the MsState sequences are considered less surprising. At first glance, this is unexpected due to the fact that surprisal difference values for substitutions in the PTB and MsState models are simply converses. However, this can be explained in part by the mismatch in lexical class between a given PTB substitution and its MsState correction. Table 4.5 shows the distribution of lexical type for both the PTB and MsState sequences in a substitution. Content and mixed sequences are commonly substituted by non-matching lexical types (at

PTB	MsState				
	Content	Discourse	Function	Mix	Other
Content	1384	36	191	35	282
Discourse	79	4212	519	28	188
Function	230	482	3162	104	1668
Mix	14	17	85	276	7
Other	259	129	499	12	691

Table 4.5: The distribution of average lexical types for substitutions. Rows represent PTB lexical type and columns represent MsState lexical type.

28% and 30% of the time). 'Other' sequences are substituted by non-matching lexical types more often than not, at 57%. Discourse sequences, which show the expected mirroring of model predictions, substitute for non-matching sequences only 16% of the time.

The top twenty substitution errors are plotted in Figure 4.6c. The most common substitution pairs are function and discourse substitutions in which the substituted word is less surprising than the corrected word. These words are also notable for their phonetic similarity (e.g. 'i' and 'i've'). Content words are the third most common type of substituted word, but the substituted sequence pairs have a more uniform frequency distribution (the most common substitution of this type is 'use/used' with 23 instances). Table 4.6 shows a random sample of twenty content-content substitution pairs. The surprisal difference of content-content substitutions is not strongly positive or negative ($\overline{n}$=0.07).[2]

[2]The middling suprisal difference of content-content substitutions is likely influenced by a substantial number of typos (e.g. buffalos/buffaloes). Typos are often unseen in the LM and thus exceptionally surprising.

PTB seq.	MsState (corrected) seq.
opera	opry
favorites	favorite
burnt	burned
machines	machine
order	old
fellows	fellas
corner	corners
stuff	such
make strike	makes right
call	cost

Table 4.6: A random sample of content-content substitutions.

4.4 Conclusion

The results in this chapter provide an examination of the relationship between the predictability of words and transcriber errors in a corpus. Interestingly, it is not the case that transcribers are simply likely to transcribe word sequences that are more expected. When transcribers hallucinate a word or sequences of words, they are likely to produce a transcription that is more expected than the corrected version. Hallucinations may simply make a word sequence more expected or better align with a listener's mental grammar. However, when transcribers miss a word or word sequence, the reference sequence is on average lower in surprisal. There are several potential reasons why missed sequences have low surprisal. One reason is simply a matter of low surprisal corresponding to high frequency tokens. Frequent words appear more often, and if all word types have similar error rates, frequent tokens will be misheard more often. Zayats et al. (2019) found that the relationship between error rates and log

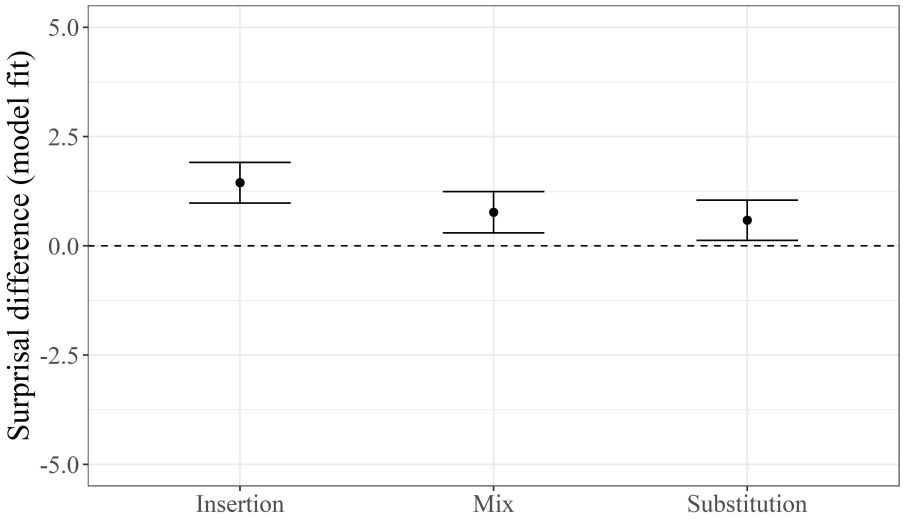
frequencies of function and content words was linear. Another possibility is that frequent words are shorter in duration and often reduced, and are therefore easier to miss.

Considering substitutions, there is some weak support for the idea that substituted words are generally less surprising than their corrections. The PTB model suggests that substituted words are less surprising, while the MsState model does not find corrections definitively more surprising. Further studies should consider substitutions in more detail.

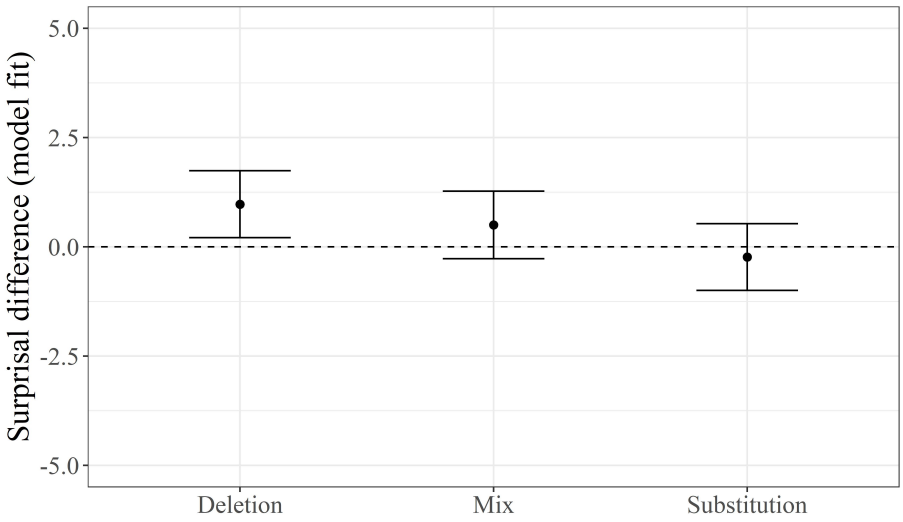
Furthermore, lexical interactions are very important when considering the relationship between predictability and errors. Function insertions and deletions show the highest surprisal difference, and this effect is strong even when controlling for the highest-frequency words. Zayats et al. (2019) show high-frequency tokens are more likely to lead to errors. Function word errors are by far the most numerous in the corpus, and they have a strong influence on predictability differences when considering error types.

The current study relies on forward-direction language modeling, where the expectation of a word is based on previous words in context. While these models align with the idea of a 'left-to-right' processing mechanism of speech, studies find that the following context of a word can aid in its disambiguation (Grosjean, 1985; Bard et al., 1988), and speech perception models such as TRACE (McClelland and Elman, 1986) account for these effects. Future work could take advantage of backwards or bi-directional models to consider how the predictability based on following context affects transcriber errors.

This study takes a novel approach to studying perceptual errors with the use of very large corpora of speech. Transcription differs from traditional laboratory studies in speech intelligibility, where the participant responds to isolated words or sentences by, for instance, producing a target word or choosing from a list of alternatives. Transcription is a complex process which requires perception and production of speech and 'translation' of natural language into written forms. Transcription errors may arise throughout the process due to intelligibility issues, memory limitations, or production errors (i.e. typos). The following chapter uses a more common psycholinguistic methodology, the forced-choice task, to examine the replicability of speech transcription errors.

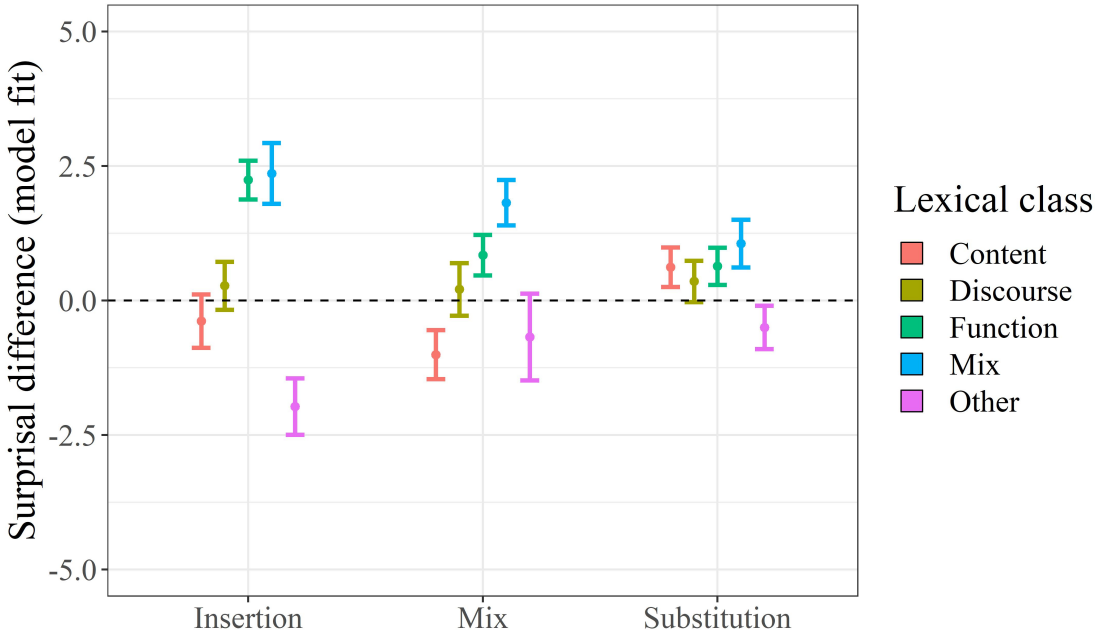


(a) PTB insertion/substitution model. A positive surprisal difference indicates that the PTB sequence is less surprising than the MsState sequence.

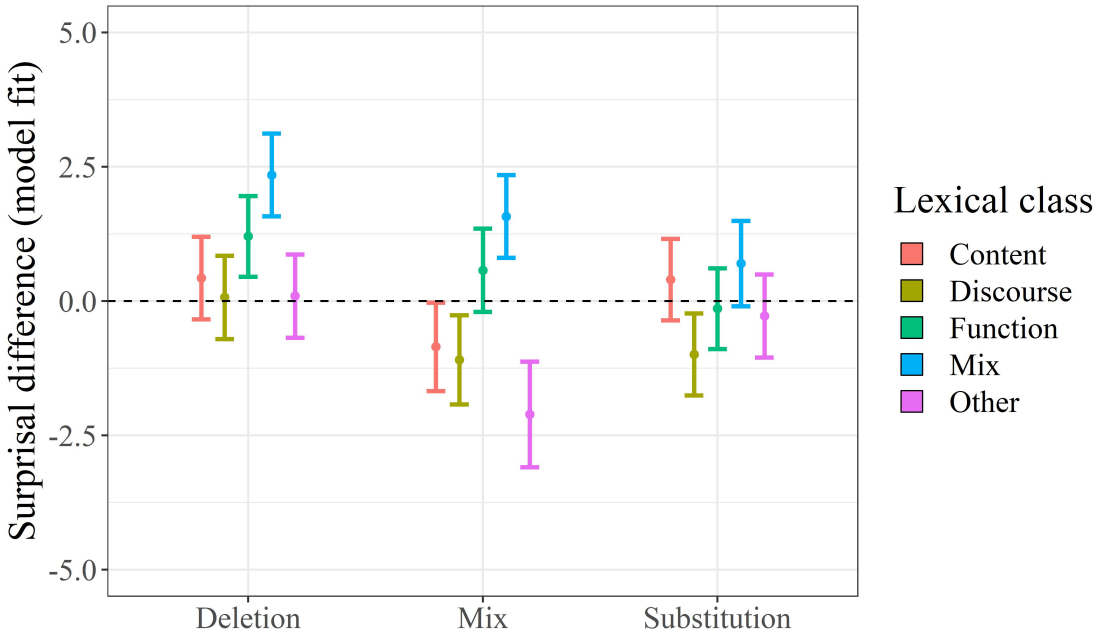


(b) MsState deletion/substitution model. A positive surprisal difference indicates that the MsState sequence is less surprising than the PTB sequence.

Figure 4.4: Model predictions using error type linear models. The mean and confidence intervals are shown.

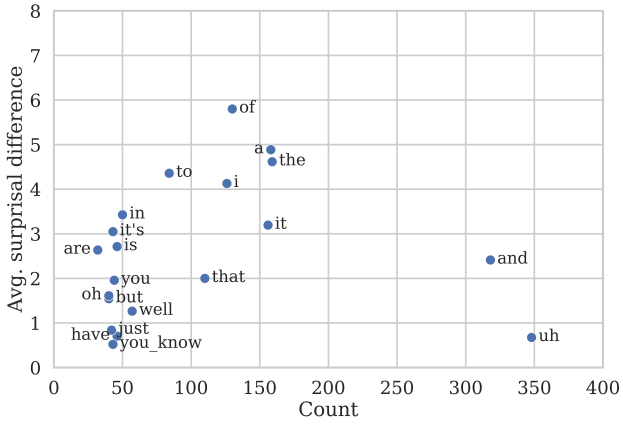


(a) PTB insertion/substitution model

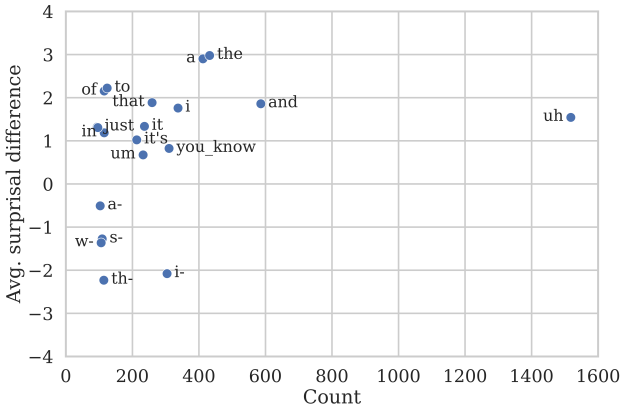


(b) MsState deletion/substitution model

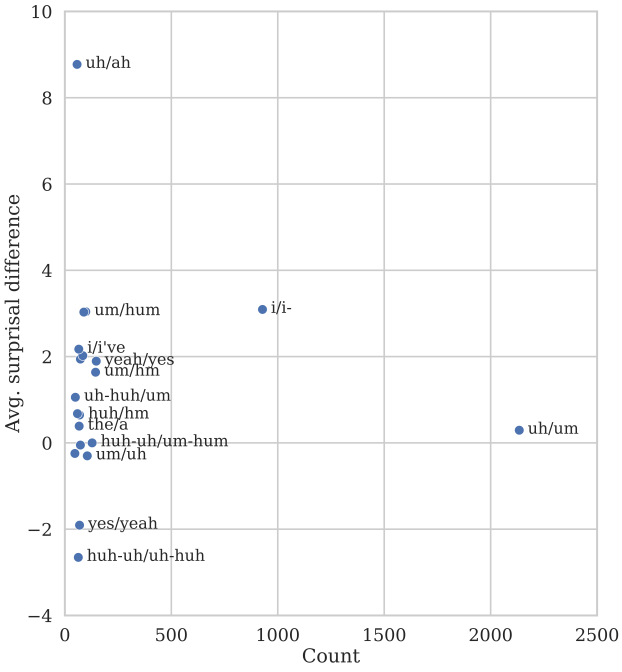
Figure 4.5: Model fit for linear models with lexical class interactions.



(a) Insertion errors



(b) Deletion errors



(c) Substitution errors

Figure 4.6: Top twenty errors by count and average surprisal difference.

Chapter 5

CROWD-SOURCING MISPERCEPTIONS FROM CONVERSATIONAL SPEECH

5.1 Introduction

This chapter considers the replicability of speech transcription errors using crowd-sourcing. It further examines the relationship between the accuracy of transcriptions and linguistic features such as predictability, lexical type, and sentence position.

One of the major challenges in studying speech perception is the collection of data, and one common way to study perception is by examining speech perception errors (misperceptions). Table 5.1 presents misperception data from various sources. There is a continuum regarding how misperception data is collected, with controlled laboratory stimuli on one end, and 'naturalistic' data on the other. Curated misperceptions such as that of Bond and Garnes (1980) are collections of misperceptions where the authors and collaborators gather examples from everyday speech. In these collections, the errors have been perceived in 'naturalistic' environments, in the context of a conversation. One problem with this method of collection is a potential bias towards semantically important words, due to the fact that each example has to be noticed by the listener. Low frequency words and content words are likely to be overrepresented in such sets. On the other hand, laboratory tasks which examine perception (Pollack et al., 1960; Felty et al., 2013) are more carefully controlled and consider a wider sample of words and word frequencies. A major problem with laboratory studies is that words are often presented without context. In such conditions, the greater task difficulty involved in recognizing words without context is shown to lead to more confusability than a task which provides context (Clopper et al., 2006). In addition, the role of context in recognition cannot be effectively examined in such studies.

Dataset	Size	Source
Bond and Garnes (1980)	1K	Curated misperceptions in everyday conversation
Pollack et al. (1960)	5K	From laboratory study of monosyllabic content words in noise
Felty et al. (2013)	30K	From laboratory study of randomly sampled words in noise
Switchboard alignments (Taylor et al., 2003; Deshmukh et al., 1998; Hamaker et al., 1998)	32K	Alignment of corrected conversational telephone transcriptions

Table 5.1: Datasets of speech misperceptions including the number of misperceptions and a brief description of the origin of the speech errors.

Speech transcriptions are a widely available source of data which could be used for the study of speech perception. By comparing two transcriptions, one can collect a database of speech errors. Transcription error data shows errors in context, and may be less biased toward a particular type of semantically meaningful error than found sets. However, using mistranscriptions to study perception is novel. Potential task effects should be carefully considered. The process of transcribing is both an act of perception and production. Transcribers may listen to both conversation partners at once, and they may work on a task over several hours. Because of these factors, it might be assumed that transcription poses a higher task difficulty compared to traditional perception tasks. To consider this effect, the current study has participants judge sentences from a transcription error set using a forced-choice task, which is already used extensively in speech perception literature.

The current study targets transcriptions where transcribers have made various types of errors: hallucinations, misses, and substitutions. Participants listen to previously mistran-

scribed audio segments. Instead of re-transcribing the segments, a forced-choice task prompts the user to choose between two alternatives which appear on the screen. One sentence is mistranscribed and contains an error, while the other sentence has been corrected.

A control set of stimuli is introduced alongside the mistranscribed sentences. The control stimuli have no mistranscription errors – they were transcribed correctly in the original transcription. However, the control audio are presented with fabricated lexical errors. The purpose of the control stimuli is to establish a baseline rate of accuracy for the participants. It is expected that, given the simplicity of the task, performance on the control stimuli should be quite good. However, participants are predicted to perform poorly on mistranscribed sentences. In other words, they will be likely to reproduce the initial mistranscription. Hypotheses about reproducibility include the following:

- **Null hypothesis:** Accuracy on the mistranscription stimuli will not be lower than the control stimuli.
- **Alternative hypothesis:** Accuracy on the mistranscription stimuli will be significantly lower than the control stimuli.

The control stimuli and mistranscription stimuli are also balanced for certain lexical conditions. There are equal numbers of stimuli for the three error types (hallucination, miss, substitution), and three broad lexical types (function, content, interjection). The length of the sentence is also balanced between the control and mistranscription stimuli. These balances allows for a statistical analysis of how various lexical features affect the listener's accuracy. Of course, it is well-known that acoustic features play a role in misperception, but this study targets the role of lexical features.

One linguistic feature of interest is the predictability of a word given its context. The previous chapter found that transcribers were more likely to hallucinate or substitute a word that is predictable. One might expect that listeners are more likely to choose an incorrect alternative if that alternative is more predictable. This aligns with models of speech perception

where top-down information can influence recognition at various stages. To quantify the gap in predictability, the difference in linguistic surprisal is measured (refer to Chapter 4 for implementation details). The hypotheses regarding predictability are as follows:

- **Null hypothesis:** Participants' choice of alternative will not be correlated with that alternative's predictability.
- **Alternative hypothesis:** Participants' choice of alternative will be positively correlated with that alternative's predictability.

5.2 Methodology

This chapter uses a forced-choice methodology where participants hear a short audio and select between two sentence alternatives. Half of the sentences are taken from a set of mistranscribed sentences, while the other half are considered a control group. The stimuli creation process and statistical methods will be discussed in detail.

5.2.1 Dataset

The stimuli consist of 220 sentences of conversational telephone speech from the Switchboard project (Godfrey et al., 1992). The PTB version of transcripts (Taylor et al., 2003) is an initial set of transcriptions, and the MsState version (Deshmukh et al., 1998; Hamaker et al., 1998) includes transcriptions from the same set of audio which have been corrected to reduce the word error rate by an estimated 8%. See Chapter 3 for more information about the transcriptions used in this study.

5.2.2 Stimuli selection

The stimuli consist of pairs of mistranscriptions and control sentences from PTB and MsState transcripts. Although the 'slash units' in Switchboard may consist of short or incomplete utterances (e.g. backchannels), the stimuli contain sentences of three or more tokens.

An algorithm shown in Figure 19 is used to select pairs of similar control and mistranscription sentences. Each mistranscription m_{sen} is collected from sentences of at least three tokens, where the error is one token long.[1] The context m_c is also extracted, a substring which includes the error token and the previous and following tokens. In the case of a hallucination or substitution, the error token is the inserted word, while in the case of a miss, the error token is null. First, candidate control sentences are collected from the set of all non-mistranscribed sentences S based on matching context. Then, the best matching control sentence is selected for each mistranscription. The best matching sentences are selected based on a similarity score. This score aims to minimize differences between the lengths of the sentences and the positions of the matching context. The similarity score can be expressed as:

$$similarity = e^{-|\log(\frac{len(s_m)}{len(s_t)})|} + e^{-|\log(\frac{pos(e_m)}{pos(e_t)})|}$$

where $len(s)$ is the sentence length of the mistranscription (m) or target pair (t), and $pos(e)$ is the relative position of the error (i.e. the index of the error divided by sentence length). The sentence that maximizes similarity is added to the set of potential stimuli pairs.

The set of potential stimuli pairs is then randomized and sorted into lists by error type and lexical type. Twelve pairs of each combination of error type and lexical type, totalling 108 pairs, are manually chosen based on several criteria. Both the mistranscription and control sentence had to meet all criteria:

1. The audio is correctly time-aligned with the transcript

2. There is no audio interference from the other speaker, or loud obscuring noises (e.g. baby crying, sirens)

3. The audio was not judged to be extremely lengthy (i.e. greater than 10 seconds in length)

[1] To reduce complexity, only single-word errors are considered. Approximately 87% of errors in PTB are singletons.

Algorithm 1 Stimuli selection

Input: Set M of tuples (m_{sen}, m_c), Set of strings S

Output: Set P of tuples (a, b)

$P \leftarrow \emptyset$

$X \leftarrow$unique m_c in M

// Find candidate control sentences.

$B \leftarrow \text{map}(K; V)$

for $s \in S$ **do**

    for i in tokenized s **do**

        Compute *context* for $s[i]$

        if *context* $\in X$ **then**

            $B \leftarrow (context; s)$

// Choose top control candidate for each mistranscription.

for $(m_{sen}, m_c) \in M$ **do**

    $candidate = \emptyset$

    $sim = \infty$

    for $(k; v) \in B$ where $k \Leftrightarrow m_c$ **do**

        Compute similarity score *current* for (m_{sen}, v)

        if $current < sim$ **then**

            $candidate = v$

    $P \leftarrow (m_{sen}, candidate)$

Return P

Error type	Mistranscripton	Control
Hallucination	i **think** **[it/–]** **was** november	i **think** **[–/it]** **was** campinas
Miss	**oh** **[–/well]** **that** is nice	**oh** **[well/–]** **that** would uh-huh
Substitution	**[unfortunately/fortunately]** **i** don't have to work in those companies	**[fortunately/unfortunately]** **i** haven't uh haven't uh been inundated with that situation yet

Table 5.2: Examples of mistranscription and control sentence pairs. The shared context is emphasized in red. The error is in brackets. The incorrect reading of the sentence is shown before the slash. The correct reading of the sentence is shown after the slash.

4. Mistranscription stimuli were chosen for a range of surprisal differences across lexical classes, to approximate a normal distribution.[2]

Examples of the sentence pairs are shown in Table 5.2.

5.2.3 Crowd-sourcing procedure

Participants from Amazon's Mechanical Turk crowd-sourcing platform were asked to listen to audio and select between two alternatives, one which matched the correct audio transcription and one which contained an error. The stimuli were divided into 9 conditions, and each condition was completed by 10 participants. Each trial contained approximately 14 mistranscription sentences and 14 control sentences. Audio and sentences were randomized for each participant. The participants were paid $1.25 for completing the task, which was estimated to take 4 minutes.

[2]The previous chapter shows that surprisal difference and transcriber error has a systematic relationship.

Instructions

After agreeing to the study according to IRB protocols, the participants were instructed to put on headphones and play a test audio sample. They were then given a set of task instructions and could proceed to 2 practice trials. The instructions are shown in Figure 5.1.

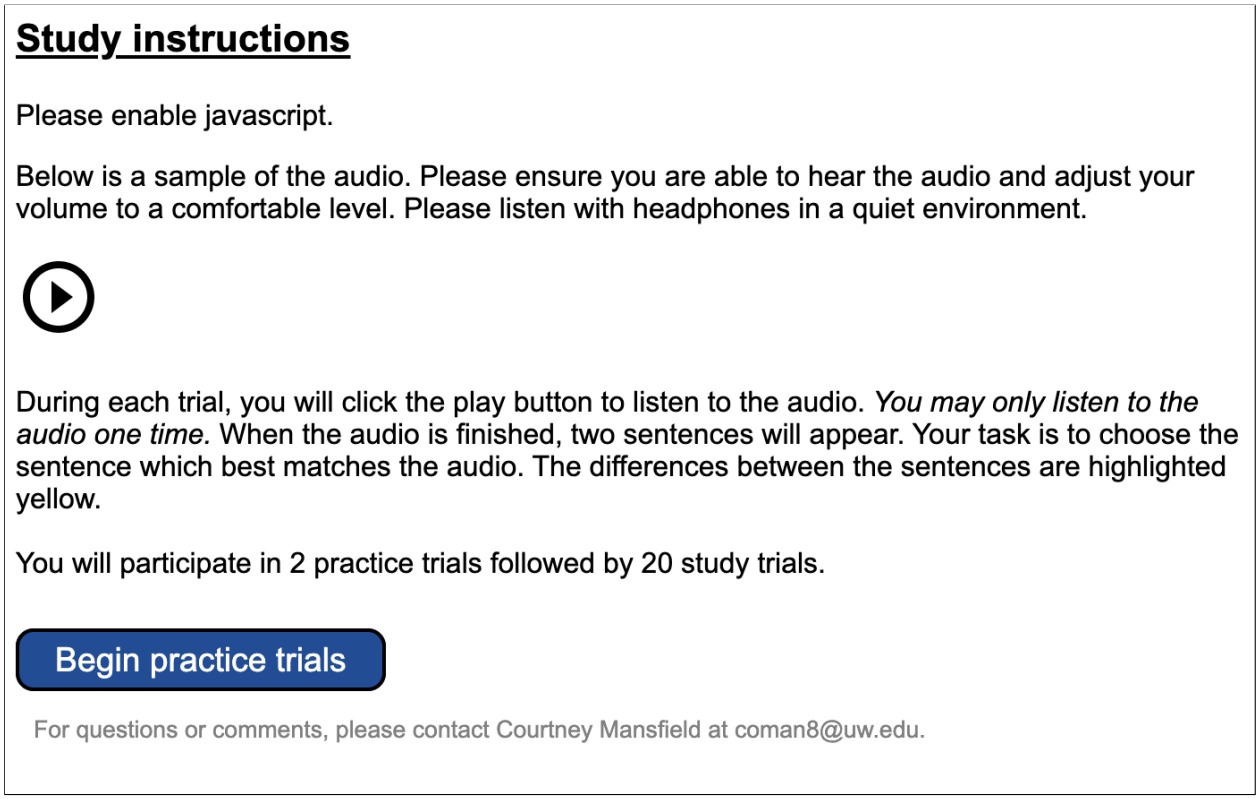


Figure 5.1: Participant instructions.

Graphical interface

Figure 5.2 shows the user interface. During each trial, the participant clicks to play the audio. They may only listen to the audio once before making their judgement. When the audio is finished playing, two sentence alternatives appear below corresponding to an incorrect transcription and a correct transcription. The difference between the sentences is highlighted

in yellow. The listener must choose the best fitting transcription before proceeding to the next trial.

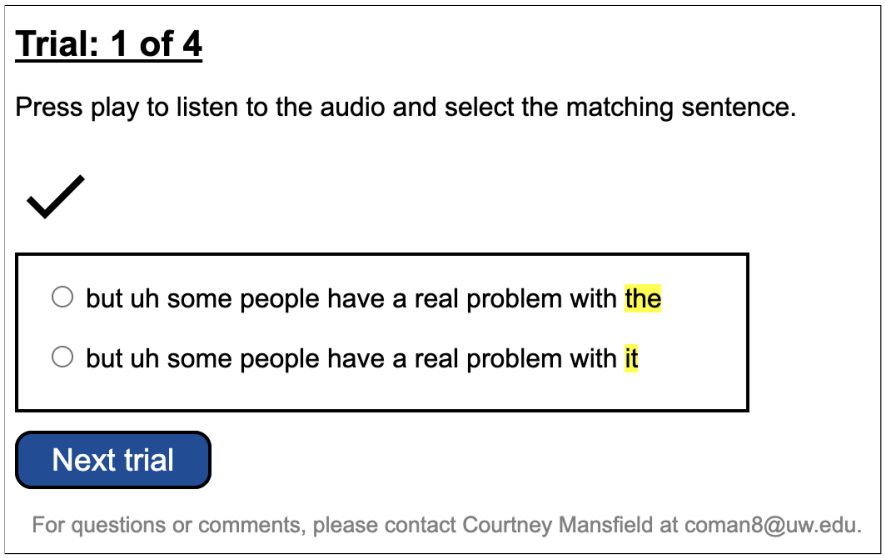


Figure 5.2: Experiment user interface.

5.2.4 Participant demographics

A total of 77 participants took the survey. One participant did not correctly fill out the demographic information, and their results were excluded from the analysis. Several people participated in multiple conditions–8 participated in 2 conditions, 1 participated in 3 conditions, and 1 participated in 5 conditions. Participants were required to be native speakers of English and to currently live in the US. Demographic information was collected including age, languages spoken, and the state where the participant grew up. The average age of participants was 37, with a minimum age of 18 and maximum age of 65. Figure 5.3 shows languages (other than English) spoken by the participants. Participants grew up in 24 states across the US, and the counts were generally representative of the populations of those states. The top states were Florida and Illinois (N=9), followed by California and Texas (N=6).

Language	Count
Spanish	5
German	2
Chinese	1
Danish	1
French	1

Table 5.3: Languages (other than English) spoken by participants.

5.3 Statistical model

To assess the differences between the mistranscription and baseline stimuli, an ordinal logistic regression (OLR) model is used to model the relationship between the factors of interest and the response accuracy. In OLR, Y is an ordinal outcome with J number of categories. $P(Y \leq j)$ is the cumulative probability of Y less than or equal to category $j = 1, ..., J-1$. Ordinal logistic regression is then formulated as:

$$logit(p(Y \leq j)) = \theta_{j0} - b_1 x_1 - ... - b_n x_n \tag{5.1}$$

where θ_{j0} is the coefficient of the j^{th} outcome category and $b_{(1...n)}$ are the coefficients of the predictor variables.

5.3.1 Model variables

The outcome variable is stimulus response accuracy, which is defined as the total number of correct sentence alternatives selected by participants per unique audio stimulus. Note that accuracy here denotes whether the alternative a participant chose corresponds to the ground truth. Because each audio is judged by 10 participants, the response accuracy can be calculated as an ordinal of integers 0-10. The distribution of stimulus response accuracy

scores is shown in Figure 5.3. The number of audio belonging to each response accuracy score varies considerably. Only 3 stimuli had a response accuracy score of 0, while 46 stimuli showed a score of 9. For this reason, the response accuracy was bucketed into more balanced categories, with N=3 buckets chosen as a well-balanced and interpretable number. The three categories are named as 'low accuracy' (correct response of 0-5), 'medium accuracy' (correct response of 6-8), and 'high accuracy' (correct response of 9-10)

Predictor variables include the origin of the sentence (mistranscription or control stimulus) and various linguistic factors. Predictor variables are listed and described in Table 5.4.

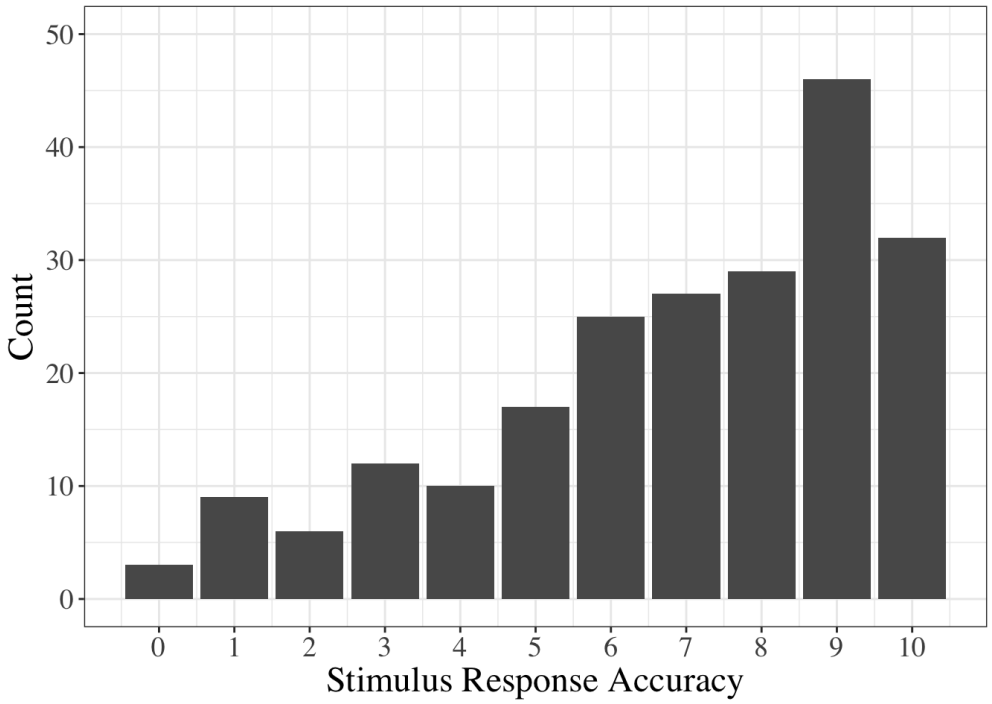


Figure 5.3: Count of stimuli corresponding to each response accuracy score. A score of 10 (shown on the x-axis) indicates that all 10 participants chose the correct sentence alternative for a particular audio stimulus.

Predictor name	Type	Description
Control	Categorical	Describes the origin on the stimulus; does it come from a mistranscribed audio (0) or a control audio (1).
SurprisalDiff	Continuous	The difference in surprisal between the error alternative and the correct alternative.
LexType		The error token's lexical category.
LexType:Content	Categorical	The dummy function lexical type (0) vs. content (1).
LexType:Interj	Categorical	The dummy function lexical type (0) vs. interjection (1).
ErrorType		The type of transcription error.
ErrorType:Del	Categorical	The dummy hallucination error type (0) vs. miss (1).
ErrorType:Sub	Categorical	The dummy hallucination error type (0) vs. substitution (1).
ErrorPos		The sentence position of the error.
ErrorPos:BOS	Categorical	The dummy middle of sentence position (0) vs. beginning of sentence (1).
ErrorPos:EOS	Categorical	The dummy middle of sentence position (0) vs. end of sentence (1).
SentLen	Continuous	The total number of tokens in the sentence.

Table 5.4: Predictor variables and description

5.3.2 Model selection

To select for the most meaningful model predictors, backwards stepwise selection based on Akaike Information Criterion (AIC) is used. AIC is formulated based on the maximum likelihood estimate of the data and the number of predictor variables used, and is known to penalize complexity less severely than comparable measures. AIC is formulated as:

$$AIC = N \cdot \log(\frac{RSS}{N}) + 2k \tag{5.2}$$

where N is the number of examples in the training data, RSS is the residual sum-of-squares, and k is the number of model parameters. The OLR models and selection are implemented with the MASS package in R (Venables and Ripley, 2002).

5.3.3 Advantages and disadvantages

OLR has several advantages over other models. It does not assume a normal distribution of classes and provides interpretable coefficients which are calculated as log odds. There are also limitations to the approach. Logistic regression assumes a linear separability between the predictor variables and outcomes. In OLR, the odds ratio for each predictor variable is shared across outcome categories. Therefore, models must meet the assumption of proportional odds, where effects of predictor variables are proportional across the outcome categories. The Brant Test (Brant, 1990) uses goodness-of-fit measures to compare the ordinal approach to a set of separately fitted logistic regressions and will be used to assess the proportional odds assumption.

5.3.4 Demographic analysis

In addition, the study examines correlations between the performance of individual participants and their demographics. There are several reasons to assess participant scores: to determine if any outliers may affect the analysis, to consider how demographic factors might have a strong effect or skew on the findings, and to lend insights into how such factors affect

performance. Demographic features include age and languages spoken (i.e. is the participant mono- or mutli-lingual). The variable of interest in the demographic analysis is the participant accuracy. The participant accuracy is the proportion of correctly chosen sentence alternatives over the total number of stimuli that an individual has rated.

To assess the relationship between age and participant accuracy, the the Kendall rank correlation coefficient is used. Kendall rank correlation is a non-parametric rank test which has an advantage in its ability to handle ties in the data. Given a set of observations $(x_1, y_1), ..., (x_n, y_n)$, all pairs $(x_i, y_i), (x_j, y_j)$ are sampled where $i < j$. The pairs are assessed as being ordered the same way (concordant) or ordered differently (discordant). The equation can be written as:

$$\tau = \frac{2}{n(n-1)} \sum_{i<j} \operatorname{sgn}(x_i - x_j) \operatorname{sgn}(y_i - y_j) \tag{5.3}$$

where τ has a range of (-1,1). A z distribution is used as an approximation of the standard distribution, which can be used for significance testing.

5.4 Results

The following section presents the results of the statistical analysis. First, the OLR model results are shown. The baseline model is presented, which includes all predictors from Figure 5.4. A reduced model is selected using AIC scores, and the reduced model coefficients and performance are discussed. Next, the findings related to mistranscription and predictability are analyzed in more detail. Finally, an analysis of demographics and participant accuracy is presented.

5.4.1 Baseline model

The baseline model was trained with the seven predictor variables from Table 5.4. The results of the baseline model are presented in Figure 5.5. The coefficient values are presented as odds ratios. The predictor odds ratio represents the odds of a particular stimulus falling into

	Odds ratio	Std. Error	t value	p value
Predictors:				
Control	13.53	0.33	7.86	< 0.0001
SurprisalDiff	0.76	0.077	-3.62	<0.001
LexType:Content	3.03	0.37	2.97	<0.01
LexType:Interj	1.33	0.36	0.80	NS
ErrorType:Del	0.92	0.36	-0.22	NS
ErrorType:Sub	1.33	0.51	0.56	NS
ErrorPos:BOS	0.58	0.60	-0.91	NS
ErrorPos:EOS	0.82	0.63	-0.32	NS
SentLen	0.98	0.03	-0.80	NS
Intercepts:				
Low—Medium	2.29	0.49	1.67	NS
Medium—High	13.06	0.53	4.89	< 0.0001

Table 5.5: Information about the model coefficients for the full model.

the 'high' accuracy category rather than the 'low' or 'medium' categories, given a particular feature. Among the predictors, the surprisal difference, stimulus origin (mistranscription vs. control) and lexical type (function vs. content, but not function vs. interjection) are deemed significant.

5.4.2 Model selection

In order to limit the parameters of the model and exclude less meaningful predictors, a backwards-stepwise regression is used. Predictive factors which do not improve model performance are removed (as demonstrated by a decrease in AIC value). The AIC values at each step are shown in Table 5.6.

Predictors	AIC
Control + SurprisalDiff + LexType + ErrorPos + SentLen + ErrorType	370.09
Control + SurprisalDiff + LexType + ErrorPos + SentLen	366.59
Control + SurprisalDiff + LexType + ErrorPos	363.02
Control + SurprisalDiff + LexType	362.12

Table 5.6: Stepwise regression of each model based on the AIC scores. A lower AIC score indicates a better fitting model.

Model	DF	AIC	Deviance	LR stat	p-value
FULL	205	370.09	348.09	–	–
REDUCED	210	362.12	350.12	2.02	0.84

Table 5.7: Comparison of the reduced and full models, including degrees of freedom, AIC, deviance, and the ANOVA likelihood ratio stat and p value.

The baseline model is compared to the reduced model in Table 5.7. The AIC value of the reduced model is smaller, although the model deviance is slightly larger. An Analysis of Variance (ANOVA) shows that the model fit is not significantly different, which suggests that the simplified model maintains a relatively good fit with a reduced number of factors.

5.4.3 Main effects

The main effects of the reduced model are presented in Table 5.8. When the audio originates from the control group instead of the mistranscription group, it is approximately 13.25 times more likely to be classified as high accuracy than medium/low. A content word is 2.9 times more likely than a function word to be classified as high accuracy than medium/low.

	Odds ratio	Std. Error	t value	p value
Predictors:				
Control	13.25	0.33	7.87	<0.0001
SurprisalDiff	0.75	0.08	-3.77	<0.001
LexType:Content	2.90	0.37	2.90	<0.01
LexType:Interj	1.37	0.35	0.89	NS
Intercepts:				
Low—Medium	2.90	0.32	3.37	<0.001
Medium—High	16.39	0.37	7.53	<0.0001

Table 5.8: Information about the model coefficients for the reduced model.

Predictability also has a significant effect in the model. For every 1 unit increase in surprisal of the correct alternative relative to the incorrect alternative, the audio is 1.3 times more likely to be predicted as low rather than medium/high accuracy. Visualizations of the model coefficients can be seen in Figures 5.4 and 5.5.

5.4.4 Model assumptions

To assess the validity of the model, the model must pass the proportional odds assumption. The Brant test provides a calculation of the odds assumption. If the probability of the χ^2 value is greater than 0.05, then the assumption of proportional odds is sustained. The Brant test values are shown in Table 5.9. Both the global test and individual tests of coefficients confirm the proportional odds assumption.

Diagnostic plots in Figure 5.6 allow for an assessment of the model. The 'sure' package in R produces diagnostics using 'surrogate residuals' (Liu and Zhang, 2018). Given an ordinal outcome Y, surrogate residuals are continuous variables based on the conditional distribution of latent variable Y. The quantile-quantile plot in the top left corner shows the model residuals

Test for	χ^2	DF	Probability
All	3.60	4	0.46
Control	0.68	1	0.41
SurprisalDiff	0.04	1	0.85
LexType			
Content	2.30	1	0.13
Interjection	2.43	1	0.12

Table 5.9: Brant test for the proportional odds assumption, with chi square value, degrees of freedom, and probability

plotted against surrogate residuals, where a valid model will show points grouped closely along the diagonal line. The surrogate residual vs. fitted values plot shows residuals grouped approximately symmetrically about zero. Additional plots show the surrogate residuals plotted for each predictor. For the categorical variables, means are close to zero and are roughly symmetrical. For the continuous variable of SurprisalDiff, the residuals are generally grouped symmetrically. Overall, the diagnostics do not point to any serious issues.

5.4.5 Predictability

The predictability of the error, quantified by the surprisal difference between correct and incorrect alternative, was a significant predictor in the model. However, the average surprisal difference varied between the mistranscription and control stimuli. As part of the stimuli selection process, the surprisal of the mistranscriptions was balanced to have a roughly normal distribution, although the mean centers slightly above 0 ($\bar{x}$=0.48). The surprisal difference of the control stimuli were not constrained, and show an average surprisal difference of -1.07. A plot of the density distribution in both categories of stimuli is shown in Figure 5.7.

To ensure that the findings about predictability are reliable, additional modeling is

	Odds ratio	Std. Error	t value	p value
Mistranscription model:				
SurprisalDiff	0.82	0.10	-1.99	<0.05
LexType:Content	6.13	0.63	2.88	<0.01
LexType:Interj	4.86	0.63	2.5	<0.05
Low—Medium Intercept	6.52	0.54	3.49	<0.001
Medium—High Intercept	32.27	0.61	5.73	<0.0001
Control model:				
SurprisalDiff	0.72	0.12	-2.73	<0.01
LexType:Content	2.23	0.54	1.48	NS
LexType:Interj	0.63	0.48	-0.98	NS
Low—Medium Intercept	0.13	0.44	-4.56	<0.0001
Medium—High Intercept	0.91	0.36	-0.26	NS

Table 5.10: Model coefficients for the mistranscription and control subgroups.

performed. Two OLR models are fit, one each for the misperception and control stimuli. The model predictors are identical to the reduced model above, with the 'Control' predictor removed. Table 5.10 shows the model coefficients. Increases in the surprisal difference are associated with lower accuracies in both models, although the effect is stronger in the control stimuli model.

5.4.6 Participant accuracy and demographics

Figure 5.8 shows the participant's performance as a function of overall accuracy. Each datapoint represents the average accuracy for one listener considering all stimuli which they have judged, including both mistranscription and controls. Participant accuracy has a mean

of 0.63, with a standard deviation of 0.13. The minimum participant accuracy was 0.35 and the maximum 0.95.

The demographic data includes second languages spoken, but only 8 of 76 participants (10.5%) spoke a language other than English. The average performance of monolingual speakers and multilingual speakers was similar ($\bar{x}$=0.63 vs. $\bar{x}$=0.59). According to Kendall's rank correlation, age and participant accuracy were not significantly correlated (τ=0.10, z=1.26, p=0.20).

5.5 Discussion

The following sections discuss the findings of the analysis in more detail. First, the relationship between stimulus response accuracy and various lexical factors is discussed. Finally, there is a discussion of the participant accuracy and participant demographics.

5.5.1 Mistranscription vs. control sentences

While five features of the stimuli were considered in the OLR model, the most impactful feature was the origin of the stimulus, or whether it was selected from the mistranscription stimuli or the control stimuli. It was hypothesized that previously mistranscribed audio would be misperceived more often, and this was found to be the case. Sentences that came from the control group were about 16 times more likely to be bucketed into the high accuracy category (where the high accuracy group signified a response accuracy of 9 or 10/10). When participants heard the control audio, they chose the correct alternative 85% of the time. When participants listened to previously mistranscribed audio, they chose the correct alternative only 53% of the time.

One consideration is whether participants tended to agree on the mistranscription alternatives, regardless of correctness. While participants were choosing the correct alternative just over half of the time in these cases, it is possible that as a group, participants typically had the *same* correct and incorrect readings. If participants show high levels of agreement, it is possible that the transcription corrections aren't always reliable, or that expert listeners

and MTurk workers have systematic differences in their judgements.

Further analysis does not support this view. Figure 5.9 shows a measure of average listener agreement for mistranscription and control stimuli. An agreement score of 1 signifies that all listeners chose the *same* alternative for a stimulus (regardless of whether that alternative was correct). An agreement score of 0 signifies that there was perfect *disagreement*, where half of the listeners chose the correct alternative, and half chose the wrong alternative. Agreement on mistranscription stimuli was low, with an average agreement of 0.4. In other words, the average mistranscription had at least 3 of 10 participants disagree with the majority choice. This suggests the audio itself is ambiguous or challenging.

The difference in response accuracy between mistranscription and control was likely related to various acoustic-prosodic features of the utterances. While several lexical features were controlled for, acoustic variability was left unconstrained. In previous studies, factors such as acoustic confusibility and reduction have been shown to affect intelligibility (Luce and Pisoni, 1998; Vitevitch, 2002).

5.5.2 Predictability and accuracy

The second hypothesis suggested a relationship between the relative predictability of an error in a stimulus and its response accuracy. In particular, participants were expected to select the alternative that is more predictable, according to the difference in linguistic surprisal. This was also found to be true. For every 1 unit increase in the surprisal difference (which signified an increase in the relative predictability of the incorrect sentence), a stimulus was 1.3 times more likely to be placed in the low accuracy response category. Because of differences in the sampling of the mistranscription and control data, separate models for each subgroup were implemented. The models confirmed the effect and its significance.

This effect is shown clearly when examining response accuracy outliers from the mistranscriptions. Below are examples of mistranscription stimuli which had a perfect response accuracy of 10/10:

(2) we can't uh seem to pay for all the **little** things we have going now *hallucination*

(3) i it wasn't that long **ago** that i was that young *miss*

(4) oh **oh** that's a neat idea *hallucination*

(5) and you have to you know be familiar with **the/it** *sub*

These examples are quite different, although two are centered on content tokens ('ago', 'little'). However, in these examples the correct alternatives are much more predictable than the incorrect transcription. They carry an average surprisal difference of -2.53.

5.5.3 Other lexical features and accuracy

Several other lexical characteristics of the error and sentence were considered. These include lexical type, type of error, the position of the error in the sentence, and the length of the sentence. Whether the error was a hallucination, miss, or substitution had no significant impact on response accuracy. The position of the error in the sentence and the length of the sentence did not play a significant role.

The lexical class of the error was found to affect the response accuracy. Function word errors resulted in the lowest performance, and such stimuli were about 3 times less likely than content error stimuli to be predicted in the high response accuracy bucket by the model. This finding is in line with previous studies which show that high-frequency words heard in context are more likely to be misperceived (Vitevitch, 2002). When stimuli include interjection errors, defined as filled pauses, back-channels, and other discourse marking words, they had a lower response accuracy than content words, but higher than that of function words.

While the control stimuli had an average accuracy of 8/10, there were several cases which had a remarkably low response accuracy of 5/10 or lower. These cases include:

(6) and one of the things we were told is **like** they had eleven thousand harm missiles which t. i. is the sole supplier for *miss*

(7) for me i **i** just i just coul- wouldn't want to do that *hallucination*

(8) oh **oh** that's what i was thinking by quick transition *miss*

These error tokens are categorized as both function and interjection words. Notably, the errors are all part of disfluencies, and two involve repetitions. In sentence 7, the second 'i' was not present in the audio, while in sentence 8, the second 'oh' was. These sentences, with disfluencies and especially repetitions, appear considerably challenging. Further work on structural disfluencies and misperception should clarify this relationship.

5.5.4 Participant accuracy and demographics

Finally, demographic factors such as age and languages spoken were considered. Previous work on speech in noise show that age is a correlate of recognition performance (Dubno et al., 1984). However, age and accuracy did not show any relationship. This may be due to the relatively young age of participants in the study (75% of participants are less than 41 years old).

The average accuracy did not differ considerably between monolingual or self-professed multi-lingual speakers. In this case, there were few participants who claimed to speak a second language, and participants were not asked to rate their proficiency in the other language. A more targeted study of bilingual speakers is needed to address the relationship between L2 and perception accuracy in such a task.

5.6 Conclusion

To conclude, this chapter uses a crowd-sourcing, forced-choice experimental paradigm to examine whether participants are likely to misperceive sentences from a CTS corpus which have been previously mistranscribed. When participants hear previously misperceived sentences, they perform close to chance at 53%. Participants judged control sentences with 85% accuracy.

Using ordinal logistic regression modeling and correlation testing, the origin of the sentence (whether it was previously mistranscribed) was a strong indicator of response accuracy, along with lexical features such as predictability and lexical type. More predictable alternatives

were likely to be chosen, and participants were less likely to make errors when the target word was a content word. Other factors, such as length of sentence or age of the participant, were not shown to affect accuracy.

The next chapter will consider both human mistranscription errors and mistranscriptions from automatic speech recognition systems. Expanding on the findings of the previous chapters, it will examine whether ASR errors are associated with similar lexical features as human perceptual errors.

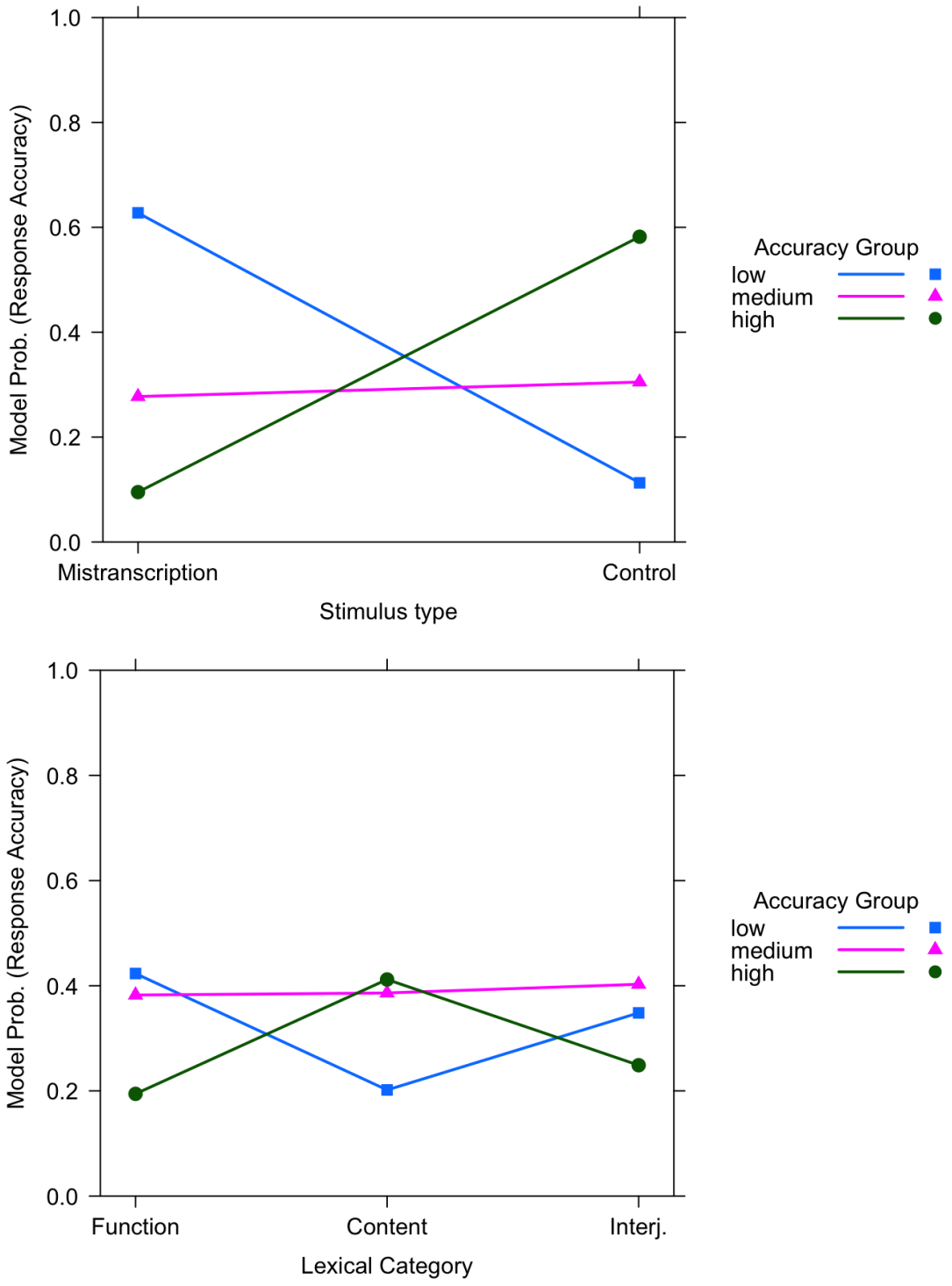


Figure 5.4: Plots of the model effects for categorical predictors. The y-axes indicate the probability of model classification of an accuracy group based on a) stimulus type or b) lexical category.

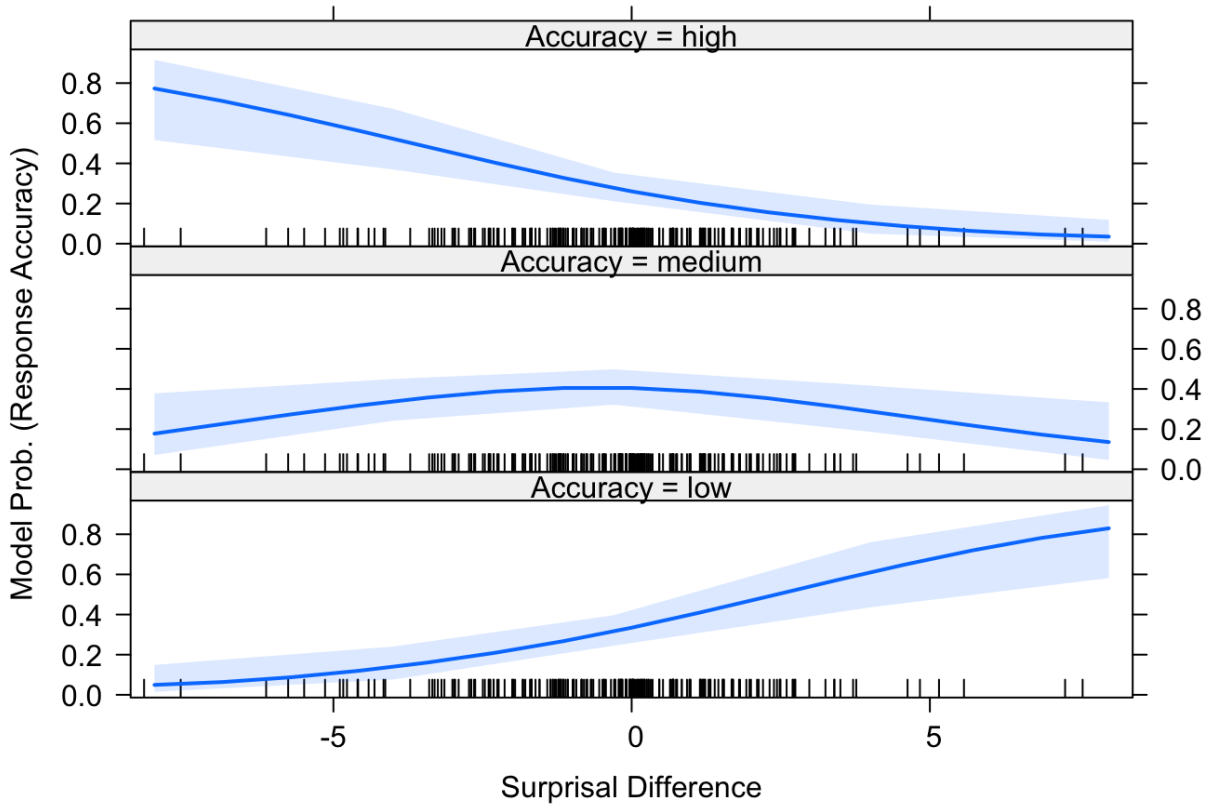


Figure 5.5: Plot of the model effects for the surprisal difference. The y-axis indicates the probability of model classification in an accuracy group based on the surprisal difference. A positive surprisal difference indicates that the correct alternative is more surprising (less predictable) than the incorrect alternative.

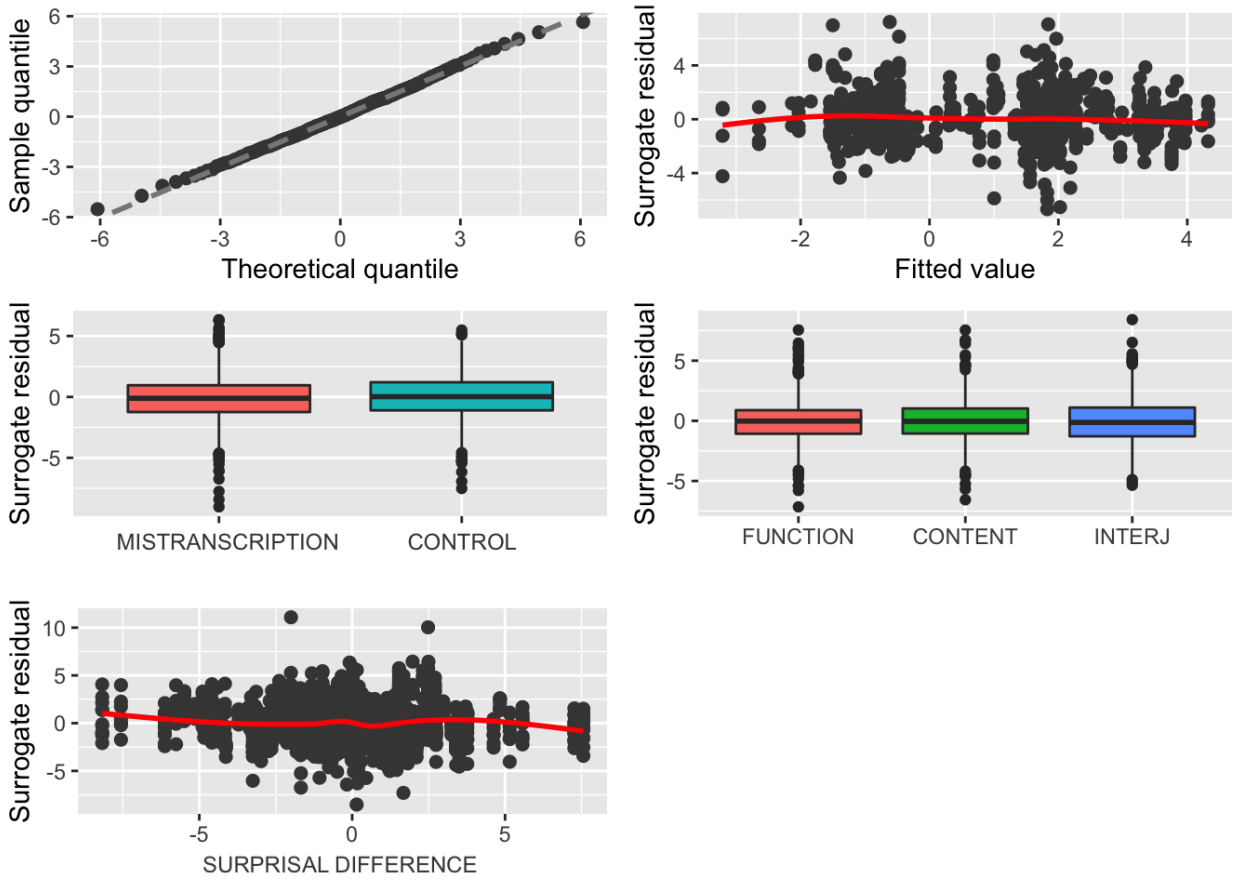


Figure 5.6: Diagnostic plots for the OLR model with surrogate residuals. At top left, a quantile-quantile plot. At top right, a residual vs. fitted values plot. In the middle row, the residuals plots for the isBaseline (left) and Lexical Type (right) predictor. In the bottom row, the residual plots for the SurprisalDiff predictor.

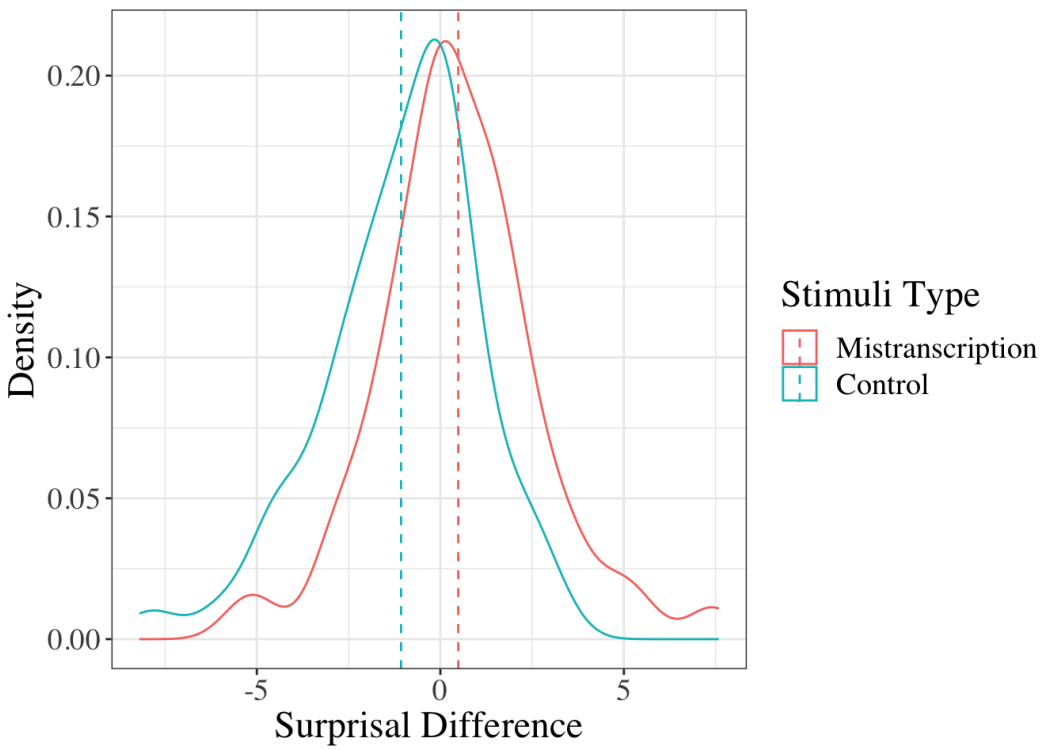


Figure 5.7: A density plot of the surprisal differences found in the mistranscription and control datasets. Dashed lines indicated the averages for each group.

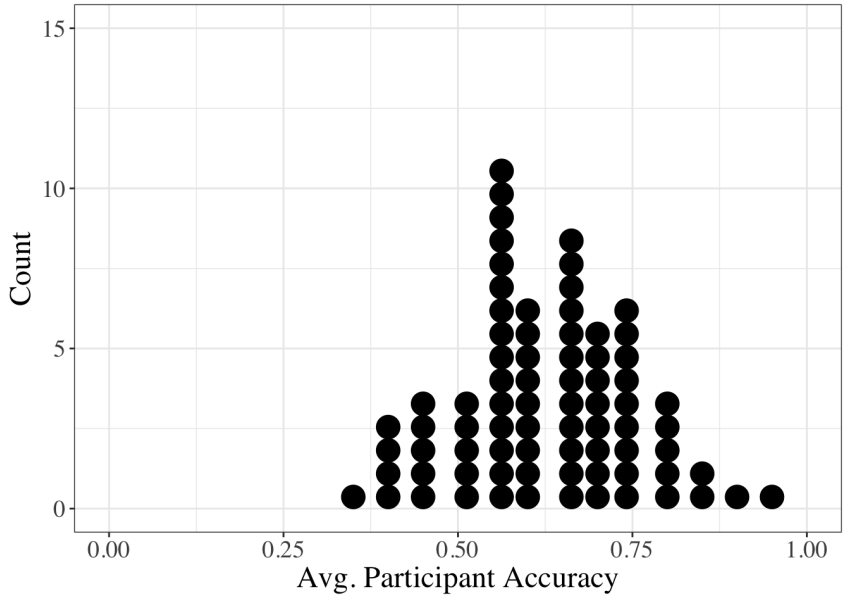


Figure 5.8: The average accuracy of each participant including both mistranscription and control stimuli. Each dot represents one study participant.

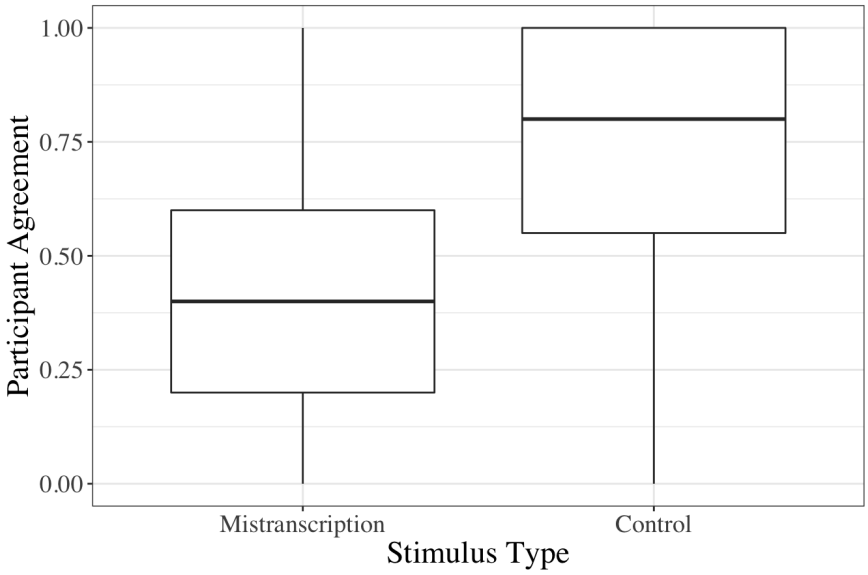


Figure 5.9: The average participant agreement score for each stimulus, categorized into mistranscription and control groups. A score of 1 represents perfect agreement.

Chapter 6

COMPARISON OF ASR AND HUMAN TRANSCRIPTION ERRORS

6.1 Introduction

Previous chapters have considered the role of various lexical features and their relationship to misperception. It was shown that predictability and other factors had an association with the errors made by humans in transcription and in a forced-choice listening task. This chapter considers human transcription errors but also introduces a set of machine transcriptions generated by a commercial ASR system. The transcriptions are part of the same CTS set and therefore provide a valuable avenue to directly examine differences in misrecogntion between humans and ASR.

Previous research has examined a wide range of linguistic factors implicated in speech recognition errors (Goldwater et al., 2010). However, limited work has been done directly comparing the linguistic features of ASR and human transcription errors. Stolcke and Droppo (2017) produced a comparison of such errors, but their work was limited to an examination of the most frequent error types and a correlation of human and ASR errors across speakers. This chapter intends to build upon their work to produce a fuller picture of the difference between ASR and human transcription errors.

This work is motivated by the importance of human transcription as a benchmarking tool for ASR. Researchers have produced ASR systems which have purportedly reached or surpassed human performance, also known as achieving human parity. Xiong et al. (2016) reduced the WER of their ASR system below that of professional human transcribers in experiments on the same audio. However, claims that ASR has surpassed human performance must be understood in context. Claims of human parity are based on the widely adapted

WER metric, which is a simple calculation based on the number of insertions, deletions, and substitutions in a transcript[1]. It is easy to conceptualize, for example, how two deleted words can differ in their severity. In the tradition of Information Theory, each word carries a certain amount of information in the signal. The deletion of a high-density word carrying more information will have a stronger negative impact on sentence meaning.

Developing better ASR metrics or evaluating the performance of a particular ASR system is out of scope in the current study (however, refer to McCowan et al. (2004); Mishra et al. (2011)). Instead, this chapter will analyse differences in a number of lexical and information theoretic features in errors that arise in ASR vs. human transcripts. This goes beyond the work of Stolcke and Droppo (2017) which focused on the most frequent error types. Understanding specific strengths and limitations of recognition in ASR systems can further aid in the evaluation and development of speech recognition systems. Likewise, understanding differences between recognition by machines and humans can provide insights into the uniqueness of human speech perception.

The rest of the chapter will be divided as follows: The following section will consider the dataset from Stolcke and Droppo (2017) which is used to directly compare human and ASR transcriptions. Included is a detailed discussion of the data cleaning and a WER comparison. Section 6.2 will describe the features which were chosen for analysis and the statistics used to analyse them. Section 6.3 will examine the results of the statistical analysis and provide visualizations of the distributional differences. Finally, Section 6.4 will provide a more detailed discussion of the results and their implications, with final conclusions presented in Section 6.5.

6.1.1 Data

In order to directly compare human transcriptions and ASR transcriptions over the same audio samples, this chapter uses the dataset from Stolcke and Droppo (2017). The dataset is

[1]See Chapter 2 for the definition of WER and further discussion.

described below, followed by the process of data cleaning. A comparison of the WER of the errors from this work and the work of Stolcke and Droppo (2017) is then detailed.

6.1.2 Dataset

The dataset from Stolcke and Droppo (2017) consists of transcripts of the NIST 2000 HUB5 CTS evaluation dataset (Fiscus et al., 2000). The dataset contains 20 held-out conversations each from the original Switchboard corpus (Godfrey et al., 1992) and the CallHome English corpus (Canavan et al., 1997; Kingsbury et al., 1997).

The human transcription data was produced using Microsoft's production transcription pipeline. The audio is transcribed in a first pass and then error corrected in a second pass. Stolcke and Droppo (2017) write that transcribers are provided with "no special instructions" regarding transcription. The transcribers were not familiarized with LDC conventions, but were privy to the purpose of the transcriptions. Transcribers listened to audio segments roughly corresponding to utterances from a single audio channel (and therefore one speaker in the dyad at a time). In this way, they received the same input as the ASR system.

The ASR data was produced with Microsoft's then state-of-the-art speech recognition system (Xiong et al., 2016). The acoustic training data uses 2000 hours of CTS. The LMs use CTS data, LDC Broadcast news data, and data from the University of Washington conversational Web corpus. The system utilizes multiple decodings from CNN and Bi-LSTM models which differ in various parameters. They use a 4-gram N-gram model to create lattices which are expanded into 500-best lists. These N-gram lists are then rescored with several forward and backward LSTMs. See Xiong et al. (2016) for a more detailed discussion of the ASR system.

6.1.3 Data Processing

Both the NIST 2000 dataset and the transcription data underwent cleaning and normalization before scoring. Stolcke and Droppo (2017) write that they performed normalization on the human transcripts to better align with the NIST conventions, but they do not provide

additional details. The normalization process between this work and Stolcke and Droppo (2017) is therefore not identical, and the resulting WER (shown in the next section) deviates from their reported WER. The data cleaning process used here closely follows the 2000 NIST Evaluation plan.[2]

Comments, punctuation, etc. are removed from the reference script. The NIST SCTK scoring option for fragments allowed for a transcription to match a reference fragment in several ways. The transcriber could either ignore the reference fragment or transcribe a token which began with the character(s), e.g. 'the' or 'those' for 't-'. Contractions were not tokenized in the scripts (i.e. 'can't' to 'can n't'). They were maintained as full words in their phonological forms. However, the scoring allowed for either a contraction or an expanded form in the hypothesis (e.g. 'don't' or 'do not'). Hyphenated words in the transcripts were expanded to multi-word forms (e.g. 'mother-in-law' as 'mother in law'). Abbreviations in the transcripts were specially handled to match the reference set. In the machine transcriptions, the abbreviations were simply stripped of punctuation. In the human transcriptions, abbreviations which were at times transcribed as a single token were expanded (e.g. 'dj' to 'd j', 'phd' to 'p h d'). Furthermore, other stylistic differences were normalized, including reductions ('gonna' was expanded to 'going to') and compounds ('everyday' could alternatively be hypothesized as 'every day').[3]

Acknowledgements and backchannels were normalized in a category '%bcack', and hesitations in a category '%hes', in line with the 2000 NIST evaluation plan. An alignment was also generated with ungrouped hesitations, in order to examine token frequency differences in Section 6.4. In the remainder of the analysis, the hesitations are grouped.

6.1.4 Scoring and WER

The NIST SCTK scoring package was used to score the reference data against the human and machine hypotheses. The WERs for the normalized data are presented in Table 6.1,

[2]See https://mig.nist.gov/MIG_Website/tests/ctr/2000/h5_2000_v1.3.html

[3]The preprocessing code is publicly available at https://github.com/cmansfield8/human-parity

	Stolcke and Droppo (2017)		Mansfield (2021)	
	Human	Machine	Human	Machine
Insertion	0.7	1.3	0.7	1.7
Deletion	8.4	4.5	8.6	3.8
Substitution	4.3	7.8	4.7	7.4
Total	**13.3**	**13.6**	**14.0**	**12.9**

(a) Callhome data

	Stolcke and Droppo (2017)		Mansfield (2021)	
	Human	Machine	Human	Machine
Insertion	0.7	0.7	0.9	1.0
Deletion	2.8	2.0	3.3	2.3
Substitution	2.7	3.4	2.8	3.5
Total	**6.2**	**6.1**	**7.0**	**6.7**

(b) Switchboard data

Table 6.1: WER for the Callhome and Switchboard portions of the transcriptions.

alongside the WERs reported in Stolcke and Droppo (2017).

There is some difference in WER between the currently reported numbers and Stolcke and Droppo (2017) due to differences in the preprocessing of the data. One important difference is the current choice to not tokenize data such as contractions, but to treat contractions as single (and psycholinguistically real) unit of the lexicon. This led to a difference in the word counts of the reference data. The reference data of Stolcke and Droppo (2017) contains 21.6K and 21.4K tokens from Callhome and Switchboard respectively. The reference used in this chapter contains only 20.2K and 20.0K tokens. Stolcke and Droppo (2017) apply

normalization to the human transcripts, while the current study uses normalization to clean both machine and human transcriptions. STCK's Wilcoxon test and Matched Pairs Sentence Segment Word Error tests did not show significant differences between the WER of the human and ASR transcriptions. However, a Sign test found differences favoring the ASR system for both the Callhome and Switchboard datasets.

6.2 Analysis of features

This section describes the lexical and probabilistic features that will be used to compare human and ASR errors. It then describes the statistical methodology used to compare errors and discusses advantages and disadvantages of the chosen methodology.

6.2.1 Features

The analysis will focus on seven features related to the transcription errors. Features are described in more detail below.

Error Type

Error types consist of insertions, deletions, and substitutions made by the transcribers. Previous work has proposed modified WER metrics which weight each error type differently or consider the information load of the error (Morris et al., 2004; Mishra et al., 2011). Their metric was better correlated with human evaluations of transcription faithfulness, suggesting that different types of errors are more or less severe to listeners.

Some parts of the analysis consider error sequences, which consist of one or more errors in a row, where individual tokens can contain multiple error types. In the case of such sequences, types are grouped by insertions (which may include additional substituted tokens), deletions (which can likewise include substituted tokens), and substitutions (sequences with *only* substituted tokens).

Lexical Category

This analysis looks at broad lexical class as a factor. There are three broad categories proposed. Function words include closed-class (grammatical) words, content (open-class) words, and 'discourse' words. The latter category includes tokens such as hesitations or filled pauses ('uh', 'um'), acknowledgments ('yeah', 'okay'), and backchannels ('uh-huh', 'mm-hm'). All sentences have been POS-tagged with a bi-directional LSTM tagger (see Chapter 4 for more detail). They are further mapped onto broad lexical categories based on their POS and a whitelist of function words which is used for certain ambiguous POS tags (see Appendix A). When sequences of multiple errors are considered, the lexical classes consist of either tokens which share a single lexical class (e.g. insertion of multiple function words belongs to the function class) or a mixed class where error tokens include multiple lexical categories.

Unigram Probability

At each token in the reference and hypothesis text, the unigram probability was calculated using SRILM's n-gram model (Stolcke, 2002). The unigram model is trained on 17M tokens from the Fisher corpus (Cieri et al., 2004). Following Goldwater et al. (2010), log probabilities are used.

Linguistic Surprisal

Linguistic surprisal is an information-theoretic measure of the expectedness of a word in a sentence, where a value of 0 represents an absolutely expected word (of probability 100%), where surprisal increases as the likelihood of the word decreases (Hale, 2001; Levy, 2008). It can be formulated as in Equation 6.1:

$$H(w_i) = -\log_2 \; p(w_i|c(w_i)) \tag{6.1}$$

where $p(w_i|c(w_i))$ describes the probability of the i^{th} word in a sequence given some prior context $c(w_i)$ according to a language model. The linguistic surprisal of the word at each

error is calculated by first generating the conditional probability of the word using a GRU language model. The GRU was likewise trained on the Fisher corpus (Cieri et al., 2004). Out-of-vocabulary words are modeled with an 'unknown' token. Tokens with less than 10 instances are modeled as 'unknown'. The unknown probability is therefore an over-estimate of the probability of unseen words in the data. More details about the model can be found in Chapter 4.

Surprisal Difference

The surprisal difference is a measure of the predictability of some sequence of hypothesis text compared to the predictability of the reference text. Chapter 4 showed that certain human transcriber errors were associated with a more positive or negative surprisal difference. For instance, sequences in the hypothesis which contained insertions were more expected than corresponding sequences in the reference. The surprisal difference gives an estimation of whether the transcriber's hypothesis is more or less expected than the reference text in cases where certain word sequences disagree.

The surprisal difference is the difference in surprisal between the reference and hypothesis. Surprisal of the hypothesis is calculated by averaging the log probabilities of a sequence starting with the first insertion or substitution token and concluding with the first non-error token in the sequence. The surprisal of the reference is calculated likewise, including the first reference deletion or substitution and the following non-error token. In other words, these are sequences bracketed by error free forward and back contexts to ensure conditioning events are comparable. This can be shown in Equations 6.2a and 6.2b as:

$$H(w_1 w_2 ... w_n) = -\frac{1}{n} \log_2 \prod_{i=1}^{n} p(w_i | c(w_i)) \tag{6.2a}$$

$$\Delta H = H(w_1^r ... w_{n_r}^r) - H(w_1^h ... w_{n_r}^h) \tag{6.2b}$$

where $p(w_i|c(w_i))$ is the probability given by the language model, and w_i^r and w_i^h are the i_{th} tokens of the hypothesis and reference sequences. Note in some cases the hypothesis or

reference sequence will be only one token long. For instance, if there is a single insertion and following non-error token in the hypothesis sequence, the reference sequence would contain only the matching non-error token.

Relative utterance position

The relative sentence position measure considers the position of the error within the segmentation or slash-unit (the sentence-like unit used for data segmentation). It is a measure of the position of the error in the reference sentence divided by the reference sentence length. An insertion before the initial word of the slash-unit would have a value of 0, while an insertion at the end of the slash-unit would have a value of 1. Errors which span the entire length of the slash-unit and would thus be ambiguous with regards to position are not assigned a position value.

Utterance length

Utterance length looks at the total length of the utterance where the error is found. It is measured per error, rather than per sentence. In other words, sentences which have multiple errors will contribute multiple datapoints to the utterance length measure.

6.2.2 Statistical Analysis

Wilcoxon signed-rank tests will be used to compare the counts of errors throughout the chapter. Density plots will be used as a diagnostic to further visualize differences in the distribution of the probability and surprisal differences between the human and machine transcriptions. Density plots are a type of probability distribution function which plots values of a particular variable on the x-axis and density on the y-axis. The area under the curve is always equal to 1. The significance testing will be described in more detail below.

Wilcoxon signed-rank tests

For an analysis of transcription errors, several potential issues must be considered. The distribution of errors are not independent, as error tokens can be in close proximity to each other in an utterance. The transcriptions from the human transcribers and machine transcribers rely on the same data. Therefore, a paired test is appropriate. A Wilcoxon signed-rank test is used to consider differences in counts between the human and machine errors. Error counts from the same utterances are extracted from the human and machine transcription data as pairs. The Wilcoxon signed-rank test considers the null hypothesis that differences between the two distributions are symmetrically distributed around zero. The alternative hypothesis states that differences are not symmetrically distributed around zero. Thus, for a given distribution $i = (1, ..., N)$, the sign function of $|x_{2,i} - x_{1,i}|$ is calculated. The pairs are ordered by difference from smallest to largest and ranked. The test statistic W is calculated as shown in Equation 6.3:

$$W = \sum_{i=1}^{N} [sign(x_{2,i} - x_{1,i}) * R_i] \tag{6.3}$$

Approximation is used to calculate a z-score and p-value. Significance is assumed for $p<0.05$. The Wilcoxon signed-rank test is non-parametric, in other words it does not assume a particular underlying distribution of the data.

6.3 Results

The following section presents the results of the statistical tests described in Section 6.2. First, the distribution of ASR and human transcriber errors by error type and lexical category is considered. Next, individual error probability and surprisal are described. Analyses of the error sequences and the surprisal difference between sequences is presented. The section concludes with findings related to utterance position and utterance length.

6.3.1 Error Distribution

There are 3,816 errors in the human transcriptions and 3,718 errors in the ASR transcriptions. There are 3,459 substitutions, 3,259 deletions, and just 816 insertions when including both human and ASR transcribers. Among the transcriptions, there are considerably more errors from the Callhome transcripts, with 2,480 and 2,427 errors from human and ASR transcripts respectively. Callhome errors make up approximately 65% of the total errors. This is not unexpected, as previous ASR evaluations have noted higher error rates in Callhome (Fiscus et al., 2000).

A plot of the relative error counts by insertions, deletions, and substitutions is shown in Figure 6.1. Significance is reported via Wilcoxon tests in Table 6.2. Across both datasets, humans are more likely to miss tokens, while ASR is more likely to insert or misrecognize tokens. Human and ASR errors are significantly different for each of the error types with the exception of insertions in Switchboard, which are similar. The largest discrepencies between human and machine transcribers are seen in Callhome deletions.

	Switchboard		Callhome	
	Z-score	P-value	Z-score	P-value
INS only	-1.78	NS	-6.64	$p < 0.0001$
DEL only	6.38	$p < 0.0001$	13.32	$p < 0.0001$
SUB only	-4.27	$p < 0.0001$	-10.74	$p < 0.0001$

Table 6.2: Wilcoxon test of error counts at the utterance level.

Another analysis of error types considers sequences of errors, where errors are present in a span of multiple tokens next to each other. Table 6.3 shows error sequences. In some cases, multiple error types are present in a sequence. For example, a transcriber may misrecognize 3 tokens but substitute only 2 tokens in its place. This would result in a 'deletion-substitution'

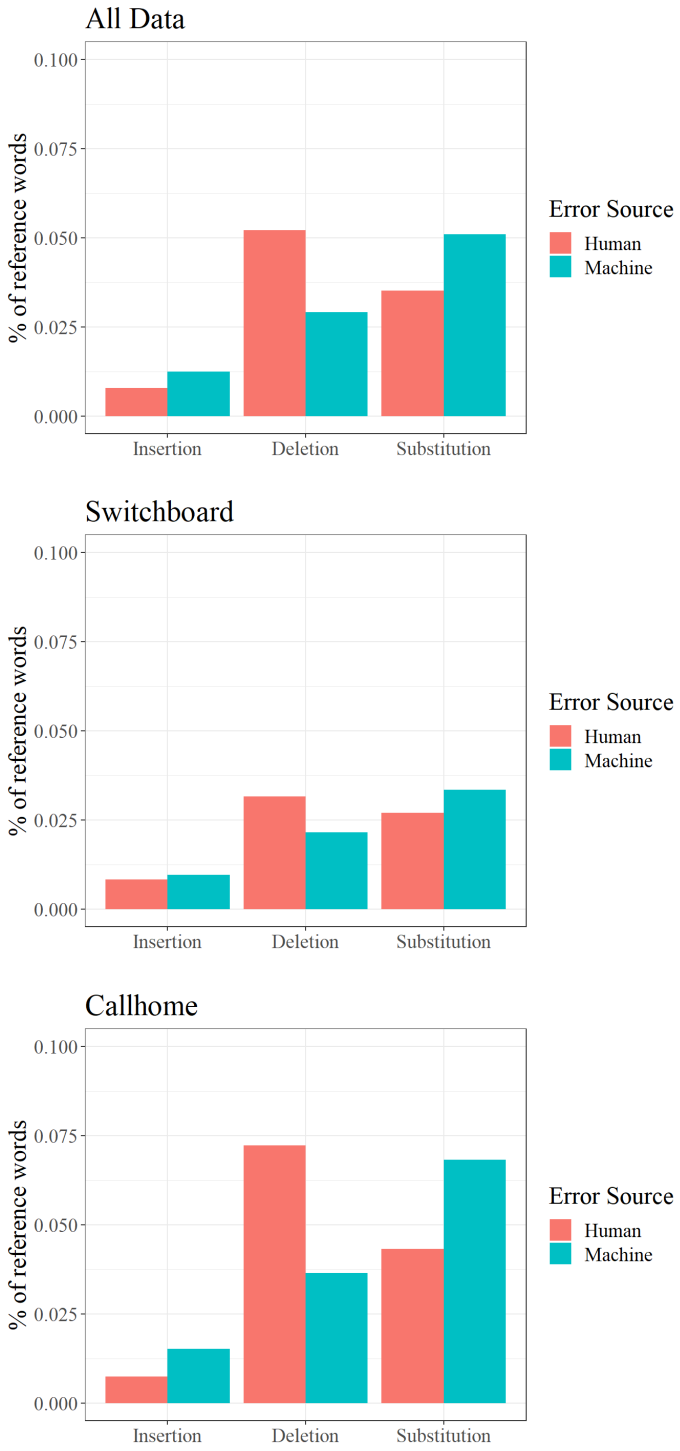


Figure 6.1: Proportion of errors of each type as a total of the reference tokens.

type error. Deletion-substitution errors are relatively common, making up approximately 10% of human errors and 12% of ASR errors.

	Human	Machine
INS only	209	239
DEL only	1120	678
SUB only	905	1219
INS-SUB mix	80	189
INS-DEL mix	0	0
DEL-SUB mix	260	314
INS-DEL-SUB mix	0	0

Table 6.3: Error sequences - count of error types.

6.3.2 Error Lexical Class

Next, the lexical class of the error is considered, including function, content, and discourse words. Differences in lexical class were considered for each error type: insertions, deletions, and substitutions. The absolute counts by lexical category for all errors are shown in Figure 6.2. In the reference data, there are approximately 23.9K function words, 14.8K content words, and 3.4K discourse words. Figure 6.3 presents the number of errors in each lexical class as a proportion of reference tokens from each corpus, which allows for comparisons across classes. This provides some idea of how likely human or machine transcribers are to make a particular error given an instance of that lexical class. Notably, the difference in lexical class proportions are greater in the Callhome corpus.

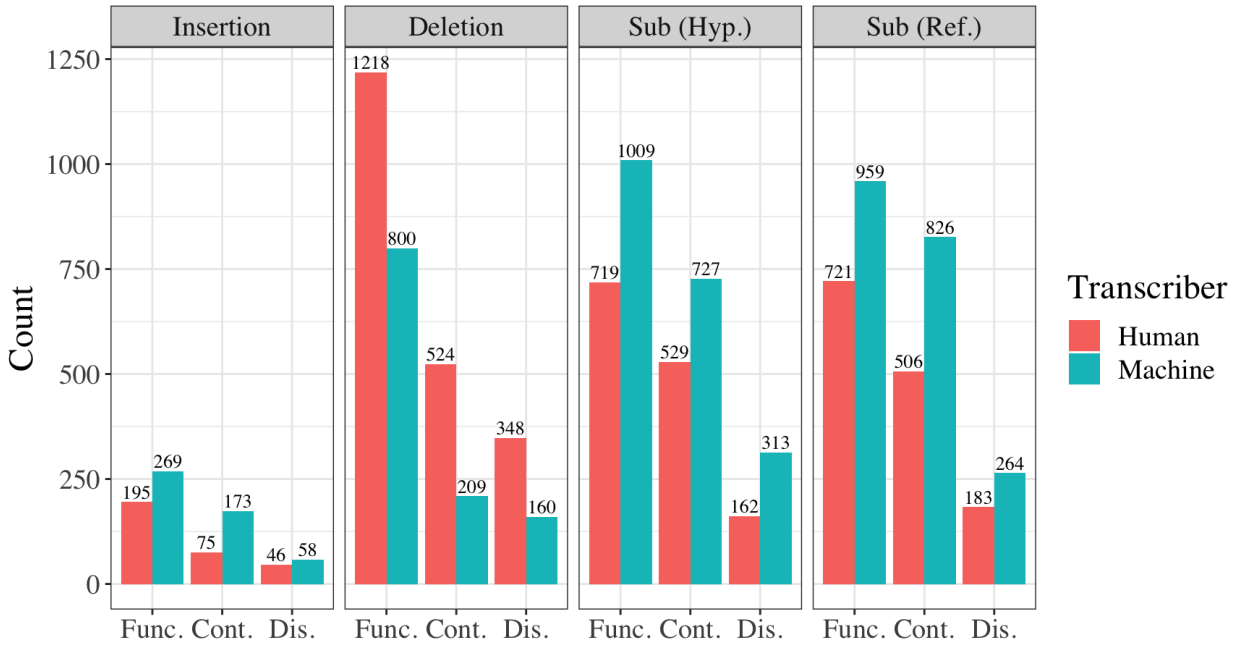


Figure 6.2: A count of the total errors made by the lexical class (x-axis: Function, Content, and Discourse) for human and machine transcribers.

Function words

While function words make up the greatest number of errors in absolute terms (see Figure 6.2), this can be explained by their high total frequency in the corpus (function words make up about 57% of all tokens). As a proportion of their frequency, function words are the least likely to be inserted and misrecognized in general. Function word insertions are balanced similarly across both Switchboard and Callhome. ASR is slightly more likely to misrecognize function words than human transcribors. Function words deletions differ most between human and ASR transcribers. There are 24% and 39% fewer ASR than human deletions in Switchboard and Callhome, respectively.

Content words

Content words errors are the second most common type of error in terms of both absolute frequency and relative frequency in the corpus. Differences in content word deletions and substitutions in the Callhome corpus are particularly remarkable. There are 61% fewer ASR content deletions in Callhome. There are notably 80% more ASR than human content misrecognitions in Callhome.

Discourse words

While discourse words make up a small amount of errors in absolute terms, they are proportionally more often hallucinated, missed, or substituted. A Wilcoxon test of discourse word errors is shown in Table 6.2. Discourse words make up the highest proportion of insertions in both Switchboard and Callhome. However, according to the Wilcoxon test, they do not differ significantly between ASR or human transcribers. Discourse words are relatively likely to be deleted by both human and ASR transcribers. However, human transcribers delete them significantly more often, as shown in Table 6.4. Humans delete more than 12% of discourse words in the Callhome corpus, compared to only about 4% by ASR. There are also significant differences in the misrecognition of discourse words by humans and ASR, with ASR significantly more likely to misrecognize them – a 60% increase over human transcribers in Callhome. Among both human and machine transcribers, discourse words are the most likely to be misrecognized.

	Switchboard		Callhome	
	Z-value	P-value	Z-value	P-value
INS	-0.49	NS	-1.441	NS
DEL	4.95	$p < 0.0001$	9.56	$p < 0.0001$
SUB (hyp.)	-3.43	$p < 0.001$	-7.66	$p < 0.0001$
SUB (ref.)	-2.49	$p < 0.05$	-4.44	$p < 0.0001$

Table 6.4: Wilcoxon test of discourse word counts for paired human and machine utterances.

6.3.3 Unigram Probability

Next, the unigram log probability at each error type is considered. The distributions of unigrams are shown in Figure 6.4. These figures do not include words which are out-of-vocabulary according to the Fisher language model.

Insertions, which are fewer in human transcriptions, are concentrated more strongly around high-probability tokens in the human transcriptions. ASR transcriptions have a greater mass of insertions errors at lower probabilities. While substitutions appear similar overall, there is a spike associated with hypothesized substitutions by ASR transcribers. These spikes correspond to the backchannels 'uh-huh' and 'um-hum'.

There are some notable differences in out-of-vocabulary words from Fisher. Note that the OOV probability is defined by the language model, and does not necessarily reflect the vocabulary of the ASR system. In any case, these words are assumed to be low-frequency content words. In substitutions, humans hypothesize more words which are OOV according to the Fisher language model, at 153 words versus 66 in ASR. On the other hand, ASR transcribers are much more likely to misrecognize an OOV reference word. While 110 OOV words are misrecognized by human transcribers, 299 are misrecognized by ASR transcribers.

6.3.4 Surprisal

The distribution of surprisal values is examined next. The plot of the the surprisal values is shown in Figure 6.5. One point to note is that the references contain a number of out-of-vocabulary words. Because the GRU model uses the history of the utterance to assign a probability, any OOVs in prior utterance context will add noise to the surprisal estimate of an error.

The shape of the surprisal distribution for insertions is similar to that of the log probability in Figure 6.4. The surprisal means are 6.49 and 7.90 for human and ASR transcriptions respectively. In the case of hypothesized substitutions, humans show a greater density of surprisal values around 5-10, while ASR has a denser distribution where surprisal is greater than 10.

6.3.5 Surprisal Difference

The surprisal difference considers surprisal averaged across all errors in a sequence, and takes the difference between the surprisal in the reference and hypothesis. A positive surprisal difference is present when a hypothesized substitution is more expected than the reference. Distributional differences are found to be significant for all types of errors. Figure 6.6 shows the density distributions of the surprisal difference.

A closer examination of each type can shed light on the findings. Figure 6.7 shows the surprisal difference for insertions broken down by lexical category. In content and mixed lexical category sequences (where 93% of sequences contain content words), ASR has greater density in the negative surprisal difference range. This closely matches the overall distribution. A negative surprisal difference describes an insertion which is more surprising than the reference.

Regarding deletions, Figure 6.8 shows a considerable discrepancy in the surprisal differences of discourse tokens. In particular, there is a large spike around the surprisal difference of -5, which corresponds to the deletion of the %bcack and %hes tokens where no other

reference words are present (-5.71 and -5.14). Human transcribers deleted these tokens in the aforementioned context 72 times, and ASR only 16 times.

The differences in the distribution of substitution errors are likewise notable. The distribution of reference lexical category is shown in Figure 6.9.[4] The function word and mixed categories roughly mirror the overall surprisal difference distributions. Discourse substitutions have a spike around 0.5. This corresponds to the substitution of single-token utterances including hesitations, backchannels, and acknowledgments. The surprisal difference in ASR is more positive than that of human transcribers.

6.3.6 Error Utterance Position

Next, the position of each error within the utterance is considered. Figure 6.10 shows the distribution of utterance position for each error type. For ASR, there are relatively more deletions at the beginning of the utterance, while human-transcribed deletions are more evenly distributed throughout the sentence.

6.3.7 Utterance Length

The last feature which will be reported is the length of the utterance at each error. ASR tends to make more insertions overall, and insertions in sentences of 10 tokens or less are relatively more common. On the other hand, the insertions of human transcribers are more frequent in longer sentences, especially in sentences of about 20-50 words in length.

6.4 Discussion

In this section, the findings for the general error distribution and unigram probability and surprisal will be discussed in greater detail. Some comments on the practical implications of the study follow.

---

[4]The substitution categories show misrecognized words. The distribution of hypothesized words in substitutions is nearly identical.

6.4.1 Error Distribution

A first point to note is that the distribution of insertions, deletions, and substitutions is unequal between human and ASR transcribers. Although the total WER is similar, ASR errors include more substitutions, while humans make more deletion errors. ASR errors appear more balanced in terms of insertions and deletions. This finding is likely an artefact related to how insertions and deletions are tuned by ASR at training time. On the other hand, humans seem more likely to miss than misrecognize words. It is possible, as suggested in Xiong et al. (2017), that when human transcribers have particular difficulty with the intelligibility of a token, they may simply not transcribe at all. The way humans process discourse words is another likely factor, as discourse words are often missed by human transcribers.

What about the distribution of open-class, closed-class, and other types of words? There are some similarities between the way humans and ASR process these words. Discourse words are the most common lexical class of error as a proportion of their frequency. Function words, with their overall high rate of occurrence, are the most common error by absolute counts. In all, lexical categories of insertions and substitutions are similarly balanced. However, there are considerable differences in deletions between human and ASR transcribers. ASR deletes content words rarely. Human transcribers delete a comparatively large proportion of discourse tokens - approximately 10%. A closer look at discourse tokens follows in Section 6.4.3.

6.4.2 Unigram Probability and Surprisal

There are several findings regarding the differences in probability distributions and in the surprisal difference between error and reference sequences. One finding is that ASR inserts more low-probability or surprising things. As described in Stolcke and Droppo (2017), the counts of the most common insertions (such as 'i' or 'and') are similar between human and machine transcribers. ASR inserts similar amounts of these frequent tokens, but also inserts a greater number of lower-probability tokens.

The surprisal difference metric finds that ASR is also more likely to insert more surprising

word sequences relative to the hypothesis. Examples where ASR inserts more surprising sequences are provided below (approximately +5 surprisal difference). Where the transcription differs from the reference, the hypothesis is indicated at the left of the slash.

(9) and uh so she says as soon as she's all through at pizza hut she's going to start **jazz exercise / jazzercise** again

(10) big **bang lee / bangly** necklaces i don't go for that[5]

(11) well i thought it was **a boo / amiable**

These insertions share a commonality: they are situated next to a substitution. In fact, the difference in insertions counts between humans and machines is mostly due to insertion-substitution combinations. There are 189 and 80 insertion-substitutions errors from machines and humans respectively. There are only 239 and 209 insertion-only errors. Another notable detail in the examples above is that the referent of the substitution is OOV. These OOV words are replaced with a string of multiple acoustically similar content words. These examples provide some clues about the general cause of probability differences in insertions.

The probabilities of substitutions differ between ASR and human transcribers. OOV tokens account for some of this difference, as ASR is much more likely to misrecognize OOV reference tokens. According to the surprisal difference, ASR more often suggests hypotheses that are more expected than the reference ($\bar{x}$=0.39). Human transcribers, on the other hand, are about equally likely to propose something more or less surprising ($\bar{x}$=-0.05). Below are several examples where ASR suggests a hypothesis that is more expected than the reference (in the surprisal difference range of 2.5 to 5):

(12) domestic i guess if there were further efforts at disarmament or arms control **like / i** that would probably decrease the threat as well right.

(13) if he's the teacher and like he **calls his son / cut off his thumb** how does he know what he's really doing

[5]The human transcriber substituted 'bangly' for 'dangly' in this sentence.

(14) not **black people / bucked teeth** but people with gap teeth[6]

There are more content substitutions than function as a proportion of their occurrence. In these cases, ASR and human transcribers are about equally likely to propose less surprising or more surprising hypotheses. Here are a few remarkable examples where ASR proposes more surprising hypotheses than the reference (in the surprisal difference range of -2.5 to -5):

(15) now i'm not well then i have to take my exams my **morals / orals** but

(16) and slept through it didn't **french / flinch**

(17) we had one of these little fold out campers with the hard **rough / roof** that you had to like crank up

Here are examples from human transcription (again, focused on content substitutions in the range -2.5 to -5):

(18) i do tax **repairing / preparing** on the side so i'm one of these people that try to get back every penny they take out yeah

(19) it's more of a college **sound / town** alternative you know like that band Nirvana

(20) and we **stood / sit** around too much you know

Although these content substitutions are similar in terms of the surprisal difference, the human transcriptions appear better grounded in real-world knowledge. For instance, it is easy to imagine grunge as a 'college sound' but 'cranking up the rough' does not have a well understood meaning.

6.4.3 Discourse Words

Discourse words (backchannels, hesitations, and acknowledgements) are the most likely words to appear in transcription errors given their occurrence. There are notable differences

[6]This misrecognition error highlights how seemingly innocuous speech input can result in racially biased or otherwise offensive predictions. See https://www.oxfordinsights.com/racial-bias-in-natural-language-processing for a review of racial bias in NLP applications.

Human	Machine
68: uh	45: uh
42: oh	24: oh
42: yeah	12: eh
30: %bcack	12: yeah
25: well	10: ah

Table 6.5: Most frequent discourse word deletions.

in the deletion and substitution of discourse words in ASR and human transcriptions. A Wilcoxon test of discourse words shows that there are significantly more deletions and fewer substitutions by human transcribers compared to ASR. The most common discourse deletions are shown in Table 6.5. This table presents hesitations in their non-normalized form; in other words, the hesitations such as 'uh' and 'um' are not grouped. Backchannels ('uh-huh' or 'mm-hm' and their spelling variants) are grouped in the category %bcack.

Humans tend to delete discourse words from single-token utterances more often than ASR. While there are 30 %bcack deletions in the human transcriptions, there are only 4 in the ASR transcriptions. One question is why humans miss discourse transcriptions more often than ASR. One explanation is that this particular group of transcribers were not given explicit instruction to transcribe such words. It may be important to give transcribers sufficient training to recognize and transcribe these forms, which are not often found in written language. It is also possible that discourse words are processed differently than other words and are therefore more difficult to recall, but there are few studies which examine this. Perception studies typically focus on 'idealized' speech. However, discourse words are shown to provide useful information to the listener; for instance, filled pauses can allow the listener to process disfluencies more quickly (Brennan and Schober, 2001).

ASR substituted more discourse words than human transcribers. Counts are shown in

Human	Machine
106: uh/um	44: uh/ah
50: uh/ah	29: %bcack/hm
23: um/uh	24: um/uh
8: oh/well	16: %bcack/um
8: uh/oh	14: uh/um

Table 6.6: Most frequent discourse word substitutions. The hypothesis is shown before the dash.

Table 6.6. Human transcribers had a large number of confusions relation to 'uh', 'um', and 'ah', but despite this had fewer discourse substitutions overall. ASR was more likely to propose backchannels ('uh-huh' or 'um-hum') for word types such as 'hm' or 'mm', with 72 backchannels hypothesized by ASR vs. only 6 by human transcribers.

6.4.4 Applications

While WER is approximately equal among the ASR and human transcriptions examined here, there are differences in the distributions of error type, token frequency, and other factors. The performance of ASR differs in ways that are distinct from human transcribers, which is important given human performance as a benchmark for ASR performance. The results of the study raise several possible points of interest:

1. The simple metric of WER is unable to capture major differences in distribution between error types and lexical categories or lexical frequency. Insomuch as these factors are related to error severity, their incorporation into ASR evaluation can help to fine-tune ASR systems and guide their development.

2. It is well known that ASR has a problem with recognition of the 'long tail' of open-class

words. In this work, these long-tail OOV words appear to have effects on the differences between inserted and substituted words in ASR and human transcripts. OOV words present a challenge to ASR, and this challenge might not be solved with simply larger and larger data sets - language change has the ability to introduce new coinages or recycle once obsolete words (see Bender et al. (2021) for more discussion on this topic). The speech data which ASR relies on can be several decades old and thus out of date.[7]

3. Discourse words are frequently missed and handled differently by human and ASR transcribers. While human transcribers missed a large number of backchannels, ASR tended to substitute such forms. Human transcribers are likely to miss discourse words and confuse 'um' with 'uh', while ASR hypothesized backchannels ('uh-huh' and 'um-hum') in place of 'hm' and 'mm'. These discrepancies motivate further study.

4. The differences in error types and lexical categories between humans and ASR appear to be exaggerated in the Callhome data. The Callhome dataset, which contains conversations between friends and family, has a more informal style than Switchboard. This illustrates how ASR may have issues adapting to different styles of speech generally. When considering the robustness of an ASR system, speech style should be an important consideration.

6.5 Conclusion

This study analyzed word errors from CTS transcriptions made by professional transcribers and compared these to errors made by an ASR system. A comparison of transcriber errors illustrated differences in features such as type of error, lexical category of the error, unigram probability, and surprisal. While the overall count of ASR and human transcriber errors is roughly equal, it is clear that the characteristics of the errors differ in meaningful ways. Future work should examine how differences in the features of recognition errors reflect error severity

[7]Note nearly 70 mentions of 'aerobics' and 0 mentions of 'yoga' in the Switchboard corpus.

in various speech applications. A deeper understanding of the linguistic aspects of recognition errors can guide ASR towards meaningful improvements through better evaluation.

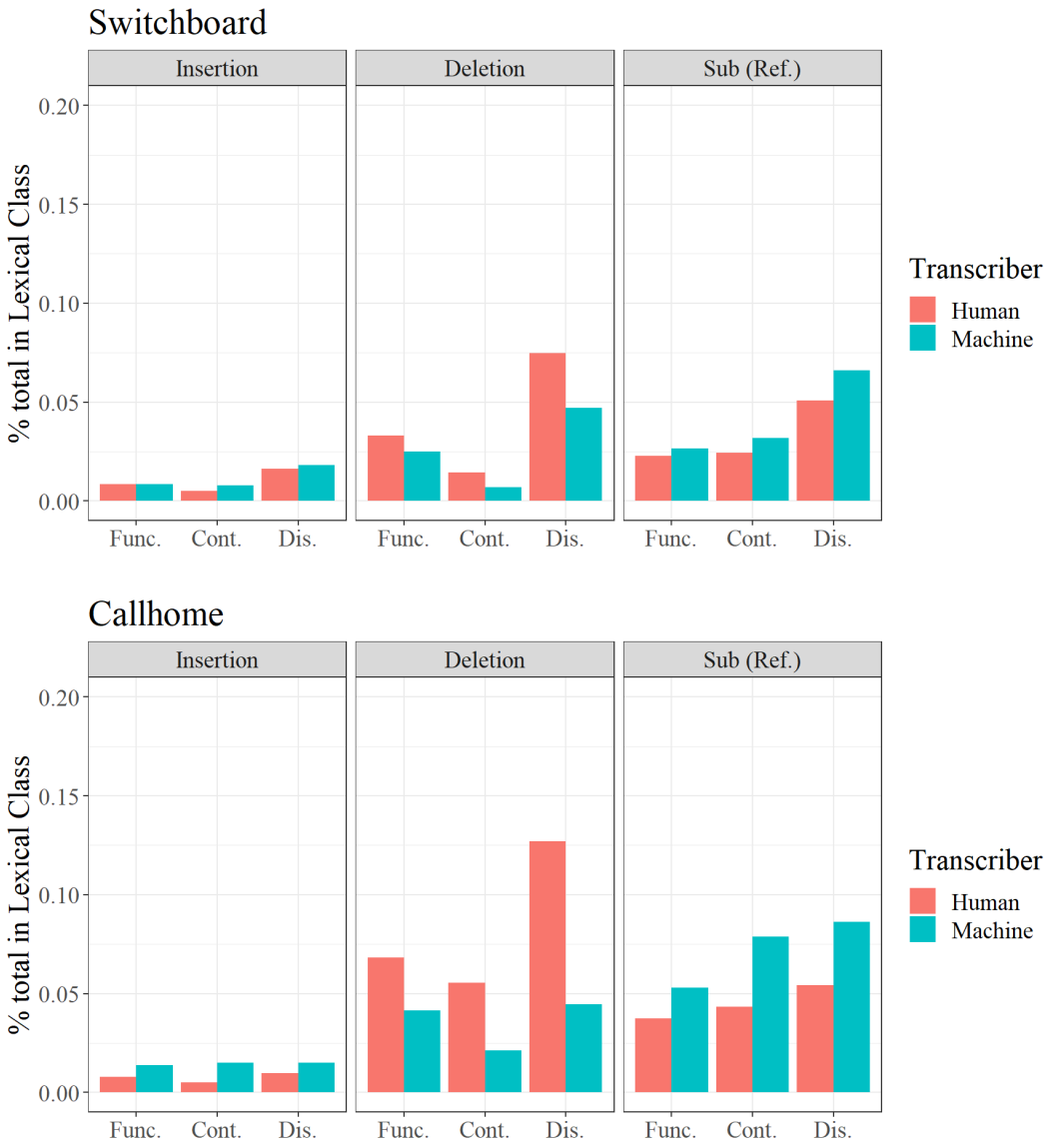


Figure 6.3: The proportion of errors made by transcribers for each lexical class (x-axis: Function, Content, and Discourse) for human and machine transcribers. The proportion is the total number of errors made of each type for the specified lexical class divided by the total number of tokens in the reference for the respective lexical class.

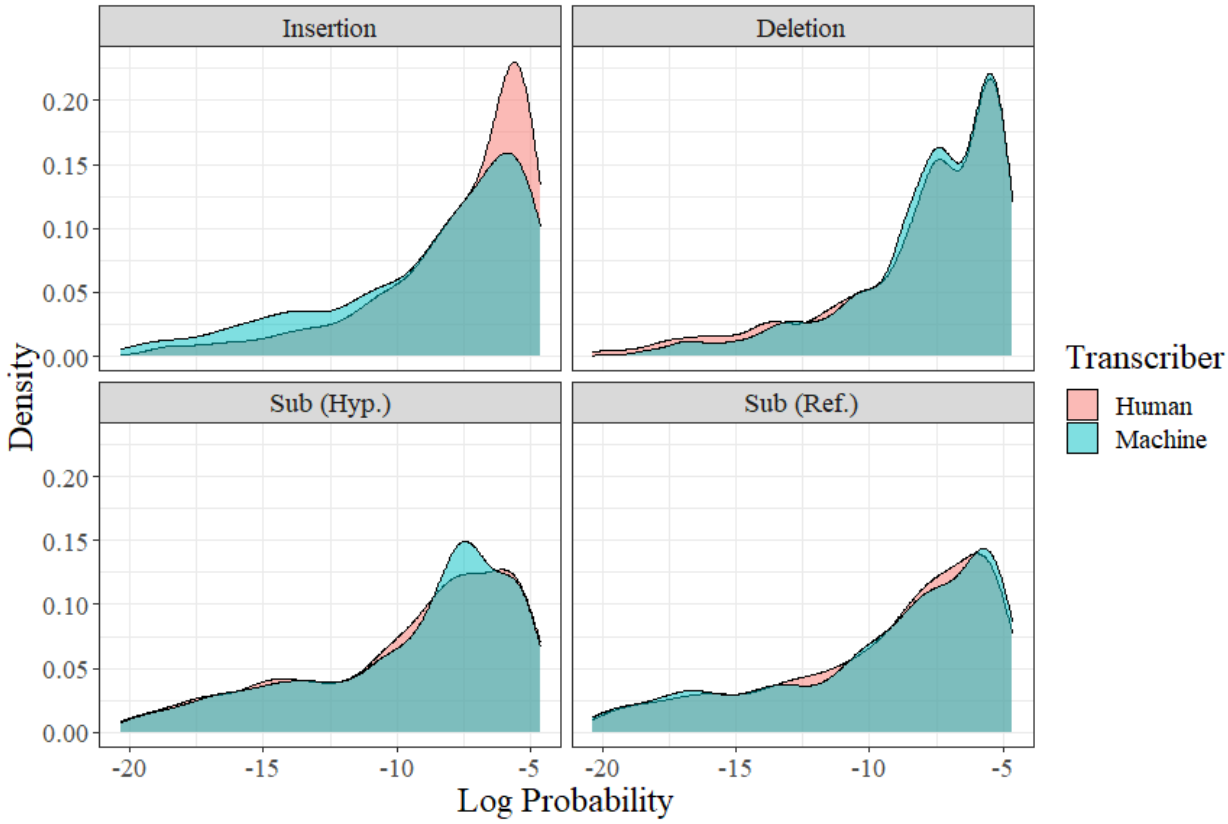


Figure 6.4: Density plot of the log probability of errors.

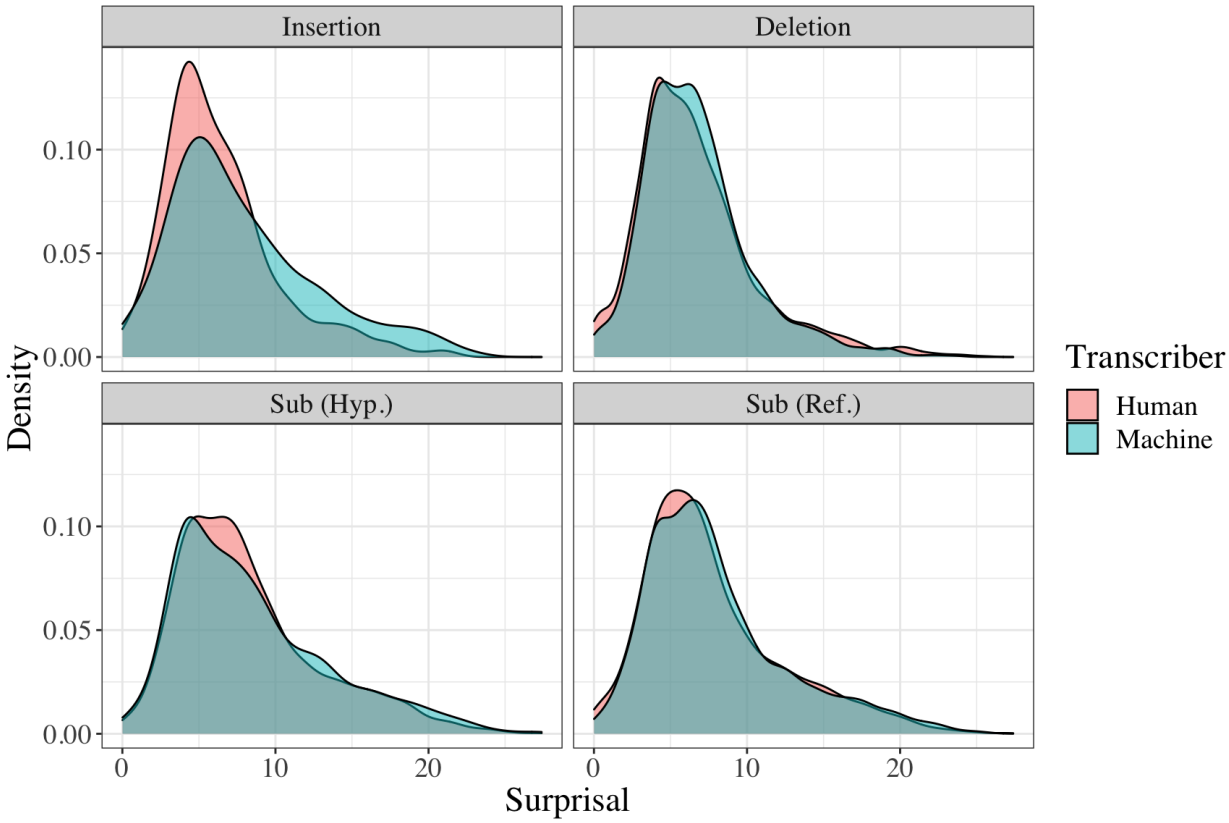


Figure 6.5: Density plot of the surprisal of errors.

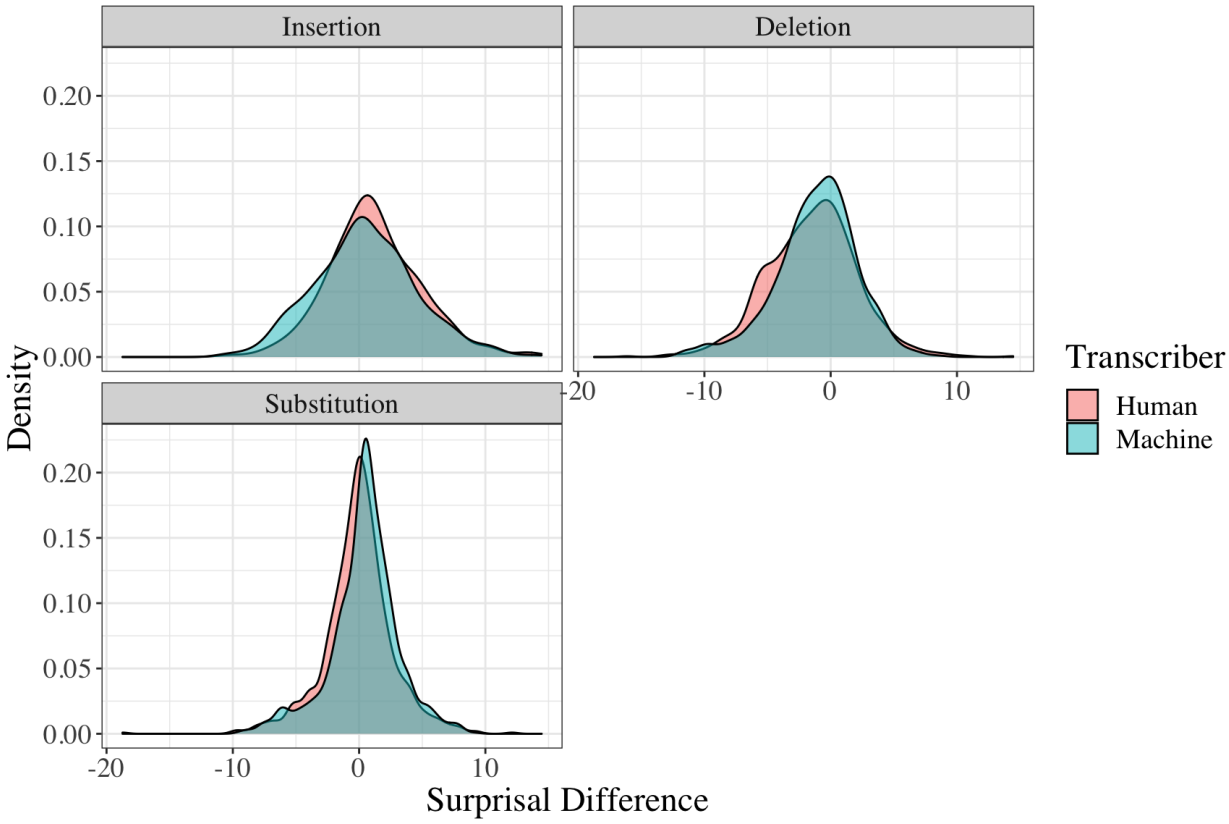


Figure 6.6: Density plot of the surprisal difference of error sequences.

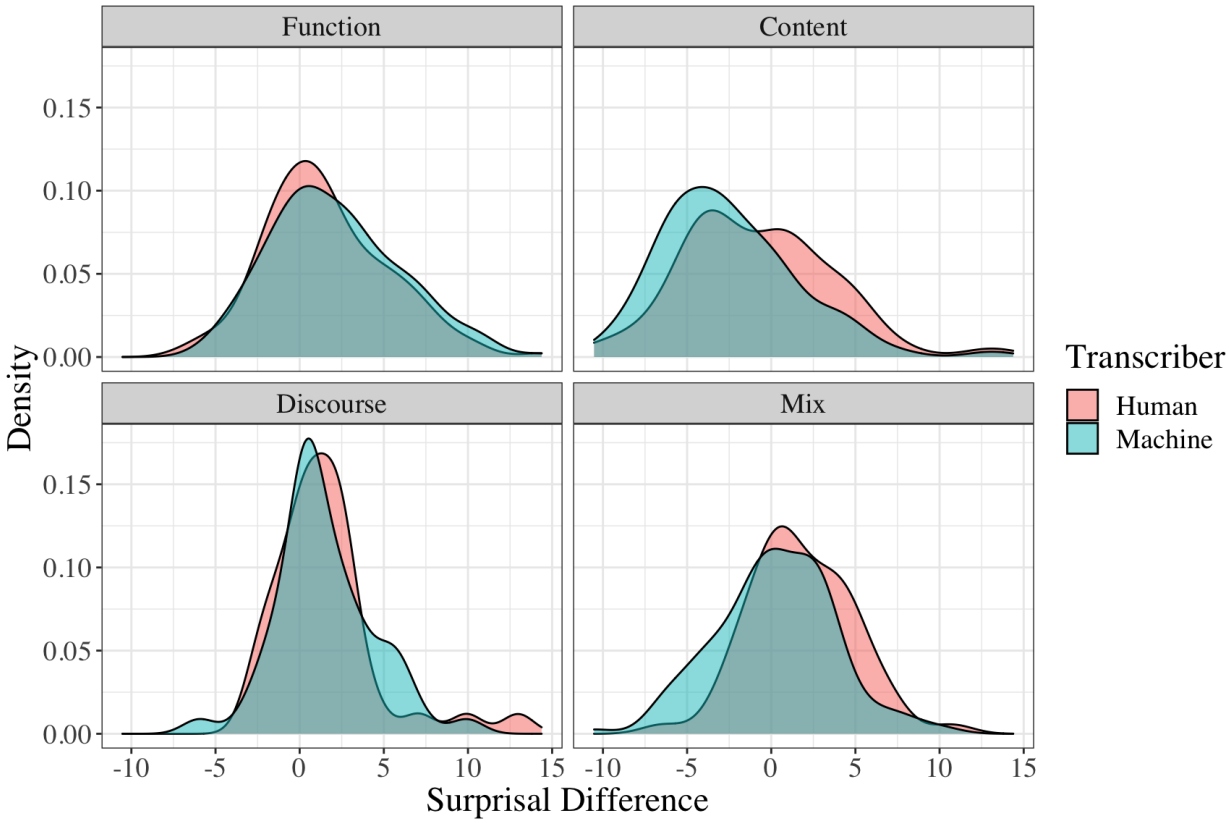


Figure 6.7: Density plot of the surprisal difference of insertion sequences.

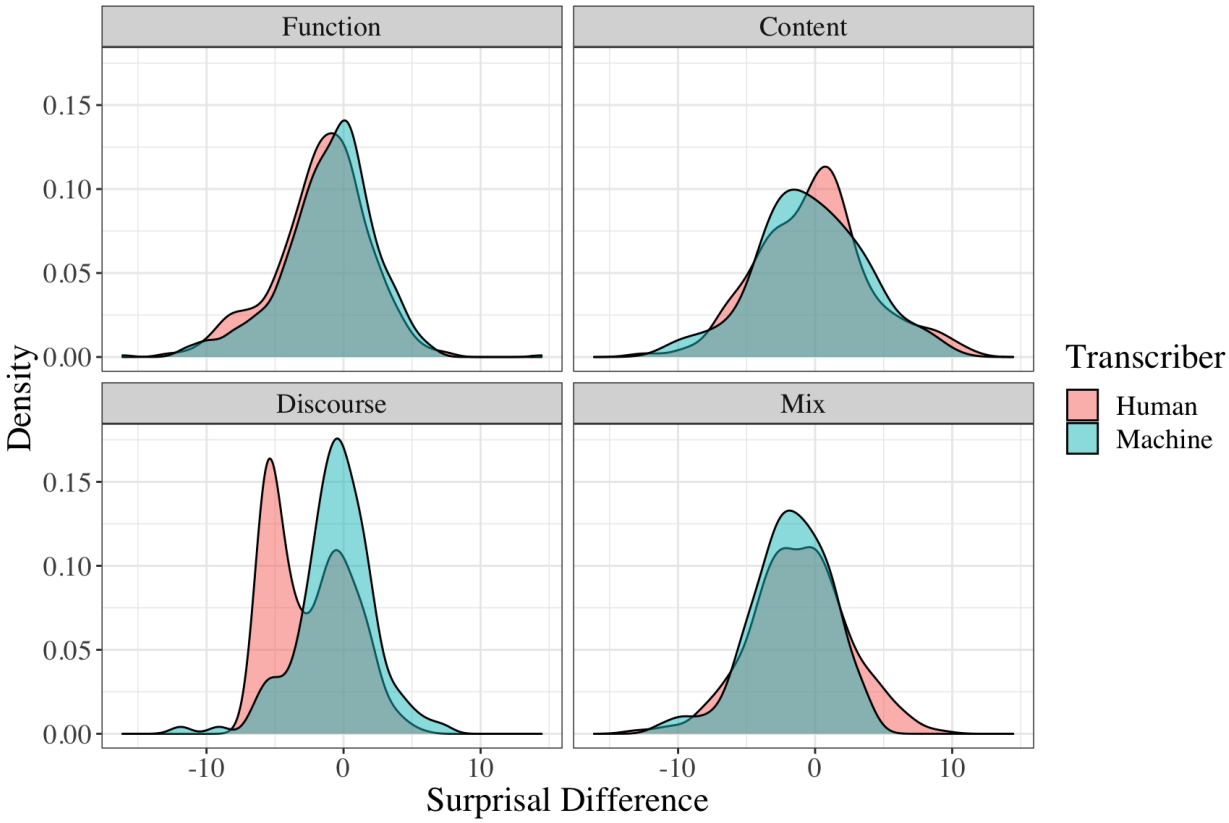


Figure 6.8: Density plot of the surprisal difference of deletion sequences.

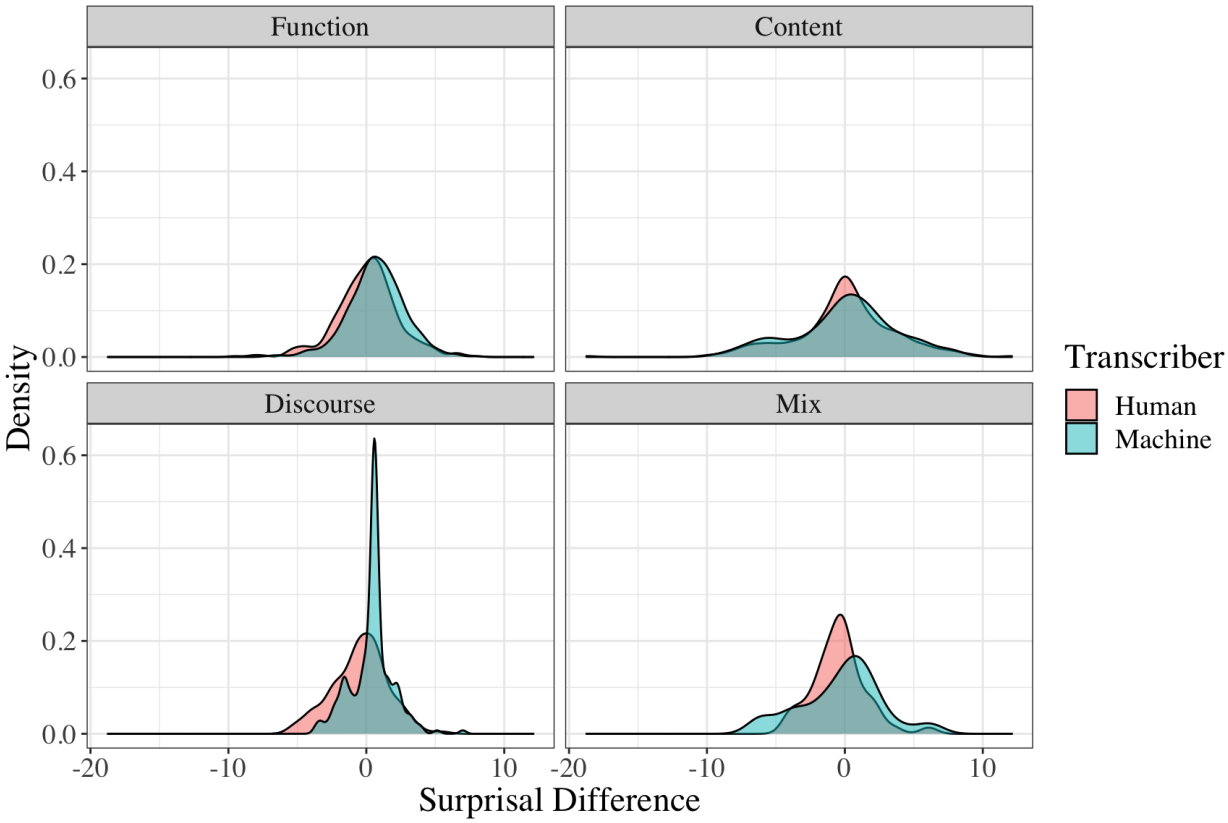


Figure 6.9: Density plot of the surprisal difference of the hypothesized substitutions.

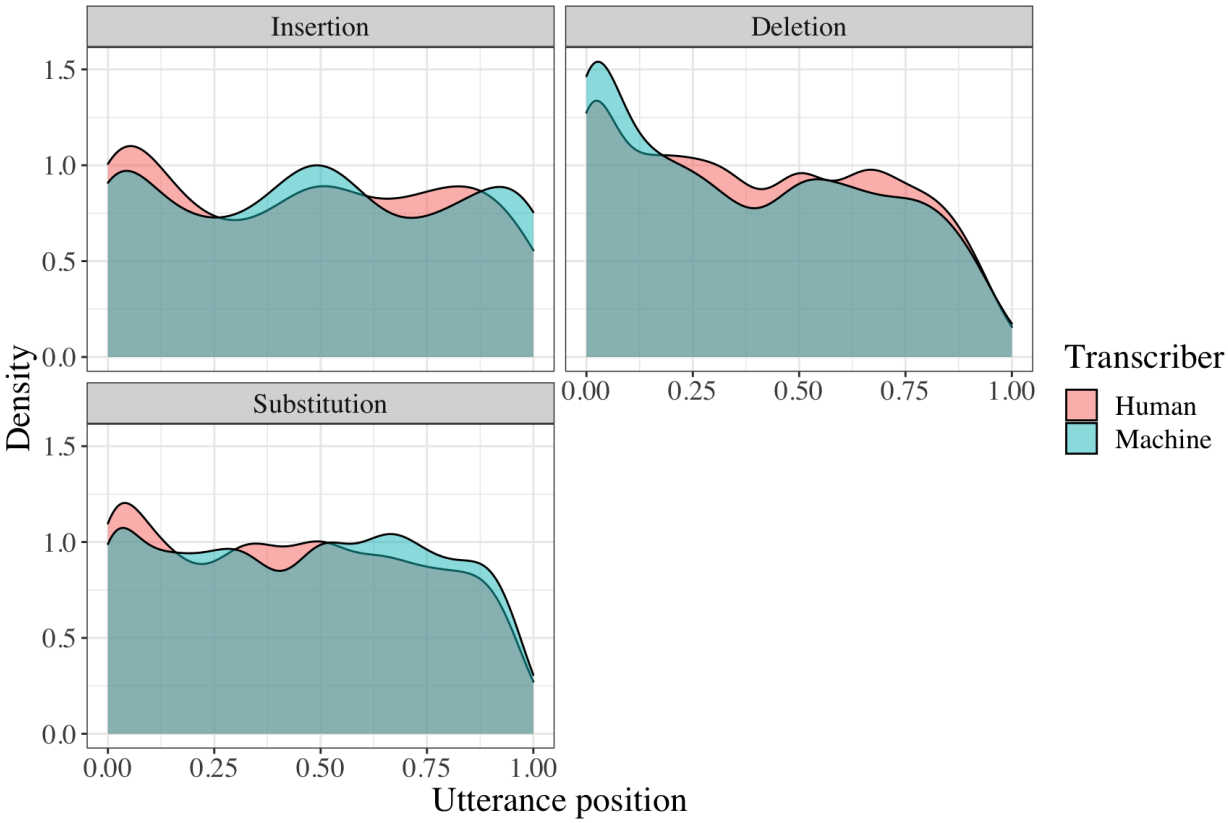


Figure 6.10: Density plot of the utterance position of the error.

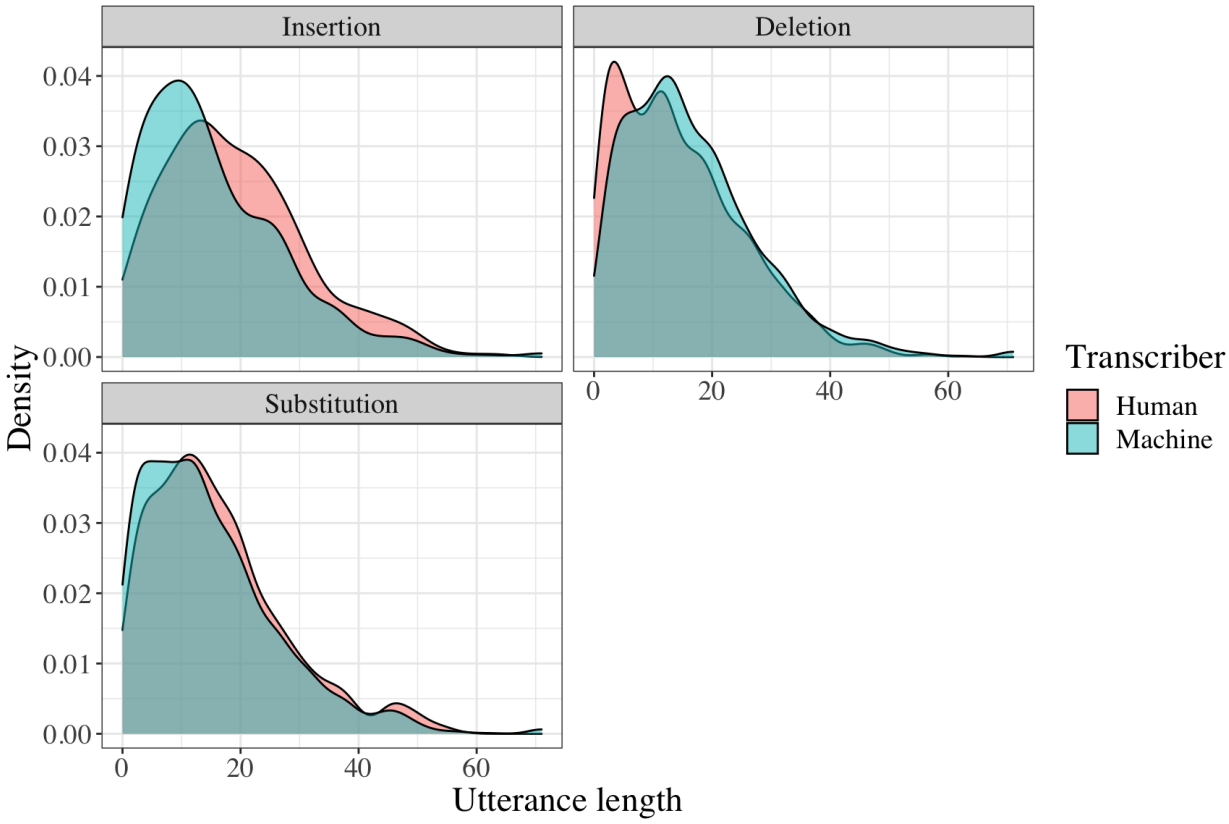


Figure 6.11: Density plot for the utterance length at each error.

Chapter 7

CONCLUSION

This dissertation has considered misperception through an examination of errors in human and ASR transcription. It has provided evidence for a relationship between misperception, predictability, and other lexical factors. The following chapter summarizes these findings and provides a brief discussion about implications and future work.

7.1 Misperception and predictability

One of the main aims of this dissertation was to consider the relationship between predictability and misperception. Previous laboratory studies showed that more predictable tokens were easier to recognize (Howes, 1957; Broadbent, 1967). However, the Smooth Signal Hypothesis (Aylett and Turk, 2004) and work on frequency in naturalistic data (Vitevitch, 2002; Zayats et al., 2019) suggest that more predictable words may be harder to recognize. Prosodic prominence would be a likely intermediary of this effect. The results found here predominantly align with the latter view. Chapter 4 and 6 showed that high-predictability function and discourse-related words had high error counts, and that these words were more likely to be involved in misperception as a proportion of their frequency in the corpus. In the crowd-sourcing study in Chapter 5, function and discourse word alternatives were more likely to result in misperception by the listener.

When transcribers made errors, their hypotheses were also found to be on average more predictable than the reference. This relationship was the strongest for hallucinations and misses. The crowd-sourcing study further supported these findings. Predictability was found to be significant when presenting two alternatives regardless of whether they were previously mistranscribed. Simply generating alternative utterances that were more predictable than an

actual utterance was more likely to lead to confusion. A look at the transcription errors shows many cases of misperception as grammatical 'correction'. These errors echo the shadowing errors found in (Lackner, 1980). A few of these errors are presented below.[1]

(21) where **do / {}** you go for steaks

(22) that's **{} / for** usually pretty good

(23) that's **a / that's** good point

These examples and the findings regarding predictability support the idea of graceful degradation. In order to efficiently process speech, people will 'edit' incomplete, erroneous, or obscured input with a statistically likely alternative. Speech perception studies tend to focus on single-word substitutions to inform ideas about lexical choice. Theories must also account for graceful degradation and the 'editing' process which can result in the hallucination of speech tokens.

To explore predictability in misperception, the metric of surprisal difference was introduced. The surprisal difference can be used to compare a hypothesis and reference sequence and determine how predictable the hypothesis is with the reference as baseline. This metric provided insights into human perception, and, as will be discussed later, was also useful for addressing ASR errors.

7.2 Transcription alignments and misperception

This work took statistics over a great number of misperceptions by repurposing datasets which are collected for speech processing applications, using transcriptions of telephone speech. In fact, the Switchboard corpus is used in training the ASR systems discussed in this dissertation (Xiong et al., 2016; Saon et al., 2017). There are reasons why a particular task such as CTS transcription may be noisy or biased when used as a tool to study misperception. For instance, quick transcription can lead to typos, which reflect errors of production by the listener.

[1]Where the transcription differs from the reference, the hypothesis is indicated at the left of the slash.

Chapter 5 produced a replication of the errors under different task conditions. In this case, a forced-choice sentence listening task was used. The model showed that previously mistranscribed utterances were 13.25 times more likely to be classified as low accuracy compared to baseline utterances. Factors related to predictability and lexical class were secondary predictors. Furthermore, the mistranscription errors were high in disagreement, showing that such errors relate to areas of ambiguity, as claimed in (Glenn et al., 2010). Based on these results, there were no obvious task-specific differences between transcription and the sentence-choice task. This supports the idea that transcriptions are an ecologically valid way to study misperception. The dataset at hand provides 31.6K errors, which is larger than any currently available corpus of misperception.[2]

7.3 Recognition in humans and machines

A major contribution of this thesis is its consideration of linguistic features in ASR errors using human transcription errors as a baseline. Previous work has been limited in this area. Goldwater et al. (2010) provides a thorough analysis but without a human performance baseline. Stolcke and Droppo (2017) compare human and machine performance but use a limited number of measures, including a correlation of speakers and a list of the most frequent error tokens.

The results of our analysis show that human and ASR errors, while having a very similar WER, have significant differences in terms of error features. First of all, there are a significantly larger number of insertion and substitution errors made by ASR, and a greater number of deletions made by human transcribers. Function words are most often deleted by human transcribers, although they are also the most frequent tokens in the corpus. Discourse words are in fact the most difficult words for transcribers. They have the highest error rate and are missed nearly 10% of the time. Human transcribers are especially likely to confuse the filled pauses ‘uh’ and ‘um’. They are confused 129 times in the human transcripts, vs. just 38 times

[2]This corpus is available at https://github.com/vickyzayats/switchboard_corrected_reannotated

in the ASR transcripts. On the other hand, ASR had a very difficult time disambiguating hesitations ('mm') and backchannels ('mm-hm'). Humans may be better able to identify these words due to their different functions in the discourse.

Although ASR was better able to recall words, there were a greater number of insertions and substitutions produced by ASR. The high number of insertions was due to combinations of insertion-substitution errors (where one token in the reference is replaced by multiple tokens). In other words, ASR was more likely to confuse words in the reference, and was therefore less precise. Precision issues were pronounced with regards to OOV words from the Fisher language model. OOV tokens were associated with differences between the distribution of error frequencies and predictabilities. There was a greater density of OOV reference substitutes in ASR.

The surprisal difference patterns for humans generally matched those of the previous studies. Insertions and deletions were highly predictable, and subsitution was about even. ASR insertions were more surprising, likely accounted for by those insertion-substitution predictions that often accompanied OOV reference words. Single token substitutions in ASR were generally less surprising compared to human transcription. This may support the idea that ASR relies more heavily on context in decoding.

Additional findings demonstrate that utterance position and length are different between human and machine recognition errors. While ASR and human transcribers may produce similar numbers of errors, there are still significant differences in recognition performance regarding particular features.

7.4 Applications and future work

This work raises several important future considerations. First of all, there was significant support for the relationship between predictability and misperception. However, these and other studies have touched on a relationship between prosodic prominance, frequency, and predictability. Laboratory studies which carefully control for each of these factors would help to better gauge the contribution of each to (mis)perception. These results and work from Bell

et al. (2009) support the idea that both frequency and predictability are important measures when modeling perception.

This dissertation also used surprisal and recurrent language models to measure predictability. Given the availability of many different types of language models with different configurations, architectures, and features (syntactic and lexical), which model best predicts misperception is an open question. There has been some related work considering surprisal models for predicting reading time (Frank and Bod, 2011). In the future, comparing both forward, backwards, and bi-directional models might help to consider the perennial question of whether listeners rely most on previous or following context during recognition.

One core finding of this work is that despite similarities in WER there can be measurable differences in human and machine recognition. Reaching human parity with regards to WER does not mean the work of ASR is done, and how researchers represent human parity to the public should be carefully considered. This is not to say that ASR cannot perform at the level of humans; in fact, this work has shown that machine recognition has surpassed human transcription capabilities in some domains. However, ASR is clearly disadvantaged in other areas.

One of the most challenging problems which is reflected in the current analysis is the problem of low-probability and out-of-vocabulary words. Tackling this problem intentionally and methodologically should be a top concern for ASR researchers. There is some promising work being done in this area already, such as considering acoustic-to-grapheme models or sub-word units to supplement acoustic-to-word approaches (Thomas et al., 2019; Inaguma et al., 2018). Speech processing in languages other than English will benefit from OOV solutions, given the much smaller datasets available for language model training. Overall, a more systematic evaluation of ASR performance can boost development by honing in on issues with the greatest impact. Given the apparent differences between human and machine recognition that were uncovered in this work, alternative WER measures like precision and recall (McCowan et al., 2004), or measures that rely on salience weighting (Mishra et al., 2011) may better direct ASR development.

7.5 Concluding remarks

This dissertation made three primary contributions to the literature. It established mistranscription as a useful benchmark for human speech misperception. It built a strong case for a relationship between predictability and speech perception using the novel metric of surprisal difference. Finally, it analyzed distributional differences in human and machine recognition errors using predictability and other lexical features. The findings of this work shed light on the unique strengths and limitations of speech recognition by humans and machines.

BIBLIOGRAPHY

Arnold, J. E., Fagnano, M., and Tanenhaus, M. K. (2003). Disfluencies signal theee, um, new information. *Journal of psycholinguistic research*, 32(1):25–36.

Aylett, M. and Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and speech*, 47(1):31–56.

Bard, E. G., Shillcock, R. C., and Altmann, G. T. (1988). The recognition of words after their acoustic offsets in spontaneous speech: Effects of subsequent context. *Perception & Psychophysics*, 44(5):395–408.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., and Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, 60(1):92–111.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency; Association for Computing Machinery: New York, NY, USA*.

Bond, Z. and Garnes, S. (1980). Misperceptions of fluent speech. *Perception and production of fluent speech*, pages 115–132.

Boston, M. F., Hale, J., Kliegl, R., Patil, U., and Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, 2(1).

Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, pages 1171–1178.

Brennan, S. E. and Schober, M. F. (2001). How listeners compensate for disfluencies in spontaneous speech. *Journal of Memory and Language*, 44(2):274–296.

Broadbent, D. E. (1967). Word-frequency effect and response bias. *Psychological review*, 74(1):1.

Browman, C. (1980). Perceptual processing: Evidence from slips of the ear. *Errors in linguistic performance*.

Calhoun, S., Carletta, J., Brenier, J. M., Mayo, N., Jurafsky, D., Steedman, M., and Beaver, D. (2010). The NXT-format Switchboard Corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language resources and evaluation*, 44(4):387–419.

Canavan, A., Graff, D., and Zipperlen, G. (1997). Callhome American English speech. *Linguistic Data Consortium*.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*.

Cieri, C., Miller, D., and Walker, K. (2004). The fisher corpus: a resource for the next generations of speech-to-text. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Clopper, C. G., Pisoni, D. B., and Tierney, A. T. (2006). Effects of open-set and closed-set task demands on spoken word recognition. *Journal of the American Academy of Audiology*, 17(5):331–349.

Dahan, D., Magnuson, J. S., and Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive psychology*, 42(4):317–367.

Deshmukh, N., Ganapathiraju, A., Gleeson, A., Hamaker, J., and Picone, J. (1998). Resegmentation of Switchboard. In *Fifth International Conference on Spoken Language Processing*.

Dubno, J. R., Dirks, D. D., and Morgan, D. E. (1984). Effects of age and mild hearing loss on speech recognition in noise. *The Journal of the Acoustical Society of America*, 76(1):87–96.

Felty, A. R., Buchwald, A., Gruenenfelder, T. M., and Pisoni, D. B. (2013). Misperceptions of spoken words: Data from a random sample of American English words. *The Journal of the Acoustical Society of America*, 134(1):572–585.

Fiscus, J., Fisher, W. M., Martin, A. F., Przybocki, M. A., and Pallett, D. S. (2000). 2000 NIST evaluation of conversational speech recognition over the telephone: English and Mandarin performance results. In *Proc. NIST Speech Transcription Workshop*, pages 1–5.

Fox Tree, J. E. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of memory and language*, 34(6):709–738.

Frank, S. L. and Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological science*, 22(6):829–834.

Frank, S. L., Otten, L. J., Galli, G., and Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and language*, 140:1–11.

Garnes, S. and Bond, Z. S. (1980). A slip of the ear: A snip of the ear? a slip of the year? *Errors in Linguistic Performance*.

Glanzer, M. (1972). Storage mechanisms in recall. In *Psychology of Learning and Motivation*, volume 5, pages 129–193. Elsevier.

Glenn, M. L., Strassel, S. M., Lee, H., Maeda, K., Zakhary, R., and Li, X. (2010). Transcription methods for consistency, volume and efficiency. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520. IEEE.

Goldwater, S., Jurafsky, D., and Manning, C. D. (2010). Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, 52(3):181–200.

Goodkind, A. and Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18.

Greenberg, S. and Chang, S. (2000). Linguistic dissection of Switchboard-corpus automatic speech recognition systems. In *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*.

Grezause, E. L. (2017). *Um and Uh, and the Expression of Stance in Conversational Speech*. PhD thesis, University of Washington.

Grosjean, F. (1985). The recognition of words after their acoustic offset: Evidence and implications. *Perception & Psychophysics*, 38(4):299–310.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics.

Hamaker, J., Deshmukh, N., Ganapathiraju, A., and Picone, J. (1998). Resegmentation and transcription of Switchboard. In *Proceedings of LVCSR Workshop*.

Howes, D. (1957). On the relation between the intelligibility and frequency of occurrence of English words. *The Journal of the Acoustical Society of America*, 29(2):296–305.

Inaguma, H., Mimura, M., Sakai, S., and Kawahara, T. (2018). Improving OOV detection and resolution with external language models in acoustic-to-word ASR. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 212–218. IEEE.

Kalikow, D. N., Stevens, K. N., and Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America*, 61(5):1337–1351.

Kingsbury, P., Strassel, S., McLemore, C., and McIntyre, R. (1997). CALLHOME American English Transcripts. *University of Pennsylvania: Linguistic Data Consortium*.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13):1–26.

Lackner, J. R. (1980). Speech production: correction of semantic and grammatical errors during speech shadowing. *Fromkin, 1980b*, pages 149–163.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Liu, D. and Zhang, H. (2018). Residuals and diagnostics for ordinal regression models: A surrogate approach. *Journal of the American Statistical Association*, 113(522):845–854.

Luce, P. A. and Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, 19(1):1.

Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19.

Marslen-Wilson, W. and Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, 8(1):1–71.

Marslen-Wilson, W. D. and Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, 10(1):29–63.

McClelland, J. L. and Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18(1):1–86.

McCowan, I. A., Moore, D., Dines, J., Gatica-Perez, D., Flynn, M., Wellner, P., and Bourlard, H. (2004). On the use of information retrieval measures for speech recognition evaluation. Technical report, IDIAP.

Michaelov, J. and Bergen, B. (2020). How well does surprisal explain N400 amplitude under different experimental conditions? In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 652–663, Online. Association for Computational Linguistics.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2):81.

Miller, G. A., Heise, G. A., and Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology*, 41(5):329.

Miller, G. A. and Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America*, 27(2):338–352.

Miller, G. A. and Selfridge, J. A. (1950). Verbal context and the recall of meaningful material. *The American Journal of Psychology*, 63(2):176–185.

Mishra, T., Ljolje, A., and Gilbert, M. (2011). Predicting human perceived accuracy of ASR systems. In *Twelfth Annual Conference of the International Speech Communication Association*.

Monsalve, I. F., Frank, S. L., and Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 398–408. Association for Computational Linguistics.

Morris, A. C., Maier, V., and Green, P. (2004). From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In *Eighth International Conference on Spoken Language Processing*.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, 76(2):165.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Pham, V., Bluche, T., Kermorvant, C., and Louradour, J. (2014). Dropout improves recurrent neural networks for handwriting recognition. In *14th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 285–290. IEEE.

Pollack, I., Rubenstein, H., and Decker, L. (1960). Analysis of incorrect responses to an unknown message set. *The Journal of the Acoustical Society of America*, 32(4):454–457.

Przybocki, M. and Martin, A. (2001). 2000 NIST Speaker Recognition Evaluation: LDC2001S97.

Roark, B., Bachrach, A., Cardenas, C., and Pallier, C. (2009). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 324–333. Association for Computational Linguistics.

Rubenstein, H. and Pollack, I. (1963). Word predictability and intelligibility. *Journal of Verbal Learning and Verbal Behavior*, 2(2):147–158.

Saon, G., Kurata, G., Sercu, T., Audhkhasi, K., Thomas, S., Dimitriadis, D., Cui, X., Ramabhadran, B., Picheny, M., Lim, L.-L., et al. (2017). English conversational telephone speech recognition by humans and machines. *Proc. Interspeech 2017*.

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika*, 6(5):309–316.

Savin, H. B. (1963). Word-frequency effect and errors in the perception of speech. *The Journal of the Acoustical Society of America*, 35(2):200–206.

Segalowitz, S. J. and Lane, K. C. (2000). Lexical access of function versus content words. *Brain and language*, 75(3):376–389.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Shriberg, E. E. (1994). *Preliminaries to a theory of speech disfluencies*. PhD thesis, University of California at Berkeley.

Solomon, R. L. and Postman, L. (1952). Frequency of usage as a determinant of recognition thresholds for words. *Journal of Experimental Psychology*, 43(3):195.

Stolcke, A. (2002). SRILM-an extensible language modeling toolkit. In *Seventh international conference on spoken Janguage processing*.

Stolcke, A. and Droppo, J. (2017). Comparing human and machine errors in conversational speech transcription. In *Proc. Interspeech*, pages 137–141. ISCA - International Speech Communication Association.

Taylor, A., Marcus, M., and Santorini, B. (2003). The Penn treebank: an overview. In *Treebanks*, pages 5–22. Springer.

Thomas, S., Audhkhasi, K., Tüske, Z., Huang, Y., and Picheny, M. (2019). Detection and recovery of OOVs for improved English broadcast news captioning. In *INTERSPEECH*, pages 2973–2977.

Vasilescu, I., Yahia, D., Snoeren, N., Adda-Decker, M., and Lamel, L. (2011). Cross-lingual study of ASR errors: on the role of the context in human perception of near-homophones. In *Twelfth Annual Conference of the International Speech Communication Association*.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S-Plus*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Vitevitch, M. S. (2002). Naturalistic and experimental analyses of word frequency and neighborhood density effects in slips of the ear. *Language and speech*, 45(4):407–434.

Wang, M. D. and Bilger, R. C. (1973). Consonant confusions in noise: A study of perceptual features. *The Journal of the Acoustical Society of America*, 54(5):1248–1266.

Wang, Y.-Y., Acero, A., and Chelba, C. (2003). Is word error rate a good indicator for spoken language understanding accuracy. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721)*, pages 577–582. IEEE.

Wicha, N. Y., Bates, E. A., Moreno, E. M., and Kutas, M. (2003). Potato not pope: human brain potentials to gender expectation and agreement in Spanish spoken sentences. *Neuroscience Letters*, 346(3):165–168.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., and Zweig, G. (2016). Achieving human parity in conversational speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M. L., Stolcke, A., Yu, D., and Zweig, G. (2017). Toward human parity in conversational speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12):2410–2423.

Yang, J. and Zhang, Y. (2018). NCRF++: An open-source neural sequence labeling toolkit. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Zayats, V., Tran, T., Wright, R., Mansfield, C., and Ostendorf, M. (2019). Disfluencies and human speech transcription errors. In *Proc. Interspeech 2019*, pages 3088–3092.

Appendix A

SUPPLEMENTAL STOP WORDS LIST

a	been	had	noone	sometime
about	being	hadn't	not	sometimes
again	did	has	nothing	somewhat
almost	do	have	now	somewhere
already	does	haven't	nowhere	such
also	don't	having	of	then
always	done	however	often	there
am	enough	in	only	together
anybody	even	instead	perhaps	too
anyhow	ever	is	quite	very
anyone	everybody	just	rather	wanna
anything	everyone	least	really	was
anytime	everything	less	several	were
anyway	everywhere	many	so	yet
anywhere	few	more	somebody	
are	fewer	most	somehow	
as	got	much	someone	
be	gotten	nobody	something	

Appendix B

SWITCHBOARD - MOST FREQUENT WORDS

a	got	me	so	um
about	guess	mean	some	up
all	had	more	something	very
an	have	much	that	was
and	he	my	that's	we
are	here	no	the	well
as	how	not	them	were
at	i	now	then	what
be	i'm	of	there	when
because	i've	oh	there's	where
been	if	okay	they	with
but	in	on	they're	would
can	is	one	thing	yeah
do	it	or	things	you
don't	it's	our	think	you_know
for	just	out	this	your
from	kind	people	time	
get	know	really	to	
go	like	right	too	
going	little	see	uh	
good	lot	she	uh-huh	