The Distribution of Disfluencies in Spontaneous Speech: Empirical Observations and Theoretical Implications

Advisor: Mark Liberman

Proposal Committee: Jianjing Kuang (Chair), Kathryn Schuler, Gareth Roberts

Hong Zhang

May 9, 2019

1 Introduction

Speech disfluencies, or hesitation phenomena such as silent or filled pauses, false starts, repetitions and repairs, are prevalent in spontaneous speech. Even normally perceived fluent speakers can have a disfluency rate somewhere between 6 and 10 disfluent words per 100 words (F. Ferreira & Bailey, 2004; Shriberg & Stolcke, 1996; Rochester, 1973). The occurrence of disfluencies in spontaneous speech is in fact not random e.g., (Clark & Tree, 2002; Shriberg, 1994; Holmes, 1988). Sentence length, complexity, lexical context, as well as prosodic phrasing have all been shown to correlate with the likelihood of observing some forms of disfluencies in natural speech (Goldman-Eisler, 1958; Tannenbaum, Williams, & Hillier, 1965; Beattie & Butterworth, 1979; Bell et al., 2003; Nakatani & Hirschberg, 1994; Lickley, 2015). On the other hand, hesitation markers can also be in the expressive armory of a speaker to signal a delay or mark the discourse structure of one's speech (Swerts, 1998; Clark & Tree, 2002). Evidence from research in the past decades has highlighted the needs of a thorough understanding of disfluencies in spontaneous speech to benefit areas that require knowledge on spontaneous speech, such as in psycholinguistic and clinical investigations of language production, sociolinguitic research in language variation and change, as well as various aspects in speech technology.

Models of speech production often cite evidence from disfluencies, because the correspondence between symptoms of disfluent speech and various linguistic variables can be used to infer breakdowns during the production process (Levelt, 1983; Holmes, 1988; V. S. Ferreira & Pashler, 2002). Since early observations from samples of spontaneous speech (Maclay & Osgood, 1959; Goldman-Eisler, 1958), studies have found evidence in support of the idea that speech production is a hierarchical process (Levelt, 1983, 1989). In these models, language production is achieved through passing down an abstract idea through several stages involving syntactic planning, lexical selection and access, as well as phonological planning and motor control for articulation, to the physical product of acoustic signals. In this picture, disfluenies inevitably happen at all the stages involved in the production process. To understand the cognitive mechanism behind speech production is thus among the key motivations in disfluency research.

Disfluent speech is not only resulted from breakdowns in the process of speech production driven by speaker's cognitive ability. Evidence has suggested that contextual variables, such as

the topic of university lectures (Schachter, Christenfeld, Ravina, & Bilous, 1991; Moniz, Batista, Mata, & Trancoso, 2014), familiarity with interlocutor (Bortfeld, Leon, Bloom, Schober, & Brennan, 2001), speaking style (Moniz et al., 2014), the nature of the task (human-human vs human-machine communication) (Shriberg, 1994) as well as the processing load of the linguistic context (Bortfeld et al., 2001; Arnold, Kam, & Tanenhaus, 2007), affect the rate and form of disfluent speech. This body of literature reports that higher disfluency rate is associated with situations with more demanding context, such as talking to a stranger, performing a harder task, or talking about more domain-specific jargon. However, some unfamiliar situations, such as during human-machine interaction, may have an effect in the opposite direction (Levelt, 1983; Blacfkmer & Mitton, 1991; Lickley, 1994). This difference between human-human conversation and human-machine interaction suggests that disfluencies can be affected by the perceived need of specific speaking task (Broen & Siegel, 1972). Results from these studies indicate that disfluencies are more than a mere by-product of performance deficit in response to contextual variation. Speakers can actively control the production of disfluencies through more careful planning based on factors involved in the speaking task itself. For instance, certain sociolinguistic variables can potentially be among these factors. Recent studies looking at factors such as age, gender and English varieties, have shown that these variables do systematically correlate with the use and frequency distribution of filled pauses (Fruehwald, 2016; Tottie, 2011, 2014). The distributional difference across age and gender has also been interpreted as a change in progress led by females (Wieling et al., 2016).

Disfluencies as both a device for message structuring and a symptom of breakdowns in the production process has broader implications beyond the interests of linguists and cognitive scientists interested in speech production and perception. Variations in the form and location of disfluent utterances can inform us about the potential cause of deficiencies in one's linguistic ability (Grossman & Ash, 2004). For example, variants of Frontotemporal Degeneration (FTD), a family of neural degenerative diseases with known effect of affecting human linguistic ability, can be characterized and distinguished in part by the surface language deficiencies. Effortful speech is a symptom for the non-fluent agrammatic variant of primary progressive aphasia (na-PPA) (Ash et al., 2010); patients with the semantic variant (svPPA) are often diagnosed with difficulties in lexical access (Ash et al., 2009; Mack et al., 2015). However, in addition to these apparent impairment with specific linguistic abilities, it has also been shown (Nevler et al., 2017) that temporal and prosodic features of such clinical speech are also crucial in distinguishing patients from healthy controls, and between different phenotypes. Using linguistic information for the diagnosis of neural degeneration is an emerging field where the role of disfluent speech has already been highlighted. However, the understanding of the relation between the disruption in speech production and the underlying functional impairment is rather limited (Boschi et al., 2017).

The presence of disfluencies in natural speech poses great challenge in human language techonology. One direct benefit from a robust understanding of the distribution properties of disfluencies is in fact to facilitate systems dealing with natural speech to accurately identify and eliminate the adversarial effects of the presence of disfluencies. Therefore substantial amount of effort has been made in automatic detection and removal of disfluencies from spontaneous speech to improve performance of both ASR and TTS systems (Liu et al., 2006; Shriberg & Stolcke, 1996; Qian & Liu, 2013; Siu & Ostendorf, 1996; Hough, 2014; Ostendorf & Hahn, 2013; Nakatani & Hirschberg, 1994). These studies both provided detailed pattern description of various disfluency phenomena with the goal of contributing to practical application (Stolcke & Shriberg, 1996; Siu & Ostendorf, 1996; Plauché & Shriberg, 1999), and developed statistical methods for the task of identification

and removal of disfluencies (Liu et al., 2006; Hough, 2014; Qian & Liu, 2013). Goldwater, Jurafsky, and Manning (2010) explicitly addressed the question of how disfluencies are related to the errors made by ASR systems. They suggested that repetition tokens, word fragments, as well as acoustically or prosodically indistinguishable disfluent segments are associated with higher error rate. Their results highlighted the need to fully explore the feature space of speaker variation to account for the observed error pattern. Although unlike early automatic speech recognition systems which were mainly trained on read speech or otherwise constrained speech format (Butzberger, Murveit, Shriberg, & Price, 1992), the acoustic or language models trained on normally produced speech may still face the problem of generalizing across different domains, such as tagging twitter, blog post and spontaneous speech (Foster, 2010). It can be more challenging in the low resource domain where suitable data is not only sparse or expensive to acquire, but also presents large amount of deviation from the standard language, such as in the setting of clinical interviews. Thus detailed understanding of distributional properties of disfluencies across domains is still necessary for overcoming the constraints in modern speech technology.

Speech disfluencies are also an integral part in evaluating and improving dialogue systems that involve human-machine conversations. The difference among the three corpora used in Shriberg (1994, 2001) demonstrated that fewer disfluencies should be expected in human-machine communication in travel planning domain. In a human-robot communication scenario, Skantze, Hjalmarsson, and Oertel (2013) show that silence and filled pauses can inhibit user activity in a map task, realized as changes in user behavior in drawing. Modeling user disfluencies has also been shown to improve the engagement of human-robot conversations and the management of the flow of dialogue. Bohus and Horvitz (2014) proposed a forcasting and hesitation mechanism that leverages human disfluency information to predict user engagement, and generate proper response to facilitate a more fluid conversation. Skantze and Hjalmarsson (2010) showed that a dialogue system that incrementally incorporates filled pause and self-repairs can achieve shorter response time and generate more naturally perceive speech, even though the generated utterances tend to be longer.

Given the large volume of work in speech disfluency from fields ranging from linguistics, phycholinguistics, sociolinguistics to natural language processing and language generation, there is still the need to further elaborate how the multivariate feature space jointly defines the distribution of disfluencies. On the one hand, pieces of information have been provided from researchers focusing on questions concerning primarily the interests within one's own field. However, due to differences in research methods, including not only experiment design, but also the classification and annotation of disfluent speech segments, cross-domain generalization of these results can be challenging. On the other, both the design issue and availability of suitable data and computing resource also limit the analyses performed on certain forms of disfluencies. In this dissertation, I will attempt to explore further into this joint feature space by addressing questions from the following perspectives: the unexplored covariates of disfluency variation, the under studied forms of disfluencies, and the overall lack of understanding beyond Germanic languages.

A deeper understanding of this dynamic and complex feature space behind disfluencies is crucial for both practical and theoretical reasons. Practically, applications involving natural human speech have the need to accommodate the presence of disfluent speech or utilize the information contained in it. For instance, inserting filled pauses in synthesized speech has been shown to improve the perceived naturalness of the system (Adell, Escudero, & Bonafonte, 2012). Information contained in disfluent interview response can be used for disease diagnosis. On the other hand, theoretically speaking, such an understanding will not only help resolve, or dismiss, the dispute

over whether disfluencies, or hesitation markers more specifically, should be considered part of human linguistic apparatus which, at least partially, convey lexical meanings (Clark & Tree, 2002) or more of a by-product when one is trying to maintain fluency (Lickley, 2015), but also inform us the dynamic role that disfluencies play in structuring the content of speech and buffer outside disruptions.

1.1 Research questions

The primary deficit in previous research is the imbalanced attention received by different forms of disfluencies with regard to their natural distribution. Although hesitation phenomena involve a wide range of and interrelated types of disfluencies or speech repairs, more focus has been placed on silent and filled pauses than repetition and repair. This is especially true when it comes to large scale corpus studies. One direct consequence is that the knowledge on the distribution property with respect to the immediate linguistic context of more complex hesitation phenomena is rather limited, compared to what have been established for filled pauses. On the other hand, among the studies dedicated to repetitions and other repair phenomena, less attention has been paid to explore factors that beyond the lexical or syntactic environment of repairs. With the observations in, for example, Shriberg (1994, 2001), it should be argued that individual variation, and the sociolinguistic factors behind, should be more systematically examined to fully account for the variations in repetitions and repair. Therefore, one question that I would like to raise and address in this study is how the variation in repetitions and repairs can be characterized and explained.

With regard to the feature space involved in speech disfluency, it is less understood how elements in conversations, such as features of the interlocutor, conversation topics, etc., contribute to variation in the surface variation. Regarding certain hesitation phenomena as sociolinguisitc marker, such as filled pauses, has received more attention in the past two decades. More recent studies such as Wieling et al. (2016) have demonstrated an interesting gender distinction in the choice of fillers which can be attributed a trend explained by a change in progress. Less understood in this domain is how the topic of conversations, as well as the interlocutors, affect variations in disfluencies, and how such meta-linguistic information interact with the immediate linguistic contexts in which disfluencies are realized. Thus the other question that I would like to address in this study is what's the role of the previously under-studied discourse and sociolinguistic variables, including conversation topic and interlocutor accommodation, in the realization of disfluencies, and further explore how these variables can be jointly understood with the previously established sociolinguistic and linguistic variables.

The last, but not the least, problem that I would like to raise is the lack of cross-linguistic comparison of the disfluency phenomena in the field. Looking back at the literature, our understanding of disfluent speech is predominantly based on studies in English, with a handful papers concerning German, French, Portuguese, Hebrew, Mandarin and Japanese (Fox, Maschler, & Uhmann, 2010; Fox, Hayashi, & Jasperson, 1996). The lack of linguistic diversity constraints researchers from discovering and exploring the disfluency phenomena that bear language-specific characteristics. For example, cross-linguistic studies on repetitions have generally acknowledged that function words, especially those immediately preceding a content word in a constituents, are more frequently repeated than other word classes (Fox et al., 1996, 2010; Clark & Wasow, 1998). However, it is not clear what would happen to a language that relies predominantly on morphological devices to realize agreements and indicate spatial or temporal relations. On the methodlogy side of the

problem, attention to the cross-linguistic aspects of disfluencies, or self-repairs more specifically, is mostly from conversation analysis, such as in the form following Schegloff, Jefferson, and Sacks (1977). Their focus on the limited set of examples is constrained by the generalizeability of their results to account for broader range of variations. Like the other aspects in the study of disfluencies, implications of this cross-linguistic perspective can be drawn across multiple disciplines related to naturally produced human speech. For instance, from the point of view of language production and psycholinguistics, careful investigations of disfluency patterns in languages with more structural diversity can help researchers to refurbish the production models, and understand the potential language-specific constraints on human language production. Thereofore, in this study, I will tap into this question by looking at repetitions in Czech. It is hoped that this effort can lead to more fruitful research on this or related topic.

The three questions, as seen from the description above, are highly entangled. Underlying these questions is the need to enrich our current understanding of disfluencies comprehensively from all disfluency types and considering simultaneously both linguistic as well as discourse and sociolinguistic variables. To acknowledge this need, I will address the proposed questions from two broad perspectives: from a macro-analysis perspective, where sociolinguistic, discourse and other contextual factors will be considered, and from a micro-analysis perspective, where the focus will be on the immediate linguistic contexts in which disfluency phenomena are observed. Interpretation of the results will be based at the joint of information obtained from the two broad arbitrary perspectives. Thus distinction is made mainly for facilitating the structuring of arguments without further theoretical implications. The goal is to provide a systematic description of the fact from a representative sample of speech, basing on which interpretations can be made to suit domain dependent interests.

1.2 Methods

Research in speech disfluency has been conducted through both quantitative observations of speech corpora and careful analyses of speech produced in controlled lab experiments. Both research methodologies have their unique advantages over the other, while facing particular challenges of its own.

Corpora of collections of spontaneous speech have been a primary source for disfluency research. The apparent reason in favor of using corpora of spontaneous speech is that speech disfluency primarily occurs in spontaneous and under-prepared speech. Given its relatively rare occurrence (as only 6% to 10% as the generally acknowledged disfluency rate) and great potential for individual variation, the amount of speech that is required to capture enough variance tends to be large. Thus corpus based research caters well with these demands. The widely used speech corpora include the Switchboard (Godfrey, Holliman, & McDaniel, 1992), ATIS (Dahl et al., 1994), and AMEX corpus (Kowtko & Price, 1989). These corpora represent a wide range of scenario where communication tasks tale place. For example, Switchboard consists of unguided spontaneous telephone conversations between two speakers under a provided topic, while ATIS represents human and machine oriented speech in the scenario of travel planning. Studies based on such wide representations of speech data (e.g., Shriberg (1994, 2001); Shriberg and Stolcke (1996)) have been fruitful in identifying linguistic and contextual or discourse variables that correlate with surface variations of speech disfluency.

However, corpus-based studies of speech disfluency are often faced with two major challenges:

the lack of properly annotated data and insufficient control for the environment in which the speech was produced. The first challenge is mainly due to the lack of awareness of creating properly annotated disfluencies during data collection. However, with disfluency information included in the transcription, significant performance gain has been reported in automatic part of speech tagging M. Johnson and Charniak (2004). Same applies when disfluent speech segments are removed (Kahn, Lease, Charniak, Johnson, & Ostendorf, 2005). The second challenge, though is of lesser concern, poses questions to the interpretation of the observed disfluency patterns. As it is often the case that corpora of spontaneous speech are comprised of conversations relatively freely conducted by task participants, causal interpretations between the linguistic variables and disfluency events are even harder to establish (Schnadt & Corley, 2006).

Experimental work has also been conducted to explore the nature of speech disfluency (Zvonik & Cummins, 2003; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; Arnold et al., 2007; Bortfeld et al., 2001; Schnadt & Corley, 2006). A typical production experiment is set up in a way that participants are expected to produce speech guided by certain speaking tasks. One research strategy in looking at silent pause in read speech is to look at synchronous speech (Zvonik & Cummins, 2003; Krivokapić, 2007), where speakers are asked to read some text either along or in synchrony with a partner. This method is meant to control for individual variation of text reading in an experiment setting. Some other research engages participants in tasks whose completion requires communication with a partner (Schnadt & Corley, 2006; Bortfeld et al., 2001; Arnold et al., 2007). For example, Schnadt and Corley (2006) adopted a network task developed from Levelt (1983) and Oomen and Postma (2001), where the task is based on describing a network structure. Bortfeld et al. (2001) asked pairs of speakers to describe and match sets of objects, where the processing load was controlled for by manipulating the familiarity of objects, familiarity between two speakers, and so on.

Language production tasks are widely used as tools for neural degeneration diagnosis. For example, the Boston Naming Tast (Goodglass, Kaplan, & Barresi, 2000) and picture description tasks such as the Cookie Theft picture. However, the goal of these tasks is mainly as prompts to elicitate spontaneous speech within a more constraint environment without assuming factorial conditions. Therefore they are fundamentally of a different nature than the experimental methods listed above.

Although speech produced in more controlled lab settings could provide many desirable properties for variable extraction and causal interpretation, it faces the problem of higher limitation on the amount of accessible data from experiments and lack of direct connection between lab conditions and real-world situations. This limitation can be illustrated through one heavily studied question: what is the effect of lexical frequency and sentence complexity on disfluency production. In a well-controlled experiment, several studies (Tannenbaum et al., 1965; Beattie & Butterworth, 1979; Jescheniak & Levelt, 1994) find significant effect of lexical frequency or context predictability on the fluency of speech or shorter response time in object naming. However, when lexical frequency is defined as the extent to which people agree on the name of particular nouns, strong effect of this measure of familiarity is also observed (Hartsuiker & Notebaert, 2009). Then the question of frequency effect on fluency becomes whether it is the speaker experience or global frequency of words that affect the fluency of speech. In another study (Tsiamtsiouris & Cairns, 2013), sentence length and structural complexity has been shown to interact with the disfluency of produced speech through an experiment of repeating sentences of 6 and 30 words long. Unfortunately,this study failed to discuss other covariates that might as well explain the observed group

difference, or how much variation in disfluency can still be explained by their controlled variables when other unobserved effects are present.

In addition, the use of lab speech doesn't necessarily solve the problem of proper annotation. While it can be claimed that careful transcriptions of disfluency events are made possible as the speech data is under the full control of the researcher, the transcriptions are often compromised with the size of speech material and hurdles for cross-lab generalization. Since most speech does not happen in well controlled environment as responses to various production tasks, practical implications from lab studies might be limited.

1.2.1 Research methods for the current study

Comparing the two streams of research methodology, corpus-based analyses is still preferred over fine-controlled lab speech, for at least three reasons. First, the two major challenges faced with corpus data, namely the lack of known control and inadequate annotation, can be mitigated in feasible manners. The concern for the lack of causal interpretation can be circumvented if the right data is used in the study, such as through consistently annotating disfluent speech, proper sampling method, and using large quantity of balanced speech data. The issue of disfluency annotation can be solved at least in part through semi-automation. On the one hand, hesitation markers such as filled pauses and full word repetitions can be relatively accurately identified automatically, while applying certain optimization procedure can significantly reduce the time needed to annotate other disfluent phenomena. Second, using publicly available corpus data facilitates collaboration within the field across labs and institutions, and preserves the integrity of research results. Finally, research from the naturally produced speech in realistic communicative settings can more easily be translated to applications that benefit various research communities and beyond. The observed patterns from careful description of naturally produced speech can also be more informative to answers to questions about speech production in psycholinguistic research. With these apparent benefits in mind, corpus-based analyses is adopted in the current study.

1.3 Corpora selection

Three corpora are selected to explore the proposed questions in this study: Fisher (Cieri, Miller, & Walker, 2004), SCOTUS (Yuan & Liberman, 2008) and Czech Spontaneous Speech Corpus (Kolár et al., 2005). In this section, I will elaborate on the reasoning behind this selection.

1.3.1 Fisher

Fisher (Cieri et al., 2004) is a corpus of telephone conversations created in response to the unique needs for automatic speech recognition (ASR) systems. The corpus is built with consideration of controlling a broad range of factors that are essential in representing daily conversational speech. The entire corpus contains 16,454 conversations, totalling 2742 hours of speech. Unlike most of other speech corpora, Fisher balances speakers' age, gender, and represents a wide range of dialectal variation. To encourage the inclusion of large quantities of vocabulary, conversations were guided by 40 topics that are pertinent to both day-to-day life and current pressing social and political issues. The list of topics can be found in the appendix. The collection process also included a platform-driven protocol, with which the data collector initiates calls and matches

between potential participants who expressed interest in the selected topic. This procedure not only maximizes inter-speaker variability, but also reduces the sampling bias to a great extent. Each participant was expected to complete at most three 10-minute conversations. However, the actual number of conversations each participant contributed may vary. Unedited transcriptions are also made available in the corpus.

Given the nature of the Fisher corpus, it is particularly suitable for exploring inter-speaker and inter-contextual variations in speech production. In this study, the subset of Fisher that contains native speakers of American English who completed exactly three conversations are chosen to evaluate these variations. The selected sample contains 9471 one-sided speech from 3157 speakers. The total duration of speech is about 790 hours.

1.3.2 SCOTUS

The second corpus to be used in the present study is SCOTUS Oral Argument Corpus (T. R. Johnson & Goldman, 2009). The full corpus contains 38 years of recordings linked to transcripts of oral arguments at the Supreme Court of the United States (SCOTUS). The subset that contains the speech from 8 US chief justices in 2001 will be used for the present study. This subset was originally compiled for a speaker identification task (Yuan & Liberman, 2008). Verbatim transcriptions of the speech material after diarizatioin are available. This corpus contains about 3 hours of speech from each justice. Unlike Fisher, this corpus provides ample speech material from single speakers, thus it is a good source for more detailed analysis of individual variation.

1.3.3 Czech Broadcast Conversation Speech

Czech Spontaneous Speech Corpus (Kolár et al., 2005) consists of 72 recordings of radio discussion program called Radioforum from 128 speakers. The total speech duration is 33 hours. It broadly falls into the same speech style, spontaneous conversations, as Fisher and SCOTUS. However, noticeable difference between the mode of conversation, such as face-to-face conversation in a more formal setting, unlike that in Fisher, but less formal than SCOTUS. More importantly, the metadata annotation of this corpus follows an adapted version of the MDE annotation developed by LDC. This systam annotates Edit Disfluencies (repetitions, revisions, restarts and complex disfluencies), Fillers (including, e.g., filled pauses and discourse markers) and SUs, or syntactic/semantic units (Consortium, 2009). In the adapted version, detailed syntactic annotation of Czech is also included, which is crucial due to the complexity of Czech syntax. Thus the information provided in this corpus will ensure accurate and consistent analysis of repetitions in Czech, and the results can be comparable to that from the English corpora, when style difference is properly acknowledged.

1.4 Some definitions

Before moving to the actual corpus analyses, some clarification in terminology is necessary. The kind of disfluencies of my primary concern is same-turn self-initiated disfluencies, in contrast to disfluencies that may involve other initiated repairs. As reviewed above, the kinds of analysis in this study are highly dependent upon the amount of information annotated in the selected corpora, as the corpora are made up of different speech styles, collected following different protocol, and intended to serve very different demands. As outlined at the beginning, the Fisher corpus can be

suitable for exploring individual variation in disfluencies as a function of sociolinguistic variables as well as the effect of discourse factor such as conversation topic. The primary disfluency phenomena being addressed with this corpus are silent pauses, filled pauses, and fluent repetitions, given the constraints proposed by the amount of data and available annotated information. On the other hand, SCOTUS is mainly used to answer questions about individual variation in repetition and repair, because it contains ample speech from each of the eight supreme court justices, while maintaining a reasonable amount for proper annotation based on verbatim transcription.

Two clarifications have to be made with regard to the analysis window considering the nature of speech being analyzed. The questions to be answered are first what is the basic unit of analysis and how to define it? And what is the criteria for labeling the disfluency categories? Shriberg (1994) finds her disfluent instances from individual "sentence", where a sentence is defined as a unit that can otherwise be marked with a period or question mark. Her juedgement is somewhat arbitrary and hard to implement with large quantities of data that have not been properly segmented. In the remainder of this section, I will define the window of analysis for the three kinds of speech: telephone conversations, court debate, and radio interviews. The definitions are derived not only from the characteristics of different speaking styles, but more with considerations on the data collection process of different speech corpora. Disfluency phenomena classification and annotation are defined in later type-specific discussions, with specific considerations of the analysis and speech domain variation.

1.4.1 Telephone conversations

Key definitions concerning conversational data in the current set up are turns and utterances. The definition of an utterance can be fuzzy, especially in conversation settings. It has to be acknowledged that an utterance does not necessarily consists of a complete sentence: the sentence can be incomplete, or multiple sentences can form a large utterance group. One alternative to identifying utterances basing off rhythmic groups as well as considering the connection between pausing and syntactic constituency (Zellner, 1994). However, identifying utterance groups en mass based on these fuzzy definitions and correspondence also faces the problem of the uncertainty of the nature of the speech context.

Another critical definition in the current set up is what constitutes a silent pause? In addition to the debate over a binary threshold, in the context of telephone conversation, there could be pauses (or gaps) at the juncture of turn taking (Beattie & Barnard, 1979). These pauses are not necessarily related to problems with speech production, but could be about courtesy to the interlocutor, or floor-holding. Filled pauses, similarly, can be used to hold the floor when two speakers were not facing each other to avoid dead air. More elaborated considerations on the structure of turn-taking are discussed in (Heldner & Edlund, 2010). Apparently these silences are not the silent pauses of the primary interest in this study.

With all the considerations above, I define the silent and filled pauses in the discussion on conversational below as within-turn pauses, which means that silences at the beginning of a turn, without preceding speech segments, as well as speech segments that solely consist of filled pauses and non-content words are excluded from the analyses. A turn is then defined as the contiguous speech segments from one speaker between two speech segments of their interlocutor.

Although Fisher transcripts only consist of arbitrarily segmented chunks of speech material, mainly for the ease of transcribers rather than offering any linguistic insight, they do offer relatively

clear clue for turn identification. Figure 1 illustrate the format of raw transcripts in Fisher. With the information readily available, turn segmentation is decided based on sorting and merging the time stamps by conversation sides provided in the transcriptions. One crucial observation that facilitates this segmentation is that floor holding or back channel talking often consist of short segments (less than four words long) of fillers or non-content words (such as *that's right, yeah*), which can be discarded without disrupting the overall integrity of speech transcripts. Although relying on the time stamps provided in transcriptions is not immune to segmentation errors, by excluding floor holding and back-channel talking through a simple rule, this segmentation method should return at least almost correctly segmented turns. Silent pause identification is then based on forced alignment results using Penn Forced Aligner (Yuan & Liberman, 2008) on the turn segmented speech.

```
# fe_03_11126.sph
# Transcribed by BBN/WordWave

0.01 1.42 A: dan martin

3.53 4.40 A: hello

4.10 5.98 B: hi hi

6.37 7.28 A: hello

6.59 9.08 B: so are we free to h- can you hear me

8.47 10.17 A: yeah i can hear you [noise]

9.53 10.64 B: (( oh okay ))

10.79 11.88 B: so

12.01 14.62 B: h- what do you think of affirmative action

15.19 18.68 A: er i thought we were talking about schools

18.97 29.41 B: um it it says the topic of the day was affirmative action and how it er implemented with the policy of
having people hired based on

19.06 19.73 A: or

30.23 36.50 A: oh mine was school um public schools in america today

36.59 38.23 B: oh really

37.57 42.09 A: maybe that's part of the thing i i don't know it's my first call that i've done so

41.82 50.97 B: oh this is the second one that i've done the last time i did it it actually er we both had the topic
but you know we could talk about affirmative action in public schools
```

Figure 1: The raw transcription format of Fisher corpus

1.4.2 Supreme Court Oral Arguments

Turn identification in SCOTUS is an easier task than that in telephone conversations, as oral debates are face-to-face conversations, which eliminates back channel talking and floor holding fillers. The time-aligned verbatim transcriptions are also segmented into contiguous speech from different parties in a debate session. Thus a turn is simply a segmented transcription file. Compared to conversational speech, the turns in this corpus are more likely to contain complete sentences. This observation may facilitate analysis of disfluencies at different syntactic or prosodic junctures. However, it does not warrant changing the minimal analysing unit from a turn to an utterance. Figure 2 is a snapshot of the time aligned transcription in the corpus.

```
2907630000 2911030000 page
2911030000 2914130000 nine
2914130000 2915030000 of
2915030000 2916130000 the
2916130000 2920730000 reply
2920730000 2925730000 brief
2925730000 2933430000 %um
2933430000 2938480000 where
2938480000 2942029999 we
2942029999 2949630000 suggest
2949630000 2952630000 that
2952630000 2967080000 concerning
2967080000 2977720000 addresses
2977720000 2983420000 two
2983420000 2994970000 distinct
2994970000 3003320000 central
3003320000 3008420000 textual
3008420000 3013920000 commands
3013920000 3017570000 {sil}
```

Figure 2: The time-aligned transcription format of SCOTUS corpus

1.4.3 Radio interviews

The Czech corpus is the most well annotated corpus among the three corpora I propose to use in this study. Sentence boundaries as well as boundaries of syntactic phrases will be determined based on corpus annotation.

1.5 The structure of the dissertation

The dissertation will be structured as the following. In chapter 2, I will set up the ground by reviewing the classification of disfluencies. More attention will be paid to the types that I will focus on in this study. Chapter 3 explores the sociolinguistic and discourse variables that potentially play a role in variations in disfluency. They include the sociolinguistic contexts that have been more thoroughly examined in the literature, such as age, gender and English dialects, as well as the less discussed questions regarding the role of conversation topics and speaker accommodation. Silent pause, filled pause and repetitions will be extensively examined. Chapter 4 primarily explores the immediate linguistic contexts in which dislfuencies happen, such as the local phrase structure, lexical property and prosodic phrasing. Implications of the effect from these linguistic variables to language production models, as well as clinical applications, will also be discussed. Chapter 5 brings in the cross-linguistic perspective of this dissertation, by surveying patterns of repetition in Czech and establishing connections with what have already known to English. Chapter 6 offers a holistic interpretation of results from both the analyses from contextual and linguistic perspectives. Finally, final discussion and remarks will be made in Chapter 7.

2 Classification of disfluency

Disfluency phenomena have been discussed from different perspectives, and have followed different descriptive paradigms. The two widely approached points of view are the forms of the disfluencies and their functions in the process of speech production (Lickley, 2015). Formal descriptions of the form of disfluencies normally do not assume particular functional underpinnings

that explain the surface variation, although the two broad perspectives are never cleanly separate. In this section, I first review some of the existing classification schemes of both the form and function of disfluencies, then I elaborate on the disfluency phenomena that are the focus of this dissertation.

2.1 Classification systems of disfluency phenomena

Speech disfluency, as implied through the terminology itself, refers to disturbance or disruption during speech production. The source of this disturbance can be pathological, but is also attributable to occasional break-downs during the production process. Levelt (1989) proposed a model which points out the locations in this process which the break downs may happen, and how disfluencies can be informative for our understanding of speech production. In this model, speech production is accomplished incrementally through three basic stages: formulating the message to be delivered, organizing the linguistic materials that are essential for communication, including syntactic planning, lexical access and phonological selection, and finally controlling the motor system to produce the intended linguistic output. Disfluencies thus reflect the disruptions that occur at different stages in this process. The cause of such disruptions can be pathological.

Unsurprisingly, some early classification systems are motivated initially to serve the needs of particular clinical population. Mahl (1956) refers to the disruption in fluent speech as "disturbance" for the main interest in distinguishing normal and schizophrenic speech. This classification recognizes eight distinct categories of "disturbance": *ah, sentence correction, sentence incompletion, repetition of words, repetition of partial words (stuttering), intruding incoherent sound, tongue slip, whole or partial word omission*. On the other hand, W. Johnson (1961), with the focus on comparing stuttering and non-stuttering speech, comes up with another set of eight categories of "disfluency". In this system, *ah* is categorized as *interjection*, and other more general summaries, such as *revision*, *incomplete phrases*, *broken words* are used to replace some of the similar but more arbitrary definitions in Malh's system. It also reflects the special need for stuttering research, with the inclusion of domain specific category of *prolonged sounds*.

With the goal to describe more general "hesitation" phenomena in spontaneous speech, Maclay and Osgood (1959) adopted Mahl's first four categories, while replacing the last four with *filled pause*, *unfilled pause*, *non-retraced false start*. Mahl's *repetition of words* and *stutter* are also consolidated into one category *repeat*. They believed that this categorization represents the most of hesitation phenomena that happen in spontaneous speech. Similarly, in yet another system, Blankenship and Kay (1964) largely follow Malh's first four categories, but change the rest into *word change* and *non-phonemic lengthening of phenemes*.

More recent studies in speech disfluency often tend to cater particular needs in domains such as human language technology and cognitive science, and rely on large scale corpus analysis. These demands require more systematic and consistent annotation mechanism. Shriberg (1994) consolidates an array of classification systems (Mahl, 1956; Maclay & Osgood, 1959; Blankenship & Kay, 1964; Levelt, 1983; Blacfkmer & Mitton, 1991; Bear, Dowding, Shriberg, & Price, 1993) into a 5-category scheme consisting the following basic forms: *filled pause*, *repetitions*, *substitutions*, *insertions* and *deletions*. This categorization is followed in later studies (Heeman, 1997; Lickley, 1998; Eklund, 2004).

Shriberg (1994)'s system also acknowledges the fact that disfluency phenomena are structured. In general, a disfluent segment consists of the word/partial word/phrase/partial phrase that to be re-

paired (reparandum), the interruption point (interrundum), the repaired or repeated word or phrase, and the resumption of fluent utterance. Figure 3 illustrates the structure of a disfluency segment in an utterance.

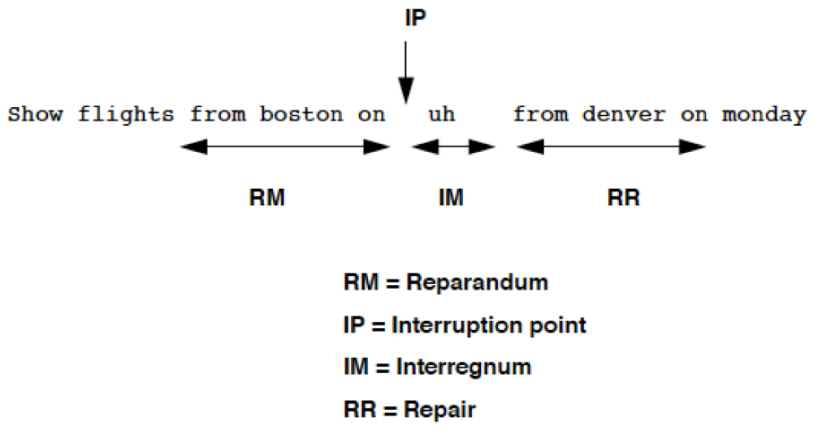


Figure 3: The structure of a disfluent region (Shriberg 1994)

Unlike the classification system of disfluencies based purely on the form variation, a functional view of a system tries to identify what is the cause of the failure in fluency. Examples of such a view include Dickerson (1972); Hieke (1981). Although these studies differ in their particular practical or theoretical goals, the classifications they adopted recognize the distinction between a way to gain more time for planning, and a strategy to re-establish fluency after a break. The functional distinction has been claimed to be distinguishable from the surface patterns in relation with the relative sequencing of repetition and filled or unfilled pauses. A more elaborated and detailed functional classification for repairs is discussed in Levelt (1983), which will be reviewed in more detail when repair disfluency is considered. Since the goal of the current study is focused more on a substantial description of the patterns, a functional classification system is of secondary interest in my discussion. Thus it will only be brought up when an interpretation of the patterns makes it necessary.

In this study, I primarily consider three out of the five generally agreed categories of disfluency (Shriberg, 1994; Heeman, 1997; Lickley, 1998; Eklund, 2004): *filled pause, repeat*, and the un-annotated but closely related hesitatioin phenomenon *silent pause*. Repair will be discussed in relation with filled pauses and repetitions, for the reasons explained below. Prolongation is yet another major disfluency type that received a lot of attention. However, I will defer to another separate study to explore its categorization and distribution, due to its close correlation with other hesitation phenomena as well as the intrinsic properties of the syllable. Therefore more fruitful analyses will be achieved only with considerations with the knowledge of other disfluency phenomena as a whole. I pay more attention to the categorization itself and how each individual disfluency phenomenon relates to variables of interests to sociolinguistic and psycholinguistic research rather than the structure of the disfluencies itself annotated or referred to in the corpora.

2.2 Silent pauses

Silent pause refers to the brief period of silence during speech production. As natural speech presumably contains silences of varying duration and for multiple reasons, such as marking the end of a prosodic group, closure during producing a stop consonant, or hesitating in responding questions, the first question about silent pause is what duration constitutes *brief*? And the second impending question is how could one distinguishes a pause that is truly hesitant from a pause that is resulted from the phonology or prosody of one's speech?

Neither of the questions has straight forward answers. To understand what is the duration that justifies a silence period as a silent pause, we should first acknowledge the large variability in terms of pausing pattern. For example, Luce and Charles-Luce (1985) reported that the duration of closure in English word final stops can vary between 30 ms to 250 ms. On the other hand, F. Ferreira (2007) argues that pauses as short as 80 ms can be the result of planning difficulty or prosodic processing. The existence of a wide range of silent pause duration puts the significance of asserting a definitive threshold for what constitutes a silent pause into question. Nevertheless, one would still hope that a reasonably defined cut-off point could still provide some information about silent pauses and hesitation within particular domains.

In the early yet still influential work, Goldman-Eisler (1958) proposed that a threshold of 250 ms should be used in research addressing questions concerning the cognitive process of silent pauses, as pauses shorter than this threshold are generally accounted for by articulatory adjustments during speech production. Although adopted by many (Mack et al., 2015; Beattie & Butterworth, 1979; Greene & Cappella, 1986), some lower threshold values have also been used (Gee & Grosjean, 1983; Eklund, 2004; Martin, 1970). In some, a much lower threshold, as short as less than 100 ms, was used in annotation (Martin, 1970; Eklund, 2004). Butcher (1981) further demonstrates that what perceived as a silent pause does not only depend on the absolute duration, but is also conditioned on the prosodic context. Thus it is more informative to understand what is the distribution of silent pause duration, and selecting a threshold reflecting context specific properties.

To attend to the context sensitive nature of hesitation pauses, many studies ask what are the discourse, syntactic, prosodic and dialectal effects on pause duration (Levelt & Cutler, 1983; Krivokapić, 2007; Zvonik & Cummins, 2003; Kendall, 2009). Using a synchronous speech method, Zvonik and Cummins (2003) showed that silent pauses between syntactically more complex phrases and longer prosodic phrases (Krivokapić, 2007; Zvonik & Cummins, 2003) are longer. However, Krivokapić (2007) suggests that more complex prosodic structure doesn't have equivalent effect as an increase in syntactic complexity. Kendall (2009) argues that speakers from different dialect background also vary in their pause duration. On the question of silent pause distribution, through a large scale multilingual study, Campione and Véronis (2002) showed an existence of bi- or tri-modal distribution in silence distribution in spontaneous speech, and offered a sharp criticism on the handling of statistical analyses of duration contrast in most research practices.

Research into the relation between silent pause duration and its immediate syntactic or prosodic context inevitably touch on the second question: How to distinguish the pauses that result from hesitation from a fluent pause? Krivokapić (2007) and Zvonik and Cummins (2003)'s results may indicate that there is some relation between properties of fluent pause and the syntactic and prosodic structures in which the pause occurs. Attempts have been made to link the complexity of syntactic and prosodic structure to fluent pause duration (Cooper & Paccia-Cooper, 1980; Gee & Grosjean, 1983; Watson & Gibson, 2004). However, F. Ferreira (1993, 2007) argues that there

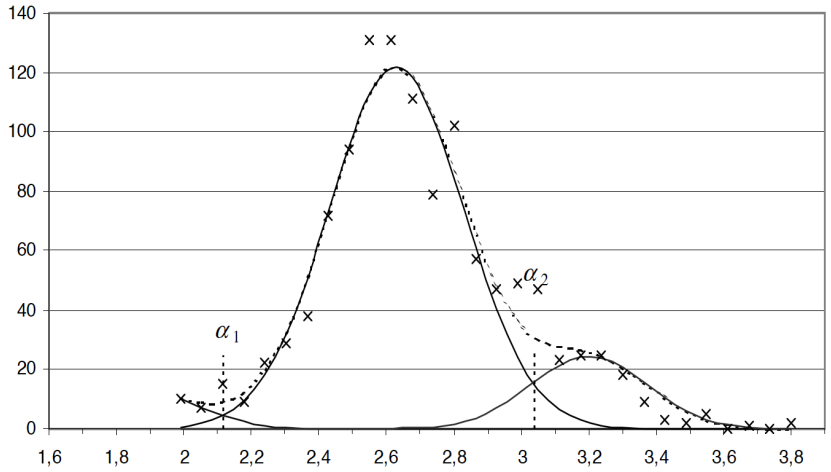


Figure 4: Bi-model distribution of pause duration in French spontaneous speech, as demonstrated in Campione and Véronis (2002). The time units are in $\log_{10} ms$.

isn't a direct relation between the structure of syntactic and prosodic phrasing. According to her proposal, pauses have to be distinguished between prosodic and performance based pauses. The distinction can be made through the relation to the phrase before and after the pause: performance based pauses are associated with the following phrase, while prosodic based ones are associated with the preceding phrase. Therefore only performance based pauses are related to difficulties in planning and can occur anywhere in an utterance. Thus performance pauses occurring at syntactic phrase junctures may reflect difficulties in syntactic planning, and within phrase pauses may be related to problems with lexical access, according to Levelt's model for language production.

Although Ferreira's distinction between performance and prosodic based pauses offers a clean explanation for understanding the pausing phenomenon, it doesn't really address the question of how to differentiate the two types of pauses in practice, especially in annotating speech corpora. Thus in practice, researchers still utilize perceptual based judgement (Nakatani & Hirschberg, 1994; Eklund, 2004; Clark & Tree, 2002) in identifying hesitation pauses, and it seems still the best practice before consistent and objective way to identify pauses has been developed (Lickley, 2015).

2.3 Filled pauses

Filled pause, which is also often referred to as filler, is probably the most easily distinguishable and heavily studied disfluency phenomenon. A crucial distinction between filled pauses and filler words has to be made clear. Filled pauses, in most disfluency studies, refer specifically to *um* and *uh* and their counterparts in other languages. Filler words, on the other hand, may include discourse markers such as *well*, *you know* etcetera, which are not a concern to the current study.

Filled pauses, with more explicit forms compared to other disfluency phenomena, are more easily identified in larger scale speech data. A more accurate frequency count can subsequently be obtained. The average frequency of filled pauses is somewhere between 1.3 to 4.4 per 100 words, depending on the corpora being analyzed (Bortfeld et al., 2001; Shriberg, 1994; Eklund, 2004; Lickley, 2015). Cross-linguistically, as demonstrated in a range of Germanic languages as well as

in French, Spanish, Hebrew, Japanese and Mandarin, a filled pause generally takes two forms: a pure (often times reduced) vowel (such as a schwa), or such a vowel followed by a nasal coda. The exact realization of the two alternatives varies in different languages. In American English, it takes the forms of *uh* and *um*, while it's often transcribed as *er* and *erm* in British English (Clark & Tree, 2002; Tottie, 2011).

The occurrence of filled pauses is generally regarded as a primary sign of hesitation. The hesitation can be the result of the speaker being uncertain (Brennan & Williams, 1995; Smith & Clark, 1993), under high cognitive demand (Arnold et al., 2007), or facing a choice (Schachter et al., 1991). Many of the structural properties of filled pause are often understood with regard to the general role of disfluencies in speech production (Corley & Stewart, 2008). As suggested by several studies (Oviatt, 1995; Bortfeld et al., 2001; Smith & Clark, 1993; Brennan & Williams, 1995; Swerts, 1998; Swerts & Krahmer, 2005), the insertion of a hesitation marker such as a filled pause may not be purely automatic. For example, lower rate of disfluency is observed in human-machine communication than human-human conversation (Oviatt, 1995), and Bortfeld et al. reported that in their highly controlled production environment, disfluency rate, particularly the rate of filled pauses, is greatly influenced by the role played by the speaker. However, these results seem to conflate the hesitations that are primarily driven by the need for planning and message structuring, and the hesitations that are resulted from contextual constraints on performance. It is also not clear whether filled pauses have any idiosyncratic properties that are distinct from disfluencis in a broader sense.

Apart from their apparent identity as hesitation markers, elements of non-randomness in filled pause distribution can be used as arguments for it being lexicalized in one's vocabulary. If filled pauses are to operate in parallel with other filler words that function as discourse markers, they should be treated equivalently as *you know* and *well* which marks transitions or turns in the discourse, and therefore argued to be encoded with explicit discourse meanings (Clark & Tree, 2002). This view of filled pause, however, is refuted by (Lickley, 2015). Citing acoustic evidence from (Shriberg & Lickley, 1993), where it has been shown that fundamental frequency of filled pauses is closely related to surrounding phrases, Lickley (2015) argues that it is unlikely that speaker intentionally insert filled pauses as signals for upcoming hesitation. If it was the case, disruptions of the prosodic structure of the otherwise fluent utterance should be expected.

The curious existence of two alternative forms of filled pauses across languages has also been put under scrutiny. As the variation in its forms may suggested, Clark and Tree (2002) argue for a differentiation between the meaning of the two filled pauses. They claim that the nasal filler corresponds to a major delay in production, while the oral version signals a minor delay. This claim is supported by the observation that *um* tends to occur at the beginning of an utterance, while *uh*'s location is more often utterance internal (Shriberg, 1994; Shriberg & Stolcke, 1996). Speakers also appear to have preference over one form than the other, and some even exclusively use only one form (Shriberg, 1994, 2001). On the other hand, abundant evidence points to different preference of one form over the other by people from different socioeconomic groups (Wieling et al., 2016; Fruehwald, 2016). The trend that younger, especially female, speakers prefer *um* over *uh* is thought to reflect a change in progress that has spread across several Germanic languages (Wieling et al., 2016).

From the review above, it can be concluded that filled pauses, like other disfluency phenomena that signal hesitation, function both as facilitators for production in case of a forthcoming hesitation and a sign of disturbance in performance conditioned on the context. There is a clear distinction

between the two forms of realizations, which can be the result of lexical difference, speaker's intentional choice, sociolinguistic language variation, or a bit of something from this list.

2.4 Fluent repetitions

Repetition is another common symptom of speech disfluency which has received much attention in the literature. In the current context, 'fluent' refers to the fact that the repetition phenomenon does not interrupt the flow of speech, or cause trouble for the listener in parsing the speech input given the presence of disfluency. I distinguish *fluent repetition* from *pathological repetition* such as what can be observed in stuttering, although they can share certain similarities (Guitar, 2013). Notice that the terminology *fluent repetition* is also used to describe an emphatic strategy that is used to emphasize or making a contrast, such as in sentences like *I really really like your idea.* This kind of repetition can be distinguished from repetitions caused by hesitation or message structuring through prosody (Lickley, 2015). Emphatic pitch, for example, can be marked on the repeated intensifiers, and other disfluency phenomena, such as silent and filled pauses, or prolongations, are not expected in the neighborhood. The emphatic repetitions are also mostly used with a limited set of content words or phrases, whereas unlikely with pronouns, prepositions and conjunctions. On the other hand, hesitation repetitions often don't carry prominent prosody, and are mostly likely to be function words (Clark & Wasow, 1998; Fox et al., 1996, 2010). In fact, Clark and Wasow (1998) reported a 10 times higher frequency of functional words being repeated compared to content words (25.2 per 1000 words vs 2.4 per 1000 words), and Lickley (1994) reported a statistics of 96% of repeated words being functional words. In terms of the location of repetition, hesitation repetitions are often at the initial of an argument (Fox et al., 1996, 2010). Cross-linguistically, at least among English, German and Hebrew, the repeated words are function words that immediately preceding the main content word of a clause. The distribution of exact lexical category of the function words, however, varies across languages, which can be interpreted as conditioned on the syntax (Fox et al., 1996). In the discussions in this study, the term *repetition* is used exclusively to refer to the fluent repetition caused by hesitation or message structuring, as reviewed above.

The forms of repetition can vary among single syllable words, multi-syllable words, multi-word phrases and word fragments. Some examples of possible variations of repetitions are listed in (1). The repeated words are only marked with boldface for the moment, although detailed annotation strategy for variations in repetition will be given in section 3.4. Although due to the correspondence between repetition and function words results in higher rate of single-syllable word repetitions in the languages reported in literature, repeating word fragments or multi-word phrases is not that uncommon. Although there isn't reported statistics on the frequency of these different types of repetitions, counts on the frequency of disfluencies involving word fragments are available in several languages. Levelt (1989) reported 22% of word fragments in a Dutch pattern description corpus; Lickley (1994) reported 36% from conversational speech in British English; and Bear et al. (1993) found 60% in ATIS corpus. The variation may be caused by difference in the nature of the corpus, but it can be speculated that among the reported word fragments, a substantial amount would involve repetitions. A careful description of the distribution of repetition types beyond word classes seems to be necessary.

(1) a. and **i i** actually don't watch any sports on television.
b. and **it's it's it's** a big question and i don't know that you know bush had an answer for that.
c. actually i am. **i'm not i'm not** so afraid of it.
d. yeah yeah they were **pr- pretty** distance portions of the state

In an account offered by Clark and Wasow (1998), the presence of repetition shows an effort by the speaker to preserve the continuity of the speech when faced with higher cognitive demand after the initial commitment to the initiated constituent. This argument explains two facts about repetition: it is more likely to occur at the beginning of long and complex clauses, and when people repeat, they tend to restart from the beginning of the constituent where interruption happened. The higher rate of function words in repetition is mainly because their location in the constituent. This theory can in principle cover some causes of repetitions, but is unlikely to be the only or a major explanation. Apart from serving as a sign to restore interruption, Hieke (1981) proposes that repetitions can be a strategy that speakers uses to coordinate the flow of speech. This distinction between passive and active control of speech is termed *retrospective* and *prospective* repetition respectively. Later acoustic analyses have suggested that the two categories can be distinguished when pause length and prolongation in the region of repetitions are considered (Plauché & Shriberg, 1999; Shriberg, 1995). A third category that parallels with Levelt and Cutler (1983)'s notion of *covert repairs* is also able to be distinguished according to Plaushe and Shriberg's study. This type of repetition can be briefly described as a 'cover up' of a pre-detected speech error before articulating the erroneous lexical item, therefore the lexical items before the error are repeated to preserve speech continuity.

However, results from these efforts in identifying the functions of repetitions are still incomplete. For example, one fairly common form of rapid repetition has not received much attention. In their corpus study of the time lapse between cut-off and repair, Blacfkmer and Mitton (1991) reveal that the time gap can be very short and frequently effectively zero to signal any noticeable delay in the speech signal. More notably, listeners are often unaware of the existence of such repetitions, which can be attested from an accuracy evaluation of careful transcriptions of spontaneous speech (Lickley & Bard, 1998). The question, then, is that is this type of repetition a sign of hesitation, just as most other repetitions are, or they are simply random additions to the output speech as a result of execution error?

One final deficiency in the current literature inventory on repetition is the lack of cross-linguistic perspective, especially a lack of understanding from languages with distinctive syntactic structure compared to that of Germanic or Romance languages. Fox et al. (2010) have already suggested the tendency to repeat function words needs to be interpreted conditioning on the morpho-syntactic restrictions and word order of the language. It remains to be seen what is the most likely unit for repetition in a language which has rather limited inventory of function words such as prepositions and personal pronouns, or ordering them after the main content word in an argument.

2.5 Repairs

At least some evidence has suggested that repetitions can sometimes be classified as a form of repair, which has been termed as *covert repair* by Levelt (1983). However, *repair* in fact covers a wider range of more complex disfluent phenomena, and is often the place of confusion. Shriberg (1994) developed a structural coding algorithm in relation to the basic structure of the disfluency

region, as illustrated in figure 3. Seven types of disfluencies were originally identified through this system. However, for more efficient representation, I combined the category *conjunction*, which essentially refers to repeating conjunctive adverbs, with *repetition*. As shown in figure 5, components in a disfluent region are identified and annotated following a fixed order. Using this algorithm, the three types of disfluencies reviewed above can be unified within more complex repair phenomena with a single structural representation, differing only in terms of what slots in the region are occupied.

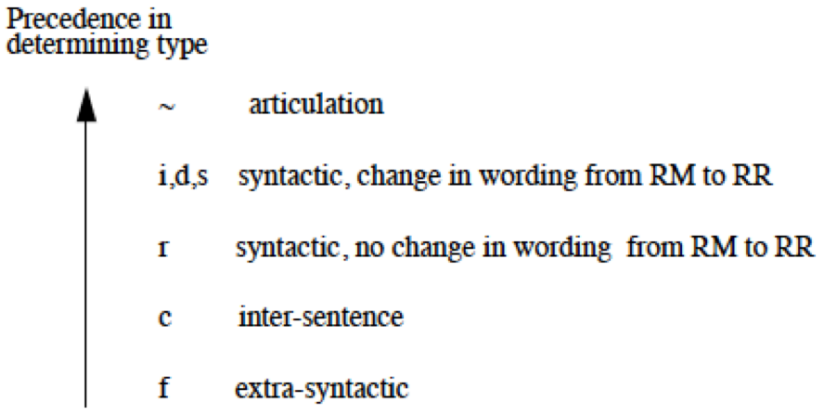


Figure 5: The classification algorithm used in Shriberg (1994). f: filled pause, c: conjunction, r: repetition, i: insertion, s: substitution

Such a structural representation implies that a faithful description that regard filled pause and repetition as special cases for repair can also be naturally well justified. Therefore to avoid confusion in terminology, I will only refer to the repairs involving replacing the phrase in reparandum (RM) with repair (RR) as *repairs* in the following discussion. In other words, the repair phenomena reviewed here refer to *hybrid, substitution, insertion* and *deletion*, as listed in Shriberg's annotation scheme summarized in figure 6, which includes instructions on what components to look for in determining the type of disfluencies.

In addition to structural representation of repairs, functional accounts have also been proposed to offer a categorization through examining the different cause of repair phenomena. In the theory brought up by Levelt (1983), five categories can be identified based on the reason of repair: **D-repair**: *refers to when original ongoing speech is aborted in exchange for something different*; **Appropriateness Repair**: *refers to when a speaker realizes that something in the speech is correct but needs to be modified for better communication*; **Error Repair**: *refers to the attempts to correct an error detected in the original speech*; **Covert Repair**: *refers to the repairs that are initiated before the error is produced, which result in repetitions of words right preceding the potential error site*. The last category is essentially all other repairs that don't fit the four established categories. According to Levelt, D-repair is equivalent to deletion. Appropriateness Repair can be initiated by the need to resolve ambiguity or offer further specification, while Error Repair can be further grouped into lexical errors, syntactic errors and phonological errors. Finally, Covert Repair may be the course to correct appropriateness or errors at the conceptual or planning stages.

However, this functional classification is highly subjective as it relies on judgements of speaker's

Classification algorithm for repair disfluency adapted from Shriberg's (1994)

Types	Must include*	Must not include	Optionally include
Hybrid	S,I,D	n/a	R,F
Substitution	S	I,D	R,F
Insertion	I	S,D	R,F
Deletion	D	S,I	R,F
Repetition	R	S,I,D	R,F
Filled pause	F	S,I,D,R	n/a

* Symbol meanings: S: substitution, I: insertion, D: deletion, R: repetition, F: filled pause

Figure 6: The classification scheme used in Shriberg's algorithm for repair annotation.

model of listener's knowledge in the discourse (Shriberg, 1994). This judgement may inevitably lead to confusions in assigning classification labels. Blacfkmer and Mitton (1991) used a simpler functional based classification system, where only two major categories: *conceptual* and *production* based repairs are distinguished. However, this approach by design is also not able to address the inherent problem of subjectivity in functional classification.

Through comparing the distribution of *deletion* and *repetition* in Switchboard, Shriberg (1994, 2001) found that speakers can be grouped into *repeaters* and *deleters* by the strategies they adopt in coping with the cognitive demands in talking while planning. She further argued that the possibility of relating this apparent strategic difference to cognitive processes in planning and production finds support from the prosody: repeaters have slower speech rate than deleters. This speed difference may be a reflection of the difference in the underlying processing speed between the two groups. However, since her primary intention was to offer a theory neutral description of disfluencies, the main take-away of this observation should be a stress on the significance of careful type descriptions of repair and repetition.

An extensive study of all the repair phenomena requires substantial effort in creating carefully annotated corpus. However, this requirement poses a major constraint on disfluency research in general. Automated annotation of disfluencies, especially repairs, is not impossible, but the performance of automatic systems is highly contingent upon the domain of annotated data that is used for training. By far, Switchboard is still the standard and primary source of annotated corpus for systems in disfluency annotation, such as Hough (2014), although advances in machine learning and natural language processing have been tremendous since 1992. A more extensive careful description of not only repairs, but disfluencies broadly, can help to push forward the efficiency of semi-automatic annotation, and benefit speech technology community by complementing the perspective that algorithmic advances may never catch. Such a description should be based on a sample with at least more than 20 speakers.

2.6 Summary

In this section, I reviewed the surface variations of the form of disfluent speech in normally fluent speakers. As discussed above, these categories are not independent from each other, and the nature of different disfluency categories determines the burden on researchers to identify and properly annotate the speech transcripts. This unfortunately limits the scope of analyses that can be efficiently performed on the categories that are harder to identify and correctly annotate. This limitation constraints researchers to conduct more extensive careful descriptions of the distribution and patterning of disfluency phenomena, especially repetition and repair disfluencies. Therefore in this study, I will combine automatic identification and extraction with semi-automatic manual annotation based on disfluency types. The primary focus of discussion will be on silent pause, filled pause and repetition. Due to its complexity in surface form variation, repairs will only be discussed in connection to the three disfluency types focused here. I will leave a detailed large scale description of repair disfluency for a future project.

3 The macro-analysis of disfluency

This chapter describes how disfluencies vary in response to changes in the broad context in which speech is produced. These factors not only include sociolinguistic variables to the interests of a sociolinguist, but also those related to the discourse of conversation. Here I first review the literature on the sociolinguistic and discourse variables' effect on disfluencies, then show to what extent these patterns are reproduced from the corpora that I currently use, and report results from previously under-explored variables.

3.1 Background

The effects of sociolinguistic variables, such as age, gender and dialectal variation have received relatively little attention in the early disfluency literature. Although discussions on various issues related to disfluency phenomena can be traced back to Maclay and Osgood (1959), and Levelt (1983); Shriberg (1994, 2001) have reported gender difference in disfluency rate distribution in both a sample of 6 Dutch speakers and Switchboard, the question of how sociolinguistic variables affect disfluent patterns was not systematically investigated until Bortfeld et al. (2001). More research efforts have been directed to this topic in the past decade, with more comprehensive comparisons of the use of filled pauses across gender, age and socioeconomic groups (Tottie, 2011; Acton, 2011), and English varieties (Tottie, 2014; Kendall, 2009). In addition to age and gender, Laserna, Seih, and Pennebaker (2014) also considered personality as a potential informative variable. The use of filled pause itself is also treated as a sociolinguistic variable (Fruehwald, 2016) which in itself is a language change in progress, where a trade-off in the frequency of *um* and *uh* has been observed. This view is also upheld by a later study (Wieling et al., 2016), where the same trend appears to persist across several Germanic languages. Sociolinguistic variables are also examined in Yuan, Xu, Lai, and Liberman (2016) in Mandarin, where gender effect has also been reported. However, even fewer studies have looked at other disfluency phenomena and include individual variation as an articulated research question. Among them, Kendall (2009) studied the distribution of silent pauses across dialect region in North America, with the goal to attach social

meanings to silent pause variation. Roberts, Meltzer, and Wilding (2009) focused on repetitions in fluent male adults for bettering the treatment of stuttering adults. McDougall and Duckworth (2017) investigated individual variation of prolongation and repetition, in addition to silent and filled pauses, for speaker identification in forensic settings.

Bortfeld et al. (2001) approached the problem through a language production experiment, in which pairs of speakers were asked to pair sets of pictures through a mixed factorial design. The factors that were controlled for include familiarity of the picture, age, gender, education, and marriage situation. They analyzed both the overall disfluency rate and several individual disfluency categories. In terms of overall disfluency rate, more disfluencies have been found in more demanding planning tasks, such as unfamiliar domains and the role taken by the participant in the picture matching task. With regard to each disfluency category, there was a change in the predominant disfluency pattern conditioned on familiarity and roles taken in the task. They also reported that older speakers produced more fillers than younger speakers, and men had higher disfluency rate overall.

Given the pioneering nature of Bortfeld et al's study, it did suggest that sociolinguistic variables are crucial components in determining how the display of disfluencies in individual's speech may vary. However, some factors in their experiment, such as familiarity of the pictures and the role played by individual speakers, reflect less on the effects they called cognitive demand, especially if one wants to generate their results from lab speech to more general settings. One apparent example is that they used geometric shapes in their "unfamiliar" category, while actual kids in their "familiar" category. From their description of the task, it is unclear whether experiment participants' perception aligned with their specified conditions, and how this set up is paired with cognitive demand.

Among corpus studies of sociolinguistic variables' effect on disfluency, the most heavily discussed topic is gender and age related variation in the use of filled pause. In some earlier work, it has been acknowledged that *er* is the second most characteristic word for male speakers and fourth most characteristic word for speakers who are 35 years of age or older (Rayson, Leech, & Hodges, 1997), through looking at British National Corpus. *Erm*, on the other hand, is among the most characteristic words for people from higher socioeconomic classes. Using a collection of corpora of telephone conversations (including Switchboard and Fisher), Liberman (2005) observed that *uh* was used more frequently among male and older speakers, whereas *um* was more frequent among female and younger speakers.

Several more recent corpus studies have been dedicated to sociolinguistic variables in disfluency production. Tottie (2011) compared between two British corpora: the British National Corpus (BNC) and London-Lund Corpus (LLC), and across multiple speech styles and speaker groups therein. Although the statistics reported in this study is exploratory in nature, it points out several directions for future exploration. In addition to the observations that men use more fillers than female and a tendency for higher filler frequency among older speakers, she also raises the question of what is the socioeconomic status' effect on the use of fillers. Through comparing the socioeconomic stratification in BNC and LLC, she proposes that people with higher socioeconomic status tend to use more fillers as well. However, she doesn't further address how these factors interact, and stops at relating these sociolinguistic variables to the role of fillers in planning.

In a later study, Tottie (2014) compared the use of fillers between American and British English. She showed that in her sample of American English, the Santa Barbra of Spoken English Corpus, male speakers do not maintain a higher rate of fillers compared to female, inconsistent with the

trend she reported for British English. She also argues for the existence of similarity between the distribution of *um* and *uh* and discourse markers, such as *well*, *you know* in her sample of American English, thus an evidence for the different discourse functions played by fillers in two varieties of English. Conversation topic and formality are also suggested to be influencing factors in filler production through this comparison. Although the variables she proposed can in principle be useful in identifying the distributional variation in the use of fillers across population groups, the corpora of her choice may not be optimal to offer unbiased claims about these factors. One fundamental problem is that the Santa Barbra corpus is much smaller in scale and sampled a different demographic group compared to BNC and LLC, even though all three corpora she used consist of conversational speech.

Acton (2011), on the other hand, documents the gender difference in *um* and *uh* distribution in American English from two spontaneous speech corpora: A speed-dating corpus collected from graduate students at an American university, and the Switchboard corpus. In both corpora, he foudn a higher rate of *um* among female speakers than male. Both Tottie and Acton's work suggest that *um* is gaining currency in their respective speaker population, and Acton further proposes that this change is persistent across gender and age group.

The initial observation of a potential change in progress in the use of *um* and *uh* as documented in Tottie (2011) and Acton (2011) has received more attention in Fruehwald (2016) and Wieling et al. (2016). Fruehwald (2016) examined the frequency of *um* and *uh* by gender and age group using the Philadelphia Neighborhood Corpus (PNC). His apparent-time analysis shows that there is an apparent increase in popularity of *um* among younger generations in the past century, and female speakers appear to lead this trend. This increase in popularity, however, is accompanied by a decrease in the relative frequency of *uh*, thus showing a trade-off between the two variants of fillers. This trend is clearly seen in 7. Expanding upon Fruehwald (2016), Wieling et al. (2016) further examined a range of Germanic languages, including English (British and American varieties), Dutch, German, Norwegian, Danish, and Faroese, in which a similar binary variation of filled pause (a vowel plus nasal version and a vowel without nasal version) is in use, through mixed-effect logistic regression using a variety of spoken and written corpora (for a detailed list of corpora, see Wieling et al. (2016)). In all the spoken corpora they examined, significant effect of age and gender on the likelihood ratio of *um* vs. *uh* use has been found in all the regression models they built for each of the corpora. However, great variation in terms of overall proportion of *um* and *uh* use is also observed across their corpora, even within a same language. This variation may be due to factors such as topic and domain variation in conversation, individual variation, dialectal variation, and some other unobservable endogenous variables. Two tentative explanations for this cross-linguistic trend of change from the perspective of language contact and a potential extra-linguistic force that enables an independent yet parallel change have been proposed to account for this interesting change in progress.

Compared to more extensive contributions for the effect of sociolinguistic variables on filled pause production, the number of studies on individual variation and topic in silent pause distribution and other disfluency phenomena is relatively limited. The range of individual variation in pause distribution was reported as early as in Goldman-Eisler (1968), both in spontaneous and read speech. Duez (1982) compared both silent pauses and other disfluent non-silent pauses of French across three speaking styles: political interview, casual interview and political speech. Each speaking style contains 5 to 7 speakers, with average speech time around 30 minutes. It is found that silent pauses are longer and more frequent in both political and casual interviews compared to

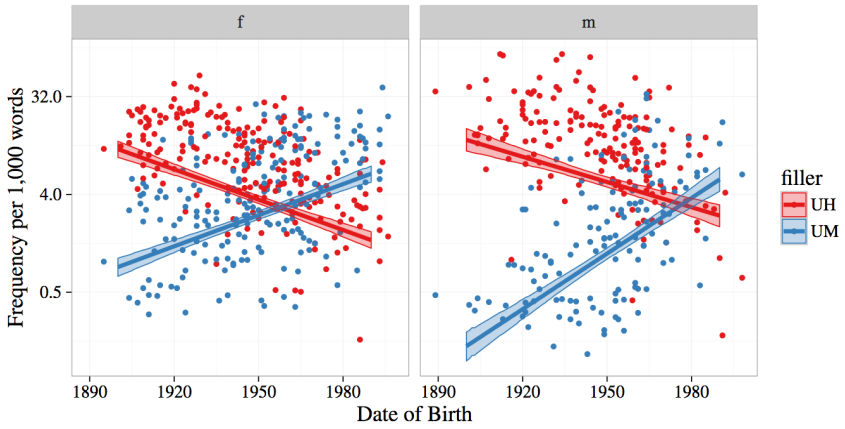


Figure 7: UM usage and UH usage trading in frequency in PNC, from Fruehwald (2016)

political speech. A wide range of individual difference is also noticed. For other disfluency types, Shriberg (1994) showed that individual preference differs between repetition and repair (repeaters and deleters) when faced with production problems, and the three speaking styles also varies in terms of overall disfluency rate and type distribution. Form variation of repetition in relation to discourse factors such as the topic and individual properties, however, is less explored.

Studies reviewed thus far have all demonstrated that age, gender and other socioeconomic factors such as socioeconomic status and education all have effects on filled pause production, especially the relative frequency of the use of two versions of fillers. A preference of using *um* over *uh* has also been demonstrated in difference sources. This trend is persistent both across varieties of English and across the Germanic family. However, as pointed out by several authors (Tottie, 2011; Acton, 2011), there exists a great potential for individual variability, the extent of which is less understood. In addition, our understanding of other discourse or extra-linguistic factors such as conversation topic is rather limited. These lesser explored factors may be the underlying variables that explain the variation contributed by age and gender, through which we can connect sociolinguistic and cognitive forces that shape human language production, and build our knowledge towards a causal inference on why the surface variations are the way they are.

In the following sections, I address the two aforementioned questions: What is the range/effect of individual variation in disfluency production, and what is the role played by conversation topic in this variability, through quantitative analyses of three forms of disfluencies: silent pause, filled pause and fluent repetition. I first examine the variation in silent and filled pause using Fisher corpus, focusing on the effect of conversation topic and the consistency within individual speakers across conversations. Later, I will turn the focus to fluent repetitions in SCOTUS in chapter 4, addressing how fluent repetitions may vary across individual speakers, and whether there is correspondence between patterns of repetition and syntactic and/or prosodic phrasing.

3.2 Silent pause

In this section, I address the question of what is the individual variation in silent pause production, and how silent pauses are affected by conversation topic. Individual variation as addressed in this section mainly refers to how likely it is for an individual speaker to vary in their pausing patterns

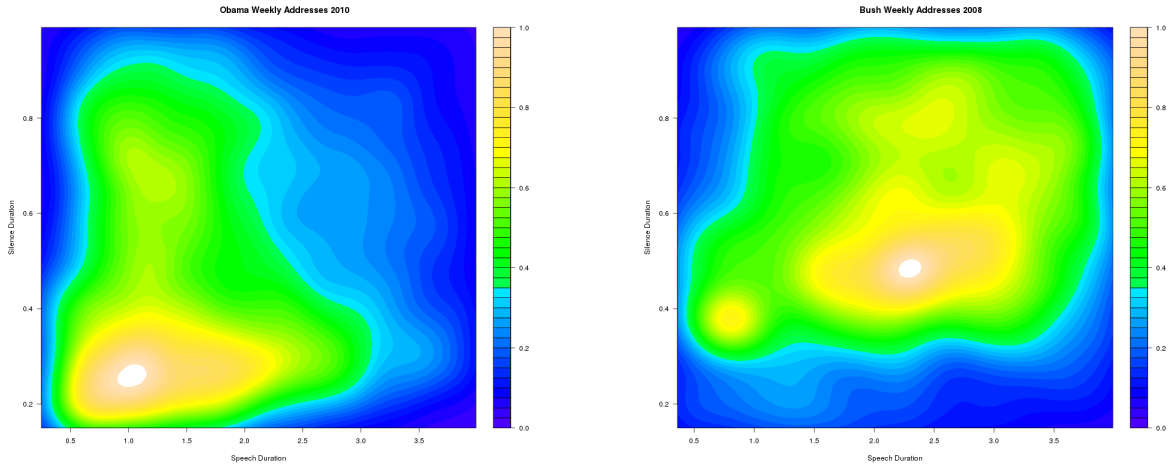
across conversations. I first raise a fundamental problem regarding research in silent pauses, then proceed with the analysis using a redefined objective quantification of silence in speech production.

3.2.1 An objective and robust representation of silence segments in speech

A challenge in silent pause research, which also regards the very fundamental definition of silent pause, is what is the appropriate threshold for separating true silent pauses that relate to hesitation or processing problems encountered during production from the silence that is the result of phonological or prosodic processes. As reviewed in earlier sections, various thresholds have been used in the literature, ranging from 80 ms to 2 seconds (F. Ferreira, 2007). Several studies have provided descriptive analyses of silent pause distribution across speaking style (Zellner, 1994) and languages (Campione & Véronis, 2002). The general findings on the question about proper selection of silence threshold are somewhat inconclusive, due to large amount of context dependent variation and subjective judgement. However, it is generally acknowledged that silence duration distribution is bimodal or multimodal, and a threshold of 200 ms can serve as a sufficient cut-off point for most purposes. In a large-scale multilingual study of pause distribution in both spontaneous and read speech, Campione and Véronis (2002) showed that the distribution of pause duration bears language-specific traits, and the choice of threshold can subjectively change the results of statistic analyses. Faced with such a wide range of variability, one might still prefer subjective judgement as the better practice in determining whether a silent segment constitutes a silent pause (Lickley, 2015; Nakatani & Hirschberg, 1994; Eklund, 2004). Nevertheless, there is a need for a more objective quantification of silent pause duration that is robust to contextual and individual variation, so that cross domain comparison can be made possible.

In this study, I rephrase the question of what is the absolute cut-off duration for categorically separating silence from speech as what is the relation between silence duration and the duration of speech segment preceding and/or following the silence. This rephrase can objectively and consistently quantify the dynamics of speech production that an absolute separation of silence from speech through a hard boundary cannot accommodate. This approach acknowledges that silence is an amalgam of linguistic, cognitive and extra-linguistic factors that reflecting both the syntactic and prosodic, and the cognitive perspectives in language production. The relative duration between silence and speech segments implicitly incorporates these multivariate space and simultaneously preserves the identity of pauses, while releasing the burden of selecting an appropriate threshold for particular purposes. On the contrary, an absolute hard cut-off point would unavoidably over or under estimate of the rate of silent pauses for individuals with varying speaking rate, and potentially misrepresenting discourse or structural pauses as hesitation pauses. Thus analyses of durational or distributional relations of silent pauses following this thresholding would be biased based on the exact context and threshold chosen.

The relation between silence duration and preceding or following speech distribution is explored through estimating the joint probability density of 2D (bi-gram silence duration plus the speech segment duration before or following silence) or 3D (considering speech-silence-speech sequence) duration space. This method is non-parametric and assumption-free, meaning that biases imposed by researchers or particular research questions can be largely eliminated, while multiple assumptions can be tested by directly working with a probabilistic distribution, controlling for the parameters of interests. In this manner, group differences in silent pauses can be easily observed from the joint distribution of pause duration and speech duration before and after the pauses, and



(a) Obama (b) Bush

Figure 8: Joint density plots of silence duration (y-axis) and following speech duration (x-axis) in seconds of president weekly address for Obama and Bush.

parameterized using dimensionality reduction methods on the joint distribution space. Thus a compact representation of the pausing pattern can be achieved for each individual. Statistics on group differences can then easily be applied.

Figure 8 plots the joint density estimation of silence duration and following speech duration for a year's worth of president's weekly radio address for Obama in 2010 and Bush in 2008 (Liberman, 2016). These plots clearly show some structure that is the result of individual variation in speaking style between the two former US presidents. For example, Obama's speech appears to have a peak at the coordinate around (1.0, 0.25), which suggests that his speech may be characterized by shorter speech segments between relatively short pauses. The secondary peak, at around (1.2, 0.7) may signal some longer pauses between paragraphs. Similar description can be made for Bush's speech, and a clear distinction between the speaking style of two presidents can be made basing off their distinctive patterns in pausing. Analysis of the relative duration between adjacent silence and speech segments in the other direction can be similarly carried out, so as the joint tri-gram speech-silence-speech duration.

With this simple demonstration, I have shown that a speaker's pausing characteristics can be captured by looking at the joint distribution of silence duration and following speech duration. This characterization can then be treated as the feature representation of individual speech. Therefore, we can perform some dimensionality reduction technique, such as Singular Value Decomposition (SVD), to achieve a compact representation of the information contained in these 2D density plots for each individual. Figure 9 is a joint density plot of the first two left singular vectors derived from over 5,000 read paragraphs in LibriSpeech (Panayotov, Chen, Povey, & Khudanpur, 2015). The input matrix to SVD is the flattened joint 100×100 2D density matrices of silence duration and following speech duration, obtained from a Speech Activity Detector (SAD) (Walker et al., 2015), for each read paragraph in the corpus. In this derived space, each combination of the values in the two latent dimensions represents a potential speaker in the population from which the initial sample is taken. The distribution in this derived space is clearly bi-modal, with a primary mode closer to the center of the graph, and a secondary mode towards the lower right corner. This bi-modal

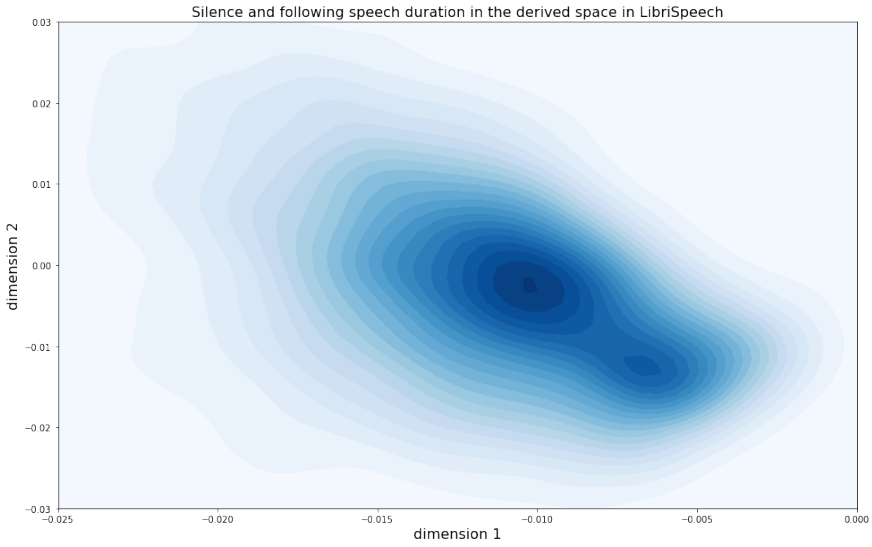


Figure 9: Plot of silence duration and following speech duration distribution of LibriSpeech in the derived 2D space.

distribution could reflect two distinctive reading strategies that readers use when contributing to the corpus. This strategic difference may be the result of genre difference, gender difference, or whether the reader has received professional training. Thus, further explorations of the underlying explanatory factors can then be carried out.

To sum up, in this section, I proposed a more objective and non-parametric quantification of silent pause distribution in speech production. This quantification method first makes reference to the speech segment duration adjacent to the silent segment duration. The resulting joint density estimations are then passed to some dimensionality reduction method, here SVD, to achieve a compact representation for each individual in lower dimensional space. Through a simple demonstration with LibriSpeech and weekly presidential address, I have shown that this quantification method is effective in capturing individual variation, as well as the underlying structure in population distribution.

3.2.2 Individual variation in silent pause and the effect from conversation topic

Here, I use the quantification method illustrated in the previous section to address the question of individual variation in silent pause distribution. As discussed in the literature, both socioeconomic variables (Tottie, 2011; Acton, 2011) and conversation topic (Lickley, 2015; Bortfeld et al., 2001) may have an effect on the disfluencies in natural speech. The question then is how these variables affect silent pause distribution?

The data I use to answer this question is the sample from Fisher corpus. I first conduct an exploratory data analysis, with the goal to explore the in-sample group differences with regard to the socioeconomic variables reported in the literature, as well as conversation topic. Then I perform a regression analysis to attest the observed group differences, and examine the potential interactions among explanatory variables. Individual variations are represented in the derived 2D space generated from the joint density estimation of silence duration and following speech segment duration for each individual speaker. The first two left singular vectors are used to construct the

2D space. The lower bound of silence duration is set at 150 ms to minimize interference from potential word-internal or within-phrase fluent pauses, but remain generous regarding all other pausing scenarios.

Gender The group difference in gender in the derived space is plotted in figure 10. The entire sample contains speech from 1499 male speakers and 1658 female speakers. The density plots suggest that both male and female speakers have an overall similar shape of distribution in this derived space, while there is a larger spread among female speakers, but also a more clearly defined center of the distribution, compared to male speakers. However, it should be expected that small, yet significant, difference exists between the two gender groups.

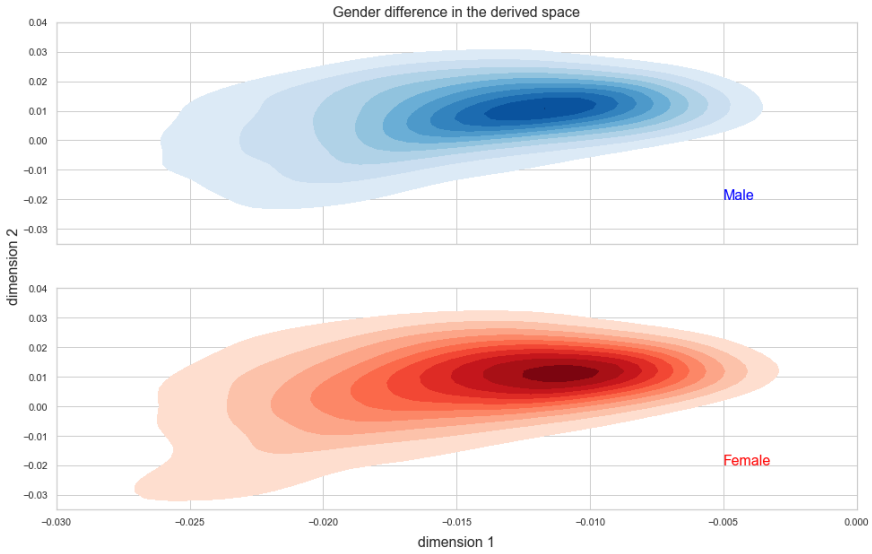


Figure 10: Gender and age difference in the relation between silence duration and following speech segment duration in the derived 2D space.

Age Group difference by age should also be expected as the literature has suggested. Six age groups are arbituarily defined: younger than or equal to 20, younger than or equal to 30, younger than or equal to 40, younger than or equal to 50, younger than or equal to 60, and older than 60 years of age.

Figure 11 plots the median values in the two latent dimensions in the derived space for each age group, controlled by gender. It can be observed that age does not seem to have a clear pattern among male speakers. However, there is a trend, although small, for speakers to move from the bottom to top along the second dimension among female speakers. Male and female speakers also appear to be roughly in two groups along the first dimension. Older groups, i.e., those older than 60 years of age, are likely to be outliers within the group of male or female speakers. Therefore with this observation, an interaction effect between age and gender on silent pause distribution can be expected. However, the effect size is also likely to be small, both across or within gender groups.

Years of education The years of education is binned into three categories in this analysis: those who received less than or equal to 12 years of education, who received less than or equal to 16

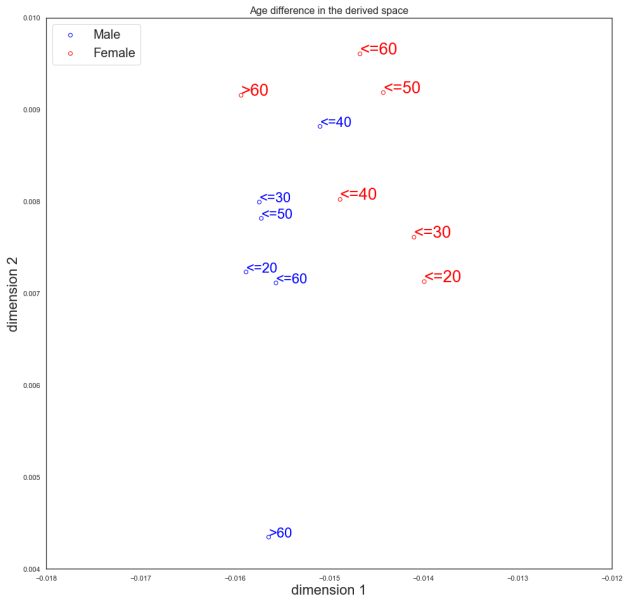


Figure 11: Gender and education difference in the relation between silence duration and following speech segment duration in the derived 2D space.

years of education, and who received more than 16 years of education. This categorization is on the one hand intended to correspond to the general treatment of education variable in socioeconomic studies: people who received at most high school education, who attended some college level education, and who have attended graduate or professional education. On the other hand, due to the structure of education system, the distribution of years of education in years approximates a step function. One caveat here is that some interaction effect of years of education and age should be expected, on top of the interaction between gender and education, as some participants in the corpus were still attending high school or college at the time of their contribution.

Figure 12 shows a clear interaction effect of years of education and gender, as among female speakers, there is a clear distinction between people who received at most college education and those who went to some graduate school. However, among male speakers, the group medians are more spread without clear pattern. Therefore, some interaction effect between education and gender can be expected, as well as the categorical effect of education level. The effect size, however, should also be expected to be small.

Dialect The last sociolinguistic variable to explore is dialect. Since it has been reported that dialectal variation does affect disfluencies in speech production both within North America (Kendall, 2009) and between American and British English (Tottie, 2014), and dialect dependent speech rate variation has also been recorded (Jacewicz, Fox, & Wei, 2010), it is worth asking if one's dialect has an effect on the silent pause distribution of their speech. In this study, I use the self reported place where participants of Fisher have been raised as the proxy for their dialectal background. In figure 13, the median values of the first two dimensions in the derived space for each state are plotted. States with too few observations (less than 10) are excluded from this plot due to the potential high variance. The state variable can be considered as more close to a randomly selected sample, thus reflecting the overall population distribution in North America.

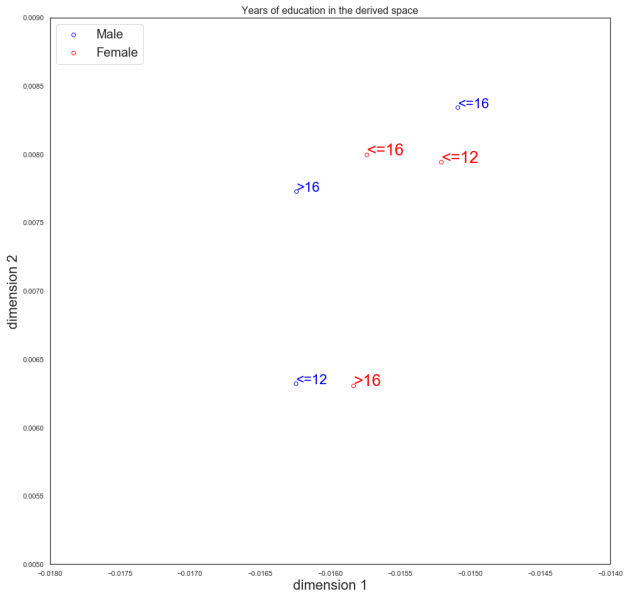


Figure 12: Density plots of the joint distribution of silence duration and following speech duration comparing medians of groups with different education background in the derived space. Education group is plotted conditioned on gender.

In figure 13, a distribution of the medians across the states seem to be randomly spread in the derived space, approximating a Gaussian distribution with uniform yet differing variances in both dimensions. Further examination of the distribution of the medians doesn't reveal any correspondence to the actual geographic relations among known dialect regions in North America. Therefore, there is evidence that pausing, or the temporal structure of telephone conversations, does not vary across English dialect regions in North America.

Topic The Fisher corpus contains conversations conducted under 40 different topics, which were provided by the data collector, but voluntarily selected by the participants. As mentioned in chapter 1, this data creation process may introduce biases from two perspectives: The topic selection process by data collectors was not intended to construct mutually exclusive topics; rather the goal was to facilitate the unrolling of conversations. Thus, topic categories were not intended to be orthogonal to each other, and overlaps between topics are unavoidable. On the other hand, the topic selection process by participants introduces the second layer of bias, such that some topics are selected more often than the others, and the kind of selected topic is apparently a function of individual speaker's personal preference. Hence, an interpretation of topic effect has to take these biases into consideration. Nevertheless, the relative large sample size in Fisher can somewhat mitigate the effects from these biases, and the results are still informative given these biases are properly considered.

In figure 14, each dot represents the median values of the joint distribution of conversations under the given topic. Although the overall shape of the distribution of the medians is approximately Gaussian, the variances appear to be non-uniform in the plotted dimensions. For example, mild evidence for two weakly separated clusters can be argued. Therefore considering some random effects from individual speakers, a main effect of conversation topic on silent pause distribution

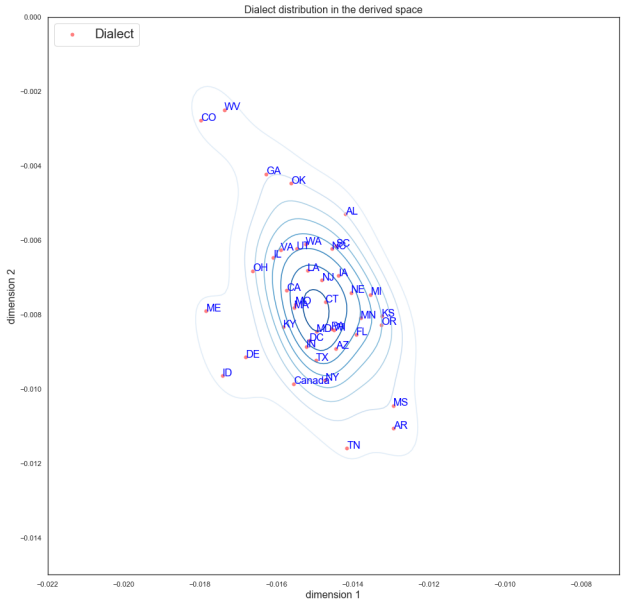


Figure 13: Density plots of the joint distribution of silence duration and following speech duration comparing medians of groups of self reported state where participants have been raised in the derived space.

can still be expected. One additional caveat is that this effect can be washed out by the existence of colinearity among some of the topics. Thus it is necessary to reduce the topic space to reach a better understanding of how certain topics behave in particular patterns.

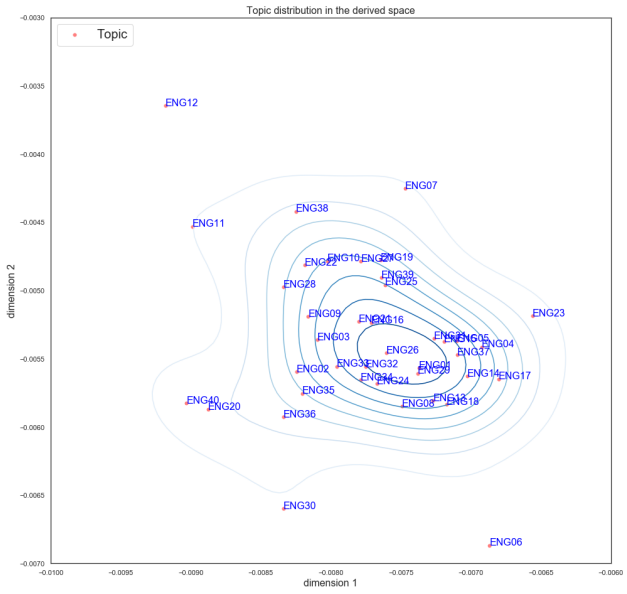


Figure 14: Density plots of the joint distribution of silence duration and following speech duration comparing medians of groups with different conversation topics in the derived space.

Variation across three conversations The last aspect of individual variation to address in this study is to what extend individual speakers would vary in terms of their silent pause distribution

across conversations? The discussion of this question will be deferred to the regression analysis, where repetition is modelled as a random slope to the full mix-effect model.

Summary So far, I have demonstrated possible effects from the sociolinguistic variables and conversation topic on silent pause variation. In the derived space, gender appears to have limited effect on silent pause variation, but this effect is still expected to be significant. Age and years of education have been shown to interact with gender, where systematic change related to age has been found among female speakers, and variation in response to years of education among male speakers has also been observed, although these effects are also expected to be small. However, the structure of topic and dialect distributions are less clear.

Regression analysis A linear mixed effects model has been fitted to test the hypotheses formulated through the exploratory data analysis presented above. In this regression, the response variable is the ratio of total silence duration over speech segments duration per speaker per conversation. The same threshold for determining silence, 150 ms, has been used. This measurement is an aggregate of the 2D density estimation for each speaker in each conversation projected to a single dimension, which can be conceptualized as a representation of the average amount of silence contained in one's speech in a given condition. Values in the derived space are not used mainly due to issues with interpretation, and the potential un-uniqueness of singular values. The explanatory variables include the factors explored in the derived space, plus the ID of the call, which represents the call repetitions, the interaction between Age and Gender, Education and Gender, the three-way interaction among Age, Education and Gender. The random effect is specified as a random intercept for speakers and a random slope for repetitions. Therefore individual difference in conversation repetitions is effectively considered in the model. The continuous variable *Education* is transformed to categorical representation, following the same strategy shown in the exploratory analysis above due to little variation within this variable. The model is fitted using the *lmer* function in the popular R package *lme4*.

Results of this regression analysis are reported in table 1. Under the view of a traditional F-statistics, all of the variables and interactions explored above are significant at least at $p < 0.05$. However, by this standard, call repetitions and the interaction between age and education are not significant in predicting the variation in the amount of silence within a segment of continuous conversational speech. The random slope of call repetition has very small variance ($\sigma = 0.0013$), suggesting that on average there is little variation across three conversations.

Looking at the effect sizes, male speakers on average have about 5 percent higher silence rate in their speech, while an increase by 1 year of age leads to a decrease of 0.02 percent of silent rate, when everything else is held equal. Thus ignoring all other factors, speakers across age groups do not vary much in the proportion of silence with regard to speech in their speech. This is also true when the interaction between age and gender is considered. The aggregated effect, when gender is controlled, is still only about 0.1 percent. The interaction effect of gender$\times$education has a relatively larger magnitude, where for male speakers, one who has completed some graduate or professional education on average has 2.2 percent lower silence rate compared to someone with only high school diploma or less. The rate is 1.4 percent lower compared to college graduates.

Table 1: Results from the mixed-effect analysis[a b]

	Df	Slope	Sum Sq	F value
Call ID	2	0.003	0.03	2.39
Topic	40	NA	0.56	2.33***
Sex(Male)	1	0.05	0.64	106.71***
Age	1	0.0002	0.12	20.73***
Educ	2	0.008	0.009	7.28***
State	51	NA	0.46	1.52*
Sex:Age	1	0.006	0.06	9.71**
Age:Educ	2	0.00	0.00	0.12
Sex:Educ	2	0.02	0.04	3.59*

[a] The *'s represents the significance level in a classical sense. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

[b] Education compares High School to Graduate School.

Discussion In sum, the regression analysis essentially confirms the observations made in the derived space. As predicted previously, the observed effects of age, gender, education, and their interactions are relatively very small, maybe with the only exception of gender and education among male speakers. However, it is harder to give exact interpretations for the other two variables with large number of categories. One possible fix might be to reduce the dimensionality of topic and dialect, then try to find associations between the content being discussed in reduced topic space and silence distribution, as well as between regional dialectal variation and silence distribution. I will defer these analyses to later stages of the dissertation work.

3.3 Filled pause

Similar to the discussion on silence distribution, I first compare between groups within each of the proposed variable, but separately for the two variants of filled pauses: *um* and *uh*. Then I present two regression models independently built for the two filler words. The measurement for filled pause in this section is the word frequency per 100 words for each individual in each conversation for the exploratory analysis.

Age and Gender Figure 15 plots the frequency of *um* and *uh* as a function of speaker age, controlled for gender. The regression lines were the mean estimators of a Generative Addative Model (GAM) fitted for each gender group under each condition using Poisson regression. The grey bands represents the 95 percent confidence band for the estimated mean in log space. The y-axis in each graph represents the log frequency of using *um* or *uh*, and x-axis represents age treated as a continuous variable.

The two graphs in figure 15 clearly indicate an opposite trend of change of the frequency of two fillers: for *um*, the relation between frequency and age is slightly negative linear, while for *uh* the relation is almost perfectly positive linear, except for the oldest and youngest males group. The observations for the oldest age group are relatively sparse, while it's not the case for the youngest. Female speakers also almost consistently have higher estimated frequencies for *um* across all ages,

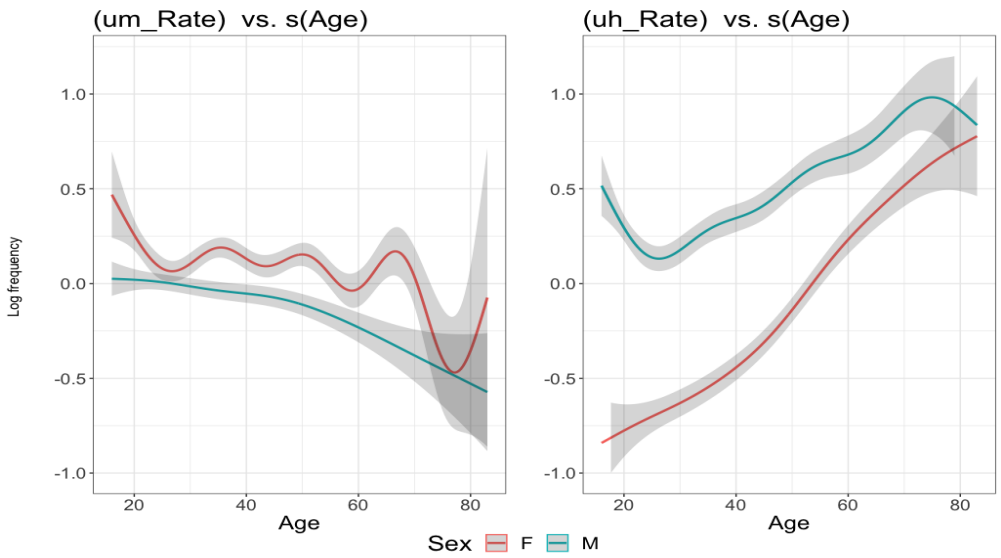


Figure 15: Log filled pause rates plotted as functions of age, grouped by speaker gender.

while lower for *uh*. This trend essentially replicates what have been reported in Fruehawld (2016) on a different data set, and in Wielding et al (2016) with a different treatment of the age and frequency variables. Interestingly, the decreasing rate of *uh* frequency, as age changes from older to younger, is higher among females, while the increasing rate is almost parallel between the two gender groups in the age range of about 225 to 60.

However, two details are worth mentioning. First, the trend for the change of filled pauses as a function of age is clearer and more stable for *uh* than *um*. The increase in popularity of *um* is actually not apparent among younger speakers (younger than 50 years of age). Therefore the seemingly increase in the popularity of *um* can be driven primarily by the low rate among older speakers. This can be problematic since the higher variance among older age groups may indicate the existence of unobserved heterogeneity. Second, the between gender frequency gap of *um* is also much narrower than the gap in *uh*. This difference in the trend of change with relation to gender may be aligned with the trend of a decreasing in the use of both *um* and *uh* over time as reported in Wieling et al. (2016) on Switchboard. The surface change or trade-off between the popularity of two variants of filled pauses can be argued to be driven by the higher variability in the use of *uh*, which is further attributable to other contextual or idiosyncratic factors that are not accounted for in the current and previous studies.

Education The relations between the frequency of *um* and *uh* and education, controlled by gender, are plotted in figure 16. The box plot for *uh* does not show clear variation across education level in both gender groups. However, speakers with only at most a high school diploma have slightly lower frequency of *um*, and this seems to be true regardless of gender. The small difference in frequency distribution also seems to be stable. On the other hand, speakers with post secondary education background tend to have higher *um* frequency, and the between gender difference appears to be much smaller compared to *uh*. One possible explanation to this difference is a potential interaction between age groups and education level: people that are older may be less

educated compared to younger age groups. Thus the apparent effect of education on *um* frequency may just be corroborated with the age effect that I have just shown.

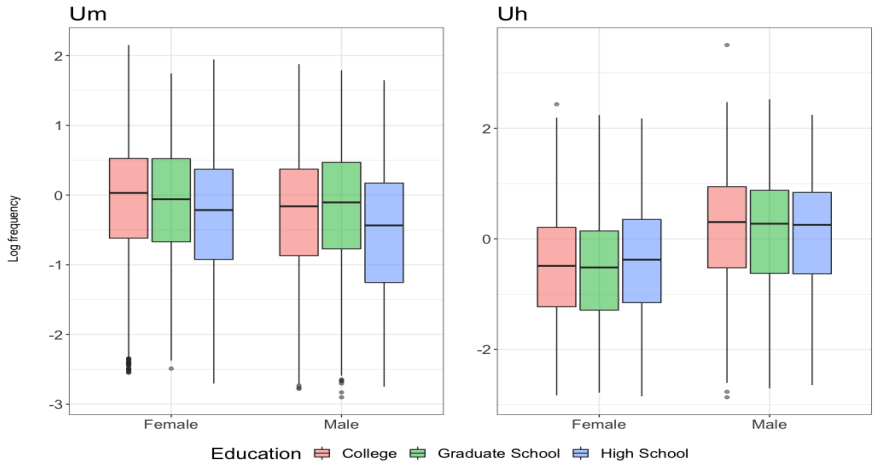
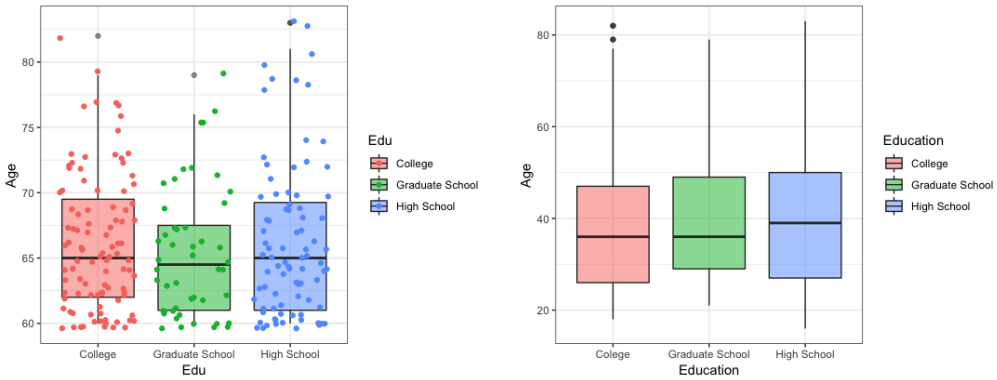


Figure 16: Box plot of *uh* (on the left) and *um* (on the right) frequency by education and gender.

Figure 17 plots age distribution by education groups: High School, Some college, and attended some Graduate or Professional education. If age is indeed the main factor underneath the observed difference between education levels, then one would expect an overall older age among High School graduates. In particular, since people older than 60 years of age have the highest *uh* frequency, one might expect more high school graduates in this age group as well.



(a) Speakers older than 60 years of age

(b) All speakers

Figure 17: Age distribution by education level

A very striking difference, however, is not found in figure 17 When people who are 60 or older are plotted separately or when the entire sample were plotted. Nevertheless, there does seem to be a slightly larger average age among people with only high school education, even though the variance among that group is also bigger. This difference is in fact confirmed through a one-way test of variance ($F = 22.878, p < 0.001$). However, the mean difference is only 3 years: 39 years of age for high school graduates, compared to 36 years of age for both of the other groups.

Dialect Variation across English dialects is plotted as the median *um-uh* frequency pair for each state against the joint contour of *um-uh* frequency pooled across the entire sample. As figure 18 suggests, on average, there is a trade-off relationship between *um* and *uh* frequency. This relation indicates that for each individual speaker, there is likely to be a preference when choosing the filler word in conversation. The peak density on this graph essentially follows the direction of x-axis, which indicates that there are more predominantly *um* users in this speaker sample. Larger variance in *um* frequency across speakers is also apparent from the graph.

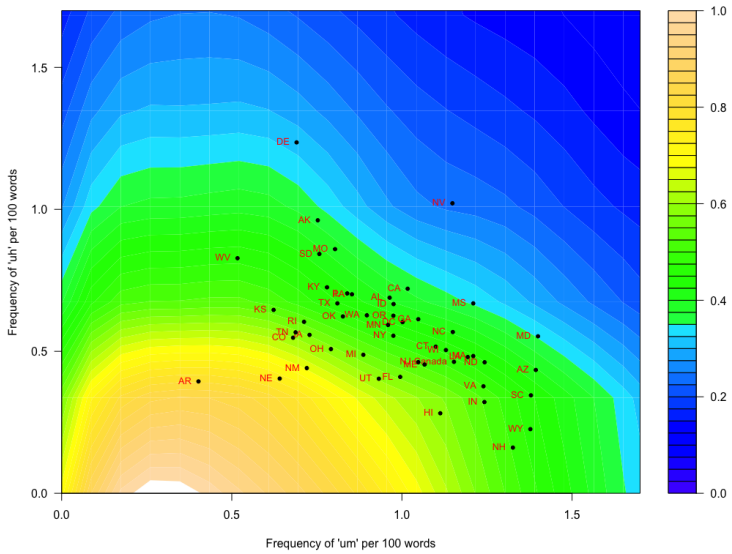


Figure 18: Contour plot of *um* frequency (x-axis) and *uh* frequency (y-axis) overlayed with speaker states.

By examining the distribution of states on this overall contour plot, it is found that the medians roughly follows the sample distribution and fail to show clear cluster structures. The distribution of states also appears to be at random, as no alignment between adjacent states and the acknowledged dialect regions can be identified. Thus, it is not likely that dialects would have a systematic effect on the choice of filled pause.

Topic Filled pause frequency of the two filler forms is plotted similarly across topics as what has been done with dialects. In figure 19, it can be observed that there is a larger variation along the *uh* frequency dimension of the median values. This suggests that there is greater variation in the frequency of *uh* across topics than the variation along the *um* dimension. Thus it is expected that topic mainly has an effect on the use of *uh*. But the existence of certain amount of variation across some topics along the *um* dimension, especially towards the bottom of the plot, suggests that some effect on *um* is also possible. The variance along *um* dimension across topics is also non-uniform. These observations can be better explained through modeling the conversation content in different topics, which will be deferred to later dissertation work.

Correlation matrices between pairs of conversation topics are derived to explore the (dis)similarities in terms of filled pause distribution across topics. The correlations are calculated based on the estimated density function from the frequency of filled pauses in each conversation in the given topic. The two types of filled pause are treated separately.

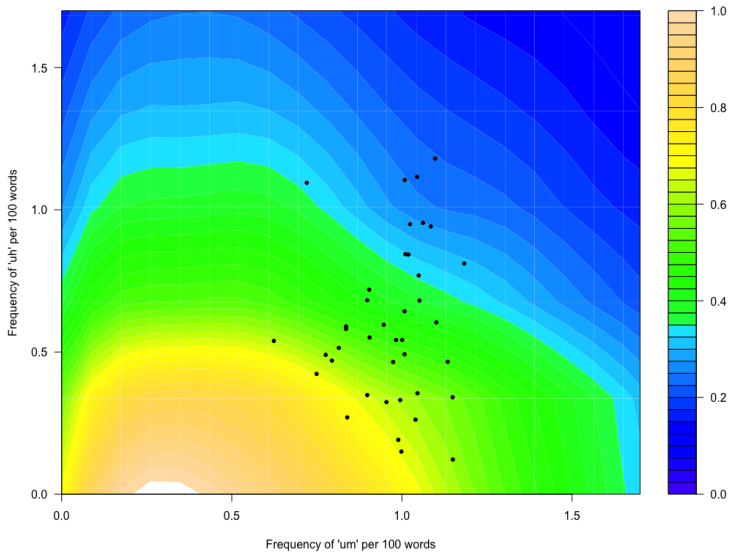


Figure 19: Contour plot of *um* frequency (x-axis) and *uh* frequency (y-axis) overlaid with conversation topics.

Figure 20 plots the two correlation matrices. In these plots, the lighter the color, the higher the correlation between topics. Figure 21 summaries the cumulative distribution of correlation scores separately for the two filled pauses. One apparent difference between the two forms of filled pause is that the overall pairwise correlation of filled pause frequency between across topics in the frequency distribution is lower for "uh" in comparison with "um". In fact, the frequency distribution of "um" doesn't seem to vary much across topics. The second observation is that the pairs of more (dis)similar topics also differ between the two forms of filled pauses. For example, in figure 20, topic 26, 27, 28, 29 and 30 have very low correlation score (less than 0.4) with topic 7 and 8 in terms of the frequency distribution of "uh". This difference may be attributable to the content of the actual conversation, as topics 26 to 31 are about more serious political or social issues such as Airport Security, Middle East and Foreign Relations, as well as Education and Family. On the other hand, topics 7 and 8 are on some hypothetical situations. However, the frequency distribution of "um" among these topics are highly correlated with correlation being over 0.9. One possible explanation to this difference is that the two forms of filled pause have different functions in the coordination of spontaneous conversations: "um" may more likely be strategically used as a device for message structuring purpose, while "uh" tends to signal the variation in speech production due to changes in the discourse.

Summary From the exploratory data analysis above, it can be expected that apart from dialect, other proposed explanatory variables, including age, gender, education and topic, will affect the choice of filled pauses in spontaneous conversations. The two variants of filled pause appear to have different sensitivity in response to the changes in the dimensions discussed above. In the rest of this section, I present two regression models, for *um* and *uh* independently, to address the question concerning what's the effect size of each of these variables, as well as the potential interactions. The two filled pauses are modelled separately, rather than jointly such as in Wieling

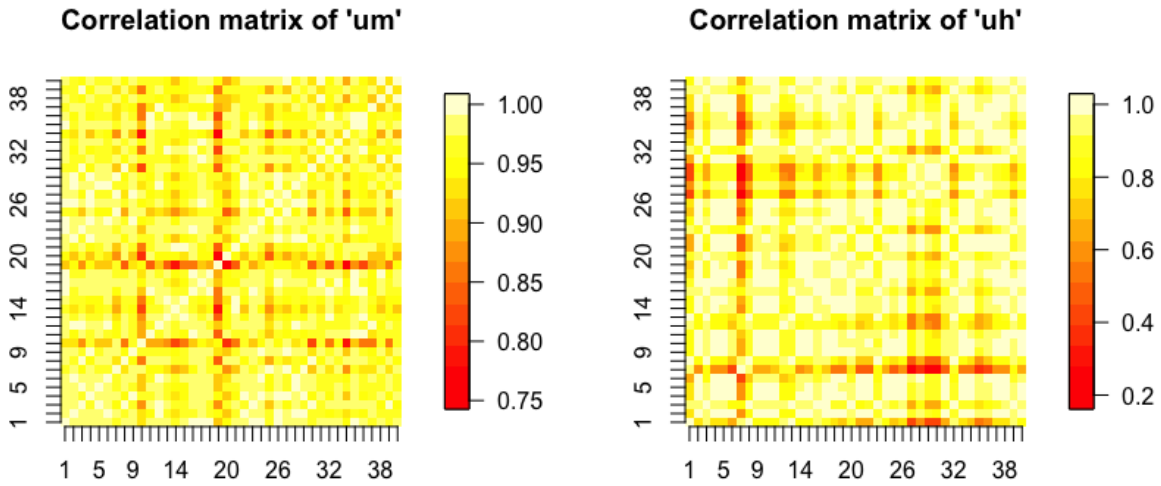


Figure 20: Correlation matrices of filled pause frequency between topics. Axes indicate topic numbers.

et al. (2016), is mainly for the existence of primarily *um*-ers and *uh*-ers in the sampled corpus, and the analysis in this study concerns each speaker as one independent observation, rather than pooling across speakers to estimate group means. Therefore if the relative frequency of *um* and *uh* were taken, no valid ratio would be found for many observations. The potential high co-linearity between *um* and *uh* frequency also deems considering one filler as the explanatory variable of the other inappropriate.

Regression models Here I report the results from two Poisson mixed-effect regression models fitted for *um* and *uh* independently. The Poisson model regresses the log frequency of each fillers onto the space defined by the exploratory variables explored above. Speaker's idiosyncratic behavior in response to the conversation task and the variation across repetitions of the task is modeled as the per-speaker random intersect and per speaker per repetition random slope. Thus within and cross speaker variation is accounted for in the model.

Table 2 summarizes the results for the filled pause *um*. As expected, except for the variable State, which is used to represent dialect variation, all other explored variables are estimated to have significant effect on the log frequency of *um*, if a threshold of $p = 0.05$ is chosen. As for the effect size, male speakers on average have 0.15 fewer *um* counts per 100 words of speech compared to female speakers, when everything else is held constant. For male speakers, an increase of 1 year of age corresponds to on average a decrease of about 0.06 count of *um* per 100 words of speech, while this decrease is about 0.005 count per 100 words for female. The sharper change among male speakers aligns with the steeper slope for male while more curvature for female observed in 15(a). In terms of education, comparing between those with graduate degree or higher, high school

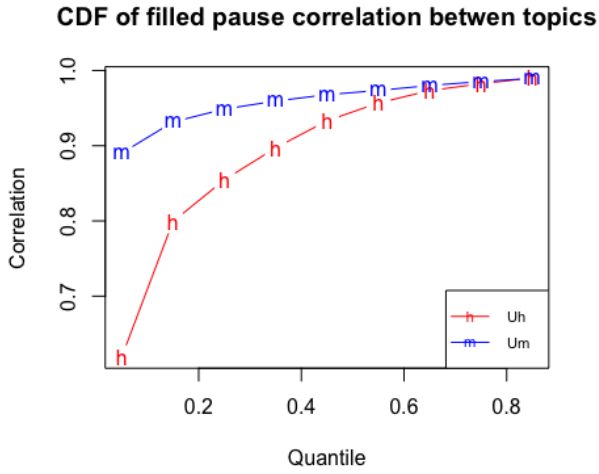


Figure 21: Quantile plot of the correlation scores.

graduates on average uses 0.28 less *um* per 100 words, when age, gender, topic, repetition and dialect are controlled. Thus the estimated effects of these variables reliably reflect the differences observed in the exploratory analysis step.

Table 2: Mixed-effect poisson regression on *um* frequency*

	Df	Slope	Sum Sq	F value
Repetition	2	-0.18	3.69	1.84
Topic	40	NA	411.41	10.29***
Sex (Male)	1	-0.09	58.34	58.34***
Age	1	-0.005	49.52	49.52***
Education	2	-0.28	50.18	25.09***
State	51	NA	72.49	1.42
Sex:Age	1	-0.05	4.45	4.45*

*Slope for *Repetition* compares the third call to first call. Slope for education compares Graduate to High School education.

In terms of the random effect, the estimated variance of slope is 0.082. Therefore, it appears that there is little variation within individual speakers across the three conversation repetitions. This shows that *um* can be a filler whose per speaker frequency subjects more to speaker factors than to conversational or contextual factors.

The second model performs the same mixed-effect Poisson regression on the frequency of *uh*, with same model specification as the previous model. As summarized in table 3, the model confirms the initial observations on the relations between each explanatory variable and *uh* frequency. In addition, the variable *Repetition* appears to be significant, which suggests that on average there is more cross-repetition variation in the use of *uh* for a given speaker.

An examination of the effect size is the following. Compared between male and female, a given male speaker on average uses 0.4 more *uh* per 100 words of speech, when everything else is held constant. In terms of age effect, for male speaker, an increase by 1 year of age corresponds to 0.01 more *uh* per 100 words, while this difference is about 0.02 per 100 words for females. Thus the trend observed in figure 15(b) is also truthfully reflected in this model. As for the random effect, the estimated variance is about 0.48, which is substantially larger than the estimate for *um*. Therefore it can be hypothesized that the use of *uh* is more sensible to contextual variables, such as the familiarity of task, the identity of the interlocutor, or the nature of the conversation topic, to name just a few.

Table 3: Mixed-effect poisson regression on *uh* frequency*

	Df	Slope	Sum Sq	F value
Repetition	2	-0.01	450.96	225.48***
Topic	40	NA	1955.92	48.90***
Sex(Male)	1	1.35	462.42	462.42***
Age	1	0.02	236.07	236.07***
Education	2	-0.06	2.48	1.24
State	51	NA	69.23	1.36
Sex:Age	1	-0.01	31.97	31.97***

*Slope for *Repetition* compares the third call to first call. Slope for education compares Graduate to High School education.

Accommodation between interlocutors A follow up question in response to the observed interactions between sociolinguistic and discourse variables to ask is what is the role played by the interlocutor in the distribution of the two filler forms? In other words, what is the accommodation effect on filled pause distribution in telephone conversations? The effect of accommodation, or entrainment, has been studied in the past from the perspective of communication mode such as differences among human-human, human-machine, and dialogue versus monologue (Oviatt, 1995), the role played by the speaker in a communication task (Bortfeld et al., 2001), and in supreme court oral arguments (Beňuš, Levitan, & Hirschberg, 2012) and converstions in a game setting (Beňuš, Gravano, & Hirschberg, 2011). To investigate the temporal aspect of turn taking in spontaneous conversations, Ten Bosch, Oostdijk, and Boves (2005) showed that durations of between-turn pauses made by speakers in a dyad are statistically related, and gender appears to have an effect on the temporal aspect of turn-taking: male-male conversations tend to have more inter-turn overlaps than female-female converstions. Oviatt, Darves, and Coulston (2004) suggested that, in a study of accommodation in human-machine communication among children, the largest adaptation comes from the pausing structure and acoustics of utterances. In human-human communications, converging patterns of pausing structure and the use of filler words or other signalling words contribute to the coordination of conversation and establishment of commonground (Beňuš et al., 2011, 2012).

To evaluate the effect of accommodation on the frequency distribution of the two filled pauses, I compare within each pair of speakers (speaker a and speaker b). The most obvious dimension for this comparison is gender: among all the variables considered so far, gender is the easiest identifiable covariate. Therefore I compare the frequencies in three different groups: male-male, female-

Table 4: Correlation of filled pause frequency in conversation between speakers controlled for gender

Filled pause type	Male-male	Female-female	Female-male
Um	0.113	0.103	0.056
Uh	0.368	0.205	0.192

female and female-male conversations. The comparisons are carried out using a sub-sample in which both sides of a conversation are present in the selected full sample. This yields a sample consisting of 685 male-male conversations, 885 female-female conversations, and 675 female-male conversations. Correlations between speaker a and speaker b in each group are reported in table 4. It is obvious from the correlation table that there is stronger correlation in the frequency of "uh" between interlocutors in a conversation than "um", and this trend holds across all three conditions.

Discussion In this section, I have explored the effects of soioeconomic variables, such as age, gender, years of education, dialect, and conversation topic on the use of filled pauses. The exploratory analysis of the frequency of *um* and *uh* first confirms the observation using other corpus data or statistic methods (Fruehwald, 2016; Wieling et al., 2016) that there is a trend for more *um* and less *uh* usage among younger speakers, which has been postulated as a change in progress led by female speakers not only in American English, but also across several Germanic languages (Wieling et al., 2016). It is also found that the rate of increase in *um* frequency is actually very slow especially among younger speakers, while the drop in *uh* frequency is steeper among female speakers compared to males. Amount of education and topic have also been suggested to affect filler frequency. However, the effect of education is mainly on the use of *um*, while the major effect by topic is on the use of *uh*. Little dialect effect has been suggested as well. Two Poisson mixed-effect regressions are then reported in support of the initial observation.

In addition to providing further evidence for a potential change in progress of filler words, this analysis also shows that the loss of popularity in *uh* in return for more frequency of *um* is not a parallel process, in terms of both the pattern of trade-off within each gender group, and across genders. This asymmetric trade-off may be a result of a different sensitivity to contextual variation for the two fillers: as suggested by figure 21 and 20, *uh* exhibits higher variability across conversation topics than *um*. Higher degree of accommodation effect has also been found in the use of *uh*, seen from the higher correlation of *uh* frequency between interlocutors. Furthermore, the regression models also offer evidence for larger expected between repetition variation for an individual speaker for *uh* than for *um*. Thus, different forces may exert different effects on the direction and magnitude of change for the two fillers. Therefore a detailed examination of the factor space and how they influence speaker's decision in choosing between the filler words is warranted. The first step will be to understand how the nature of different topics, such as the content of speaker's speech under the provided topic, affect the frequency distribution of the two filler words.

Table 5: Classification of repetitions in this study

Type	Description	Examples*
Full word repetition	Repeating complete word or phrase 2 or more times	it's a different context but **if if** it's something... and **it's it's it's** a big question... oh yeah **she loves she loves** the cats...
Partial-word repetition	Part of a word is repeated before the full word is delivered	if it's like in a **s- sexual** thing i think it that's where i draw the line... i'm not sure if **the qu- the question** i think says...
Other repetition	Partial phrase repetition, where the last word is replaced, or repetitions involving intervening filled pauses, or other situations involving repeating part of a phrase or word that not covered by the other two classes	instead of doing that **they'll play sil- they'll play politics** and say... **that's ah that's** true although it can be hard in our family sometimes... it is and **it is fascinat- it's no less fascinating** to watch...

* Examples are from Fisher corpus.

3.4 Repetition and repair

The analyses on fluent repetitions are carried out in a type-dependent fashion. Unlike the clear surface distinction of two forms of filled pauses, the form of repetitions is subject to more variability. In this study, I primarily divide repetitions into three groups: *full word repetition*, *partial word repetition*, and *other repetition*. As the names suggest, the distinction between the first two categories is mainly whether the repeated segments are full words or partial words. Other repetition, on the other hand, refers to the repetitions which do not belong to the first category. Examples of other repetition include partial phrase repetitions and repetitions that include intervening filled pauses or other hesitation markers. Although the practice of creating a garbage category in the building of disfluency classification systems is criticized by Shriberg (1994), this classification is mainly a compromise to the lack of detailed annotation in the data that I'm working with. The "garbage" category will be further discerned in chapter 4 where analysis based on careful hand annotated smaller samples is discussed. Examples of each type of repetitions are given in table 5.

In this section, I first describe the methodology for identifying repetitions from my speech sample. Then I define the parameter space within which individual variations are measured. Finally I present type dependent analyses of variations in repetitions.

3.4.1 Identifying repetitions from large speech corpora

Repeated words have been identified using a simple method based on the suffix-tree algorithm. Turns in the corpus were initially represented as lists of words. A search window size N was firstly defined so that repetitions involving a phrase of up to N words can be captured. Before applying the suffix tree algorithm, each turn was transformed into a single string of letters. Then a look-up table for each turn was made, which contains mappings from the word index in the original list representation to word-initial letter index in the transformed string. A suffix tree was constructed for each turn based on the new representation. Exhaustive searches were then performed starting from the left edge of the string using the suffix tree. A matching string was constructed using the words incrementally from the current window in the current turn. The search stopped either when one or more matches are found or all the words in the current window have been added to the matching string. Then the left edge of the window moved to the next word. Only the matched string that immediately follows the right edge of the matching string was returned as the repetition of the matching string. A tolerance value T was also introduced, to account for repetitions that involve contractions, replacement of the last word, or filled pauses. Therefore the returned repetitions include the repeated words plus T letters. The matching and matched strings were combined and translated back to an ordered word list through the look up table. If T is smaller than the length last word to be included in the returned word list, the last word is preserved.

Classification of repetitions was achieved by examining the returned repeated word lists using two straightforward rules: full word repetitions contain only complete words and the smallest repeated units. Partial word repetitions contain word fragments (indicated by "-" in the transcription) at the end of the smallest repeated units. All other repetitions are classified as *other repetitions*. This procedure identified 104,658 full word repetition instances, 24,457 partial word repetition instances, and 53,313 other repetition instances. This simple search algorithm, however, should not be expected to identify all the instances of *other repetition* from the corpus, due to its high variability. Nevertheless, it should be able to cover the majority of full word repetitions and partial word repetitions. The performance of this identification mechanism is evaluated by measuring the recall of a random sample of 500 instances from each repetition category. The evaluation results are reported in table 6.

Table 6: Performance check of the proposed repetition identification method based on 500 samples from each class

Repetition type	Full word repetition	Partial word repetition	Other repetition
Instances	104,685	24,457	53,313
Recall	0.894	0.914	0.634

An error analysis suggests that false positives in identified full word repetitions are of two sorts: emphatic repetitions and floor holding repetitions, where the repeated phrases are filler words such as *right right* and *yeah yeah*. For partial word repetitions, errors occur when the identified partial word repetition is in fact part of some more complex repair structure, such as those involving insertion and deletion. Nevertheless, false positives in these two types of repetitions can be relatively easily identified. The recall indicates that this automated method can generate a relatively large sample of accurate full and partial word repetitions to work with. However, given the challenge

presented in accurate and efficient unsupervised identification of other repetitions, especially from large collection of speech, the primary focus will be on full word and partial word repetitions in this section. Other repetitions will be discussed when detailed annotations are made available from SCOTUS.

3.4.2 The parameter space

The parameters that are considered for measuring variation in repetitions consist two parts: one is about the repetition itself, and the other the lexical and phrasal context. Two crucial questions to be addressed in this measure is the following. First, for a given repeated word, how likely it is to observe it in non-repeated speech compared to repetition? And what is the likelihood of observing repetitions in one's speech? With this consideration, I estimate the odds of observing a word in non-repeating context against the same word occurring in repetition, and the frequency of repetitions in one's total speech as measures of the repetition itself. Non-repeating context here refers to both turns that do not contain repetitions and turns that to not contain repetitions involving the same lexical items. The odds can be calculated both by lexical categories and by word lemma. Therefore, the higher the odds value, the more likely that a given lexical item or category will appear in non-repeating than repeating context. For the contextual parameters, both lexical context and phrasal context are considered. I include the term-frequency vector of the first content word after a repetition, the likelihood of observing a filled pause after a repetition, and the length of speech segment measured in the number of words following a repetition, to characterize the potential variability of repetitions. Table 7 summarizes these parameters and how they are derived.

3.4.3 Distribution in full word repetition by type

In this subsection, I report the type-dependent variation in fluent repetitions. Specifically, I first look at variations of the odds among function word classes, including prepositions, pronouns, articles, conjunctions, auxililary verbs, relative pronouns and demonstratives, as well as other function words such as quantifiers and some adverbials. As reviewed above, function words are expected to be most heavily repeated, while the absolute frequency of occurrence, relative to all the instances of repetitions, does vary (Foster, 2010). Here I first ask if the high relative repetition frequency of certain lexical categories can be translated into higher chance of being repeated compared to the repetitions of other lexical categories, and how individuals vary in their speech, as a function of age, gender, education, conversation topic and the interlocutor. Following the same pipe line in the analysis of filled pauses, after reporting the summary statistics of type and context distributions of full word repetitions, I will elaborate the observations from exploratory analyses with more rigorous quantitative modeling in the next section.

In this preliminary summary, I only report results among single word repetitions that involving repeating the word twice. More complex repetition structures, such as multi-word phrases and contractions will be deferred to the future work in this dissertation. A list of function words considered in each lexical category can be found in Appendix B.

Overall odds distribution An implicit assumption in the literature on repetitions is that the distribution of repetitions is not random. However, a direct test of this hypothesis has not been clearly

Table 7: The parameter space for quantifying repetitions in fluent speech

Parameter type	Feature	Derivation	Shape
Repetition feature	Odds of being fluent	For $word_i$ in all repeated words in the given condition, $$Odds = \frac{\sum word_i \ elsewhere}{\sum \ repeated \ word_i}$$	float
	Frequency of repetition	$$\frac{Number \ of \ repetition}{Total \ number \ of \ words}$$	float
Context feature	Term-frequency vector	A word vector of the first content word following repetitions[i]	$1 \times \lvert V \rvert$[ii] vector
	Frequency of observing a filled pause after repetition	$$\frac{Number \ of \ filled \ pause}{Number \ of \ repetitions}$$	float
	Length of the following speech segments	Number of words after repetition until the first silent pause of 250 ms or longer	integer

[i] Stop-word list is derived through inspecting the vocabulary of the corpus. It contains all the function words, plus filler words and floor holding words.
[ii] $|V|$ represents the size of vocabulary of content words that follow repetitions.

offered: the reported frequencies of repetitions are always in their relative frequencies of the vocabulary, but not with regard to the same repeated items in non-repeating context. As suggested in Blacfkmer and Mitton (1991)'s study on the timing of repair gaps, the cut-off-to-repair times may be too fast to fit into a self-monitoring model as proposed in Levelt (1983, 1989). Therefore, the first step for an account on the variation of repetition is to demonstrate the non-uniform distribution of repetition across different lexical categories. Table 8 summarizes the overall distribution of odds across 8 function word lexical categories. In the table, *odds* refers to the odds of seeing the repeated word in non-repeated context against in repeated context, and *popularity* indicates the fraction of the speakers whose speech contains repetition of words in the given lexical category. *Frequency* records the relative category frequency of the repeated pattern per 1000 words.

Table 8 first shows a large variation of the odds of being in non-repeating turns across categories of function words. The highest odds is actually about six times higher than the lowest. The second observation is that for words in a given lexical category, the chance of being repeated in a 15-minute conversational speech also varies. Almost every speaker had some repetitions of pronouns, while the majority of them repeated conjunctions, prepositions and articles as well. However,

Table 8: Odds and frequency distribution by lexical category and the corresponding fraction of speakers whose speech contains repetitions within the category

	prep.	pro.	art.	demon.	aux.	conj.	rel. pro.	other
Odds	108.071	35.54	57.34	52.89	193.20	43.08	44.48	176.10
Popularity	0.701	0.979	0.792	0.638	0.577	0.853	0.350	0.178
Frequency	75.14	106.88	48.58	25.33	75.89	63.18	7.82	13.03

relatively few speakers repeated relative pronouns and other function words, such as quantifiers, adverbs, negations, etc. On the other hand, there doesn't seem to be a correlation between the odds of being non-repeating and popularity. Although the most likely repeated category, pronoun, is also the most popular category for repetition, the more likely category such as relative pronoun, is not actually very popular as repetitions. Thus, the distributional statistics seems to suggest that the distribution of repetitions, at least across function word categories, is non-random. Variations exist both in terms of the likelihood of observing words in one lexical category in repeated or non-repeated contexts, and of the choice made by individual speakers. With this ground established, I can then explore how to explain such variation.

Age and gender Odds variations by age and gender across the 8 lexical categories, as well as over the entire sample, are plotted in figure 22. The vertical axis in each plot indicates the log odds of being in non-repeating context against repeating context for the category. Curves are fitted through generalized quasi-poisson regression. The overall relation between the odds of function words is plotted in the bottom right corner of figure 22. A slight downward trend can be argued for female speakers, which suggests that older females are slightly more likely to repeat function words. However, the pattern for male speakers is more complex, with an apparent positive relation among speakers younger than 30 years old, but largely uniform afterwards. Among the 8 lexical categories considered in this study, some more consistent patterns can be identified from these plots. For prepositions, there is a tendency of higher odds for older male speakers in the age up to 60 years old, with an odds ratio of about 1.5 estimated from the mean. Although there is a slight downward trend after 60 years of age, the variance also becomes very large. In terms of pronoun, a clear downward trend can be observed among female speakers. The average odds ratio between the youngest and oldest group is about 1.45 estimated from the mean. A steady increase in the log odds can also be found for both male and female speakers in auxiliary verbs. The odds ratio between the oldest and youngest groups is about 1.28 for male speakers, and 1.35 for female speakers. Slight negative correlation between age and the log odds of conjunctions may also be argued. There isn't a clear trend for both gender groups in other lexical categories.

In general, for most function word categories, there isn't a clear relation between the odds of repeated words occurring in non-repeating context and gender or age for most function word categories. For some, a weak effect can be argued, and they can either be gender sensitive, such as preposition and pronoun, or gender independent, such as auxiliary verbs.

Education Figure 23 plots the odds distribution by three education groups across the 8 lexical categories and the entire sample space. For all the lexical categories considered, as well as the

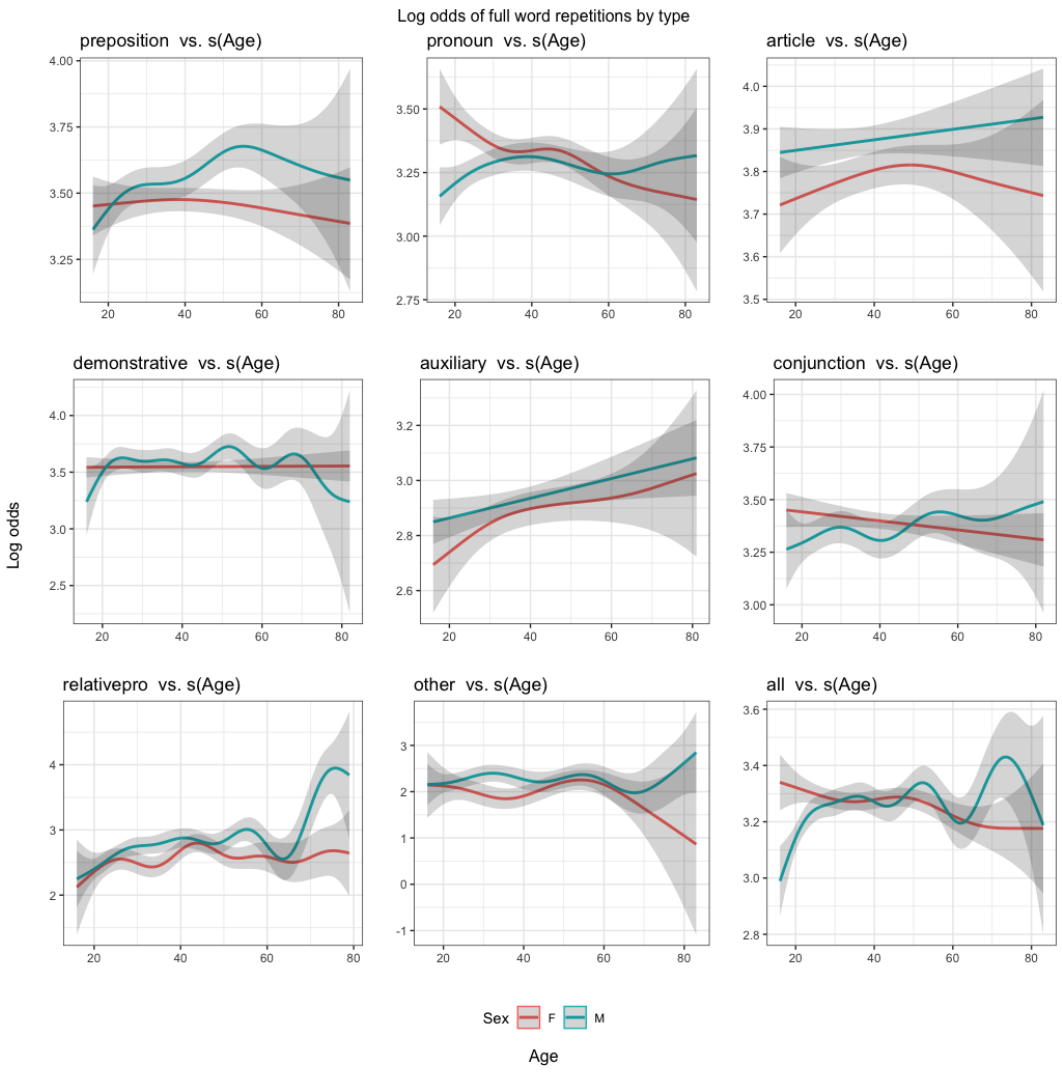


Figure 22: The relation between odds and age by lexical category

overall distribution, there doesn't seem to be a strong effect of education on the odds distributions.

Topics Variation across topics is measured as the odds within conversations produced under particular topics. The variations are displayed as bar plots by 8 lexical category and overall odds across topics shown in figure 24. In each plot, the horizontal axis represents the numerical number for the 40 topics assigned to speakers in Fisher, and the vertical axis indicates the odds calculated from repetitions observed in all the speech under the same topic. The bottom right plot shows the odds distribution across all function word categories.

The overall odds distribution is near uniform, which suggests that the variation across topics is rather limited. However, there are a few topics whose odds are apparently smaller than the others, such as topic number 5 through 7, as well as topic 17, 19 and 20. Although the topic labels for 5 through 7, as well as among topics 17, 19 and 20, can be argued as more similar compared to others, it still requires a more detailed examination of the word distribution to build the connection between topic and repetition variation. Therefore there is potential evidence for

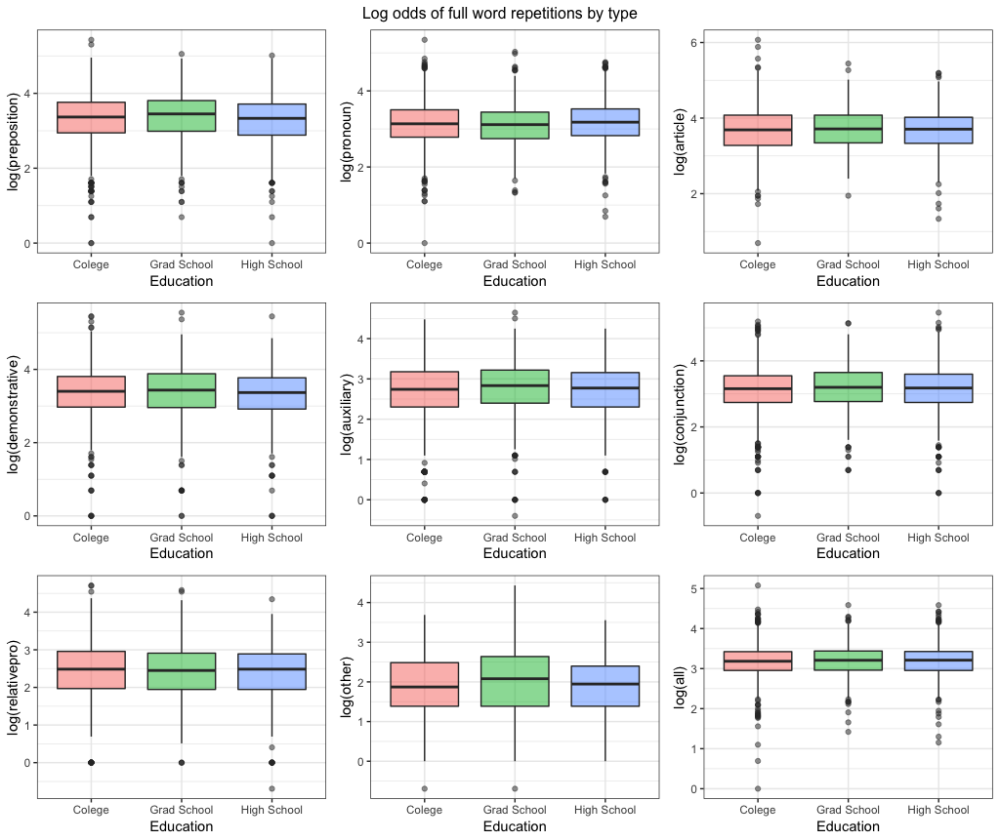


Figure 23: Box plots of the relations between odds and education by lexical category

systematic variation in the likelihood of repeating function words by topic.

More variability can be observed in type-dependent plots. The first observation is that the distribution of some function word categories better approximate uniform than others. For example, the distributions of relative pronoun, other function word categories and prepositions exhibit higher degree of variability across topics compared to other lexical categories. Thus, it can be hypothesized that conversation topics will have effects on how people might repeat function words. In particular, the lexical categories that are related to more complex sentence structures, such as relative pronouns and prepositions, tend to show greater variability, so do the categories that require more processing, such as quantifiers and negations. This increase in topic-related likelihood variation may reflect the processing load entailed by different conversation topics, and is realized in connection with the demand for more complex utterances. Again, more grounded hypotheses on the relation between topic and repetition odds can only be proposed after careful examination of the word distribution within each topic. This part will be further elaborated in the dissertation.

Future work: Accommodation and comparisons between function and content words, as well as between single word and multi-word In addition to the need for careful examination of the textual content across topics, the preliminary results presented above have left two interesting questions unaddressed, which will be discussed in the proposed work. In the first part, I will try to understand how one speaker is accommodating his or her interlocutors in terms of repetitions. The two apparent possibilities are that there are people who are more prone to adaptation, and who are

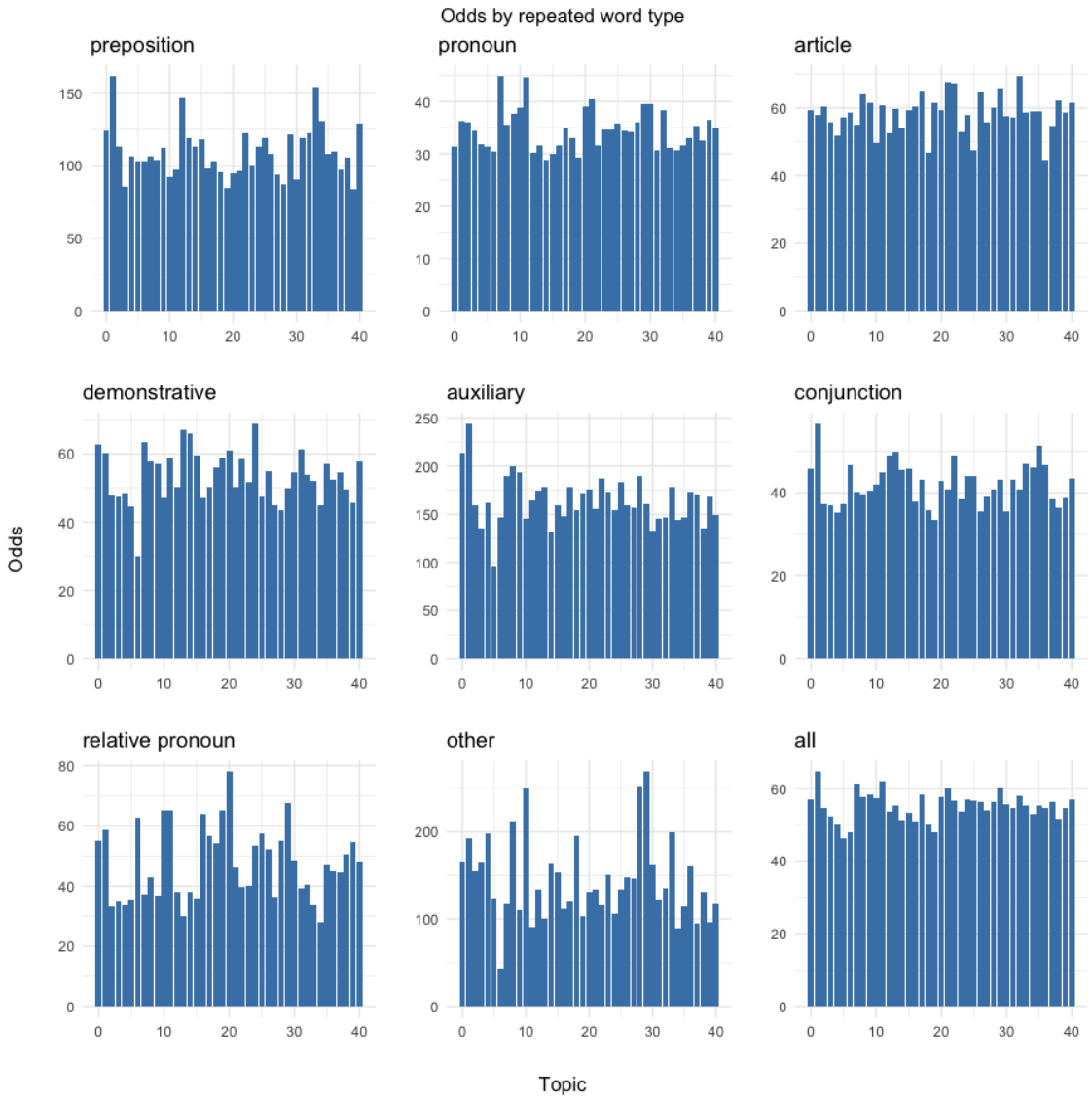


Figure 24: Bar plots of the odds across topics by lexical category

not. This difference may be related to other sociolinguistic or topic variables. The second aspect is a comparison between function words and content words, as well as comparing between single word repetitions and multi-word repetitions. The proposed study will largely follow the convention established so far.

3.4.4 Distribution of full word repetition contexts

In this section, I propose to analyze the distributional properties of both lexical and phrasal contexts of full word repetitions. The explanatory variables of interest will be the same as the analyses presented thus far. The questions to be explored include: Will the sociolinguistic variables, including age, gender and education, affect the lexical contexts in which repetitions occur? How function words and content words may differ in response to these factors? And what is the effect of conversation topic on the lexical and phrasal contexts that induce repetitions? Although pre-

vious research has examined the effects of lexical and phrasal contexts on disfluency production, such as in Holmes (1988); Clark and Wasow (1998); Arnold, Fagnano, and Tanenhaus (2003), explicit modeling of the contextual variation has not been carried out. This aspect is of particular interest to the study of repetitions due to the heavy reliance on the linguistic context in accounting for patterns of variation in repetition. In the proposed work, the lexical contexts will be represented combining three perspectives: the standard term-frequency word vectors of the following word, after filtering out stop words, the distribution of the first word after repetition without filtering stop words, and the likelihood of observing a filled pause following a repetition. Phrasal contexts will first be represented naively as the length of speech segments, measured in the number of words, following the repetition. Distributional patterns will be identified through applying some dimensionality reduction technique, such as SVD, on the initial high dimensional representations of contexts. It is hoped that an account on the variation in the linguistic context of repetitions can provide the necessary information toward a more complete explanation on why and how people repeat in spontaneous speech.

3.4.5 Distribution of partial word repetitions

In this section, I will report the results from analyzing partial word repetitions identified through the automatic method. I propose to follow the same suit as analyzing full word repetitions. However, before starting the actual analyses, a proper categorization of repeated patterns should be established. This categorization method will consult both the lexical category of the repeated partial words, as well as the frequency rank of the repeated patterns. The analyses will also be carried out in two steps: an analysis of repeated patterns themselves, and an exploration of the lexical and phrasal contexts of the repeated patterns. Results will be reported following in the same format as before.

3.5 Conclusion

In chapter three, I reported the interim results from analyses of effects of sociolinguistic variables and topic on silent and filled pause variation, considering the potential within and cross speaker variation in tandem. The results in this chapter both replicated some previous observations on the correlation between filled pause distribution and sociolinguistic variables, and supplemented new information considering the effect of topic and accommodation between interlocutors. The discussion on full word repetition also for the first time brought an under-explored disfluency phenomenon into this picture. Continuing the discussion in section 3.4, in which I reported the preliminary findings of full word repetitions with a special focus on function words, I proposed further directions for investigation into variation in partial word repetitions, and the contextual variables for both full and partial word repetitions that potentially correlate with sociolinguistic and topic variables. It is hoped that this chapter will contribute to a full picture of how disfluencies are interacting with extra-linguistic and broad contextual variables.

4 The micro-analysis of disfluencies

The focus of this chapter is to examine the linguistic contexts, such as syntactic, semantic, and prosodic variables that are supposed to be related to variations in disfluencies. As reviewed earlier, one major drive for examining the semantic, syntactic and prosodic contexts of disfluencies is from the point of the cognitive process behind speech production. Knowledge of variations in relation to the immediate linguistic context of disfluencies is also crucial for applications that involve processing spontaneous speech. I will first review the findings from both of these perspectives, then propose a study to describe and explain how disfluencies covary with linguistic contexts. Potential theoretical and practical implications from results of the proposed study will be discussed in the end.

4.1 Background

The immediate syntactic and lexical environment has been long known to affect the distribution of disfluencies. These variables include the length of sentence and syntactic complexity (F. Ferreira, 1991; Shriberg, 1994), predictability of the adjacent lexical items (Beattie & Butterworth, 1979; Tannenbaum et al., 1965; Shriberg & Stolcke, 1996), the strength of syntactic boundary (Holmes, 1988; Watanabe, Kashiwagi, & Maekawa, 2015), and the prosody of filled pause in comparison with the prosody of surrounding phrases (Shriberg & Lickley, 1993; Nakatani & Hirschberg, 1994). Disfluencies are also more likely to be observed at the initial rather than medial position of a sentence, and this trend has been shown to interact with sentence length (Shriberg, 1994), although the nature of the speaking task does influence this distribution. Following up on Arnold, Losongco, Wasow, and Ginstrom (2000)'s report that disfluencies tend to correlate with word-order choice in dative sentences, Tily et al. (2009) investigated the effect of syntactic probability on the acoustics and fluency of speech in a case study of the dative sentences in Switchboard. Their result suggests that more probable NP-PP and NP-PP combinations are less likely to contain disfluencies. However, type-dependent variation in disfluency was not discussed in their statistical model.

Why cognitive load may be a factor behind disfluencies? Because most disfluencies occur at the beginning of an utterance or phrase, and occur more often at the beginning of longer sentences or phrases (Oviatt, 1995; Shriberg & Stolcke, 1996). Disfluency rate also varies across different topics, either in constructed scenarios, such as something set to be unfamiliar (Bortfeld et al., 2001), or university lectures with different subject matter (Schachter et al., 1991; Moniz et al., 2014), in which natureal sciences used the least amount while humanities used the most in terms of filled pauses. This variation seems to be explained by the content of the lectures, as measured by the number of terms and jargon used in the lectures (Schachter et al., 1991).

At the lexical level, evidence is also supporting the assumption that cognitive load is a crucial predictor of disfluency. The rate of disfluency has been shown to correlate with word frequency (Maclay & Osgood, 1959; Levelt, 1983). Context probability has also been shown to relate to disfluency (Beattie & Butterworth, 1979; Tannenbaum et al., 1965). It is further arged by Beattie and Butterworth (1979) that word frequency plays a less important role than contextual probability of word forms, which the speaker might have the awareness to make choices. Higher rate of disfluency in lower contextual probability might just reflect this decision process. In a more recent study, Harmon and Kapatsinski (2015) looked at single and two-word repetition in Switchboard preceding main verbs and nouns, where both forward and backward transition probability

between words in the context, lexical frequency were modeled in a logistic regression model. The results suggest that speed of lexical access is negatively correlates with the length of repetition. The main determinants of lexical access speed also differ for verbs and nouns. Longer disfluencies before verbs appear to be due to significant paradigmatic competition from semantically similar verbs, while disfluencies occur when the noun is relatively unpredictable given the preceding context. From the perspective of speech perception, an eye-tracking experiment (Arnold et al., 2003) demonstrated that disfluent instructions tend to trigger an interpretation of upcoming new information in the discourse. Thus speech disfluencies also play a role in coordinating the information flow in human-human communication in both speaker's intention and listender's expectation.

Repetitions and filled pauses may correspond to different aspects of the cognitive processing of language. Oomen and Postma (2001) reported an experiment, in which speech rate was controlled for by manipulating the speed of a moving dot on screen, whose path was asked to describe. Although there was apparent increase in cognitive load in fast speech rate condition, it was only the number of repetitions that had increased, but not filled pauses. They attributed this observation to what Blacfkmer and Mitton (1991) explained as "autonomous restart capacity", where the repetition is the phonetic response to increase in cognitive load and a sign that the articulatory process is not able to keep up with the cognitive demand. In a series of experiment in Schnadt and Corley (2006) designed to further explore the question of choice, they found that hesitation disfluencies, including prolongation and filled pauses, increased when more choice were available, while the number of filled pauses didn't change when the number of choice was reduced but replaced with hard-to-name items as the way to increase cognitive load. Therefore they conclude that disfluencies are not automatic: filled pauses can be used as expressive strategies. This view has also been suggested in Smith and Clark (1993); Brennan and Williams (1995); Swerts and Krahmer (2005). Clark and Tree (2002) even claimed that the communicative function of filled pauses is encoded as part of the collateral message, and should be considered as within one's vocabulary.

However, such a view is challenged by OConnell and Kowal (2005). Through a detailed analysis of media interviews of Hillary Clinton, they found that her use of filled pauses does not necessarily signal an upcoming delay. Rather the distribution is better explained by contextual factors or individual variation. One apparent drawback with these experimental work is their inability to generalize to more realistic speech settings, and the potential confounding with unobserved endogeneity which may in fact explain their observed patterns that have been mistakenly attributed to the variables coded in the experiments. Thus corpus based analysis is necessary for drawing a full picture of the cognitive process in not only the disfluency phenomena, but also the speech production more broadly.

It can be summarized from the brief review of literature on the linguistic and cognitive variables behind speech production and disfluency that disfluencies are correlated with issues with planning, lexical access, and the motor control process during production. More demanding contexts, such as an increase in sentence complexity, difficult production task, and lower transition probability between words, are associated with more disfluent speech. Furthermore, filled pauses and repetitions may relate to different underlying cognitive mechanism, and be subject to different level of speaker's active control. However, most conclusions reached in the literature so far are either agnostic to particular types of disfluencies, or basing off evidence that does not focus on variations within each difluency type. It is therefore worth further asking, within each type of disfluencies, why and how the disfluent forms may vary? And what is the role that linguistic context plays in the observed variation? Attempts have been made to explicitly address part of this question (Schnadt

& Corley, 2006), however, as mentioned earlier, the lack of understanding of the connection between lab environment and real speech production settings warrants additional efforts using corpus data. Answers to this question will not only push our knowledge about the cognitive variables in disfluencies beyond experiment settings, but also lead to innovations in other practical fields where knowledge of disfluency can be particularly helpful.

In this chapter, I attempt to address this question by complementing our existing knowledge in each of the linguistic levels involved in the production process. Unlike most previous studies that only focused on isolated levels of linguistic analysis, I explicitly model the linguistic contexts in which disfluencies are expected, through a combination of large scale corpus study and detailed small sample analysis. Specifically, I first look at the effect of lexical and syntactic factors on filled pause distribution, as well as on the variation of repetitions. In addition to the descriptive analyses of the linguistic covariates based on large collection of speech, A detailed and efficient annotation system is developed, and deployed in a fine analysis of a smaller dataset. Results from these aspects will be interpreted from the view of both language production models and their implications for the practical fields, such as on patients with known challenges in certain stages in the production process.

4.2 Linguistic contexts

4.2.1 The syntactic context

In this section, I propose to examine the distribution and form variation of filled pauses and repetitions in regard to the phrase they occur. Before moving to original analyses, I will first confirm the distributional patterns and relations of filled pauses and repetitions reported else where with the current data. Specifically, I will examine earlier observations that hesitations are more likely to occur before more complex syntactic structures (Holmes, 1988), longer sentences (Shriberg, 1994), unexpected or infrequent lexical items (Arnold et al., 2007) and the tendency that two forms of filled pauses occur at different locations within an utterance (Clark & Tree, 2002). These relations will be examined by type dependent analyses: for each type of disfluencies of interest, what is its distributional properties in these dimensions? A large scale and relatively crude exploration of questions regarding the location within a speech segment and in relation to a speech segment boundary will be measured as the location of disfluent instance in a turn or contiguous speech segment. Sampling methods will be used to offer a more detailed picture regarding the actual phrase boundary, part of speech of the words before and after the disfluent instance, and the part of speech of repeated lexical items. The exact kinds of analyses, by disfluency type, sample type and annotation methods, are listed in table 9.

The main contribution of this section will be the distributional property of repetition and repair disfluencies. Although Clark and Wasow (1998) have established the theoretical foundation that repetitions display an effort to preserve the continuity of produced speech, more empirical evidence is needed to answer questions such as to what extent is this claim applicable? How to explain the variation in the likelihood of words in a same word class being repeated, as well as the variation in the location where a word is repeated? Are the location and word class of repeated words related to the number of repetitions and form of repetitions (full word repetition, multi-word repetition, and partial word repetition), and how? Answers to these questions will be explored, again, by first looking at type dependent repetitions with regard to their location in a turn, in a pause group, and

Table 9: The proposed variables and samples used in the analyses of disfluencies (DF)

DF type	Sample	Annotation	Variables
Filled pauses	Full sample	Automatic	Position in a turn (word index)
			Position in a pause group
			Likelihood of function words after DF
			Distribution of function words after DF
			Pause duration after DF
			Pause group duration
	Sub-sample	Manual	POS of the preceding and following words
			Constituency of the preceding and following words
Repetition and repair	Full sample	Automatic	Position in a turn (word index)
			Size of the pause group
			Likelihood of function words after DF
			Likelihood of filled pause after DF
			POS of repeated words
	Sub-sample	Manual	POS of the following words
			Constituency of the following words

the relative frequency of the same word forms that are repeated or fluent. The variables concerning repetition types include the number of repeats, number of words in a repetition, whether there is any partial word in the repetition, whether there is any filled pause within a repetition, and whether the repetition involves replacement of words. Detailed analyses will then follow using a subset of speech generated through proper sampling method.

4.2.2 The lexical context

Like the analyses in the previous section, here I will carry out the analyses in two steps. In confirming the reported effect of lexical frequency, or predictability of the lexical context, I will examine the words preceding and following the disfluent segments. Type dependent analyses will still be used. Specific questions to address include what is the distribution of word classes adjacent to disfluencies? How does the transition probability of words preceding and following disfluent segments relate to the form variation of particular type of disfluency? In addition to the frequency or transition probability based parameters normally discussed in the literature, I also propose to directly model the full distribution of lexical items in vicinity of the disfluent segments. The proposed dimensions of analyses are summarized in table 10.

Modeling the lexical distribution in the context of disfluencies directly through vector representations can uncover the dissimilarity of lexical contexts across disfluency categories. An interpretation of the discrimination can be achieved through methods inspired in topic modeling, such as by looking at the saliency or relevance (Chuang, Manning, & Heer, 2012; Sievert & Shirley, 2014) of particular terms in the distribution matrix. As suggested in the previous section as well, the expected main contribution of this section is also on the distrbitional property of repetitions.

Table 10: The proposed variables and samples used in the analyses

DF type	Sample	Variables
Filled pauses	Full sample	Raw frequency of the first content word form after DF Transition probability between the word before and after DF Transition probability between the last content word before and the first content word after DF Relative frequency of the first content word form after DF in the conversation Relative frequency of the first content word form after DF in conversations of the same topic Relative frequency of the first content word form after DF in conversations of the same speaker Term-frequency vector of the first content word form before and after DF in the conversation Term-frequency vector of the first content word form before and after DF in conversations of the same topic Term-frequency vector of the first content word form before and after DF of the same speaker
Repetitions	Full sample	Ditto the variables above, but adding the frequency and distribution measurements for the repeated words and phrases

Specifically, I will look at the type distribution of words following repeated words, such as the frequency of different types of content words, the frequency of filled pauses, and the frequency of misarticulations. Comparisons between the repeated and fluent version of same words will be carried out along the dimensions defined above. Type dependent analyses of repetition phenomena will be performed following the same feature space outlined in 10. Again, if more detailed analyses are deemed necessary, sampling methods will be applied to generate a subset of speech that will be suitable for manual annotation.

4.3 A detailed analysis of fluent repetitions and repairs

The focus of this section will be a detailed look at variation of fluent repetitions and repairs. As mentioned above, repetition and repair disfluencies are more challenging types of disfluency phenomena because their high variability in realized forms and the lack of consistently annotated corpora to carry out large scale analyses. As an attempt to address these two challenges, in this section, the discussion will be restricted to a much smaller sample, with the goal of devising an efficient semi-automatic annotation system, as well as using this system to evaluate individual variations in repetition and repair patterns. The analyses in this section will thus focus on the speech from eight US Supreme Court justices using data from the SCOTUS corpus. Variations in repetition, under this context, are explored in two stages. The first stage is to identify the patterns of interests that contrast the repetitions from ample speech of different individuals, along the dimensions in lexical, syntactic and prosodic contexts identified in earlier crude large-scale analyses. Then hypotheses on why the observed variations exit will be formed, along with proposals for possible explanations.

It is hoped that the results from this section will serve as an initial building block for future larger scale investigations into the linguistic contexts and speaker idiosyncratic factors that are behind repetitions and repairs in fluent spontaneous speech.

4.3.1 The proposed annotation strategy

The annotation will be done on the time aligned SCOTUS 2001 corpus, where the speech from each justice in court sessions has been grouped and segmented into turns. The annotation will follow an adapted version of Shriberg (1994)'s pattern labeling system (PLS), focusing mostly on the reparandum and repair annotation. In the present case, reparandum explicitly refers to the repeated segments, and repair refers to the last repetition or repair for preceding repetitions that is integrated with the following fluent utterance. In-line annotation will be adopted in the proposed project, mainly for efficiency considerations. Symbols to be used are summarized in table 11.

Table 11: The proposed detailed annotation system for repetitions and repairs

Symbol	Explanation	Example
Primary symbols		
Unmarked	Fluent word	
"+"	Repeated word	that<+
"="	The substituting word	exclusionary<=
"-"	Word fragments	ex<-
"∼"	Substituted or deleted word	expression<∼
"e"	Explicit edit or other words or vocalization between RR and RP	%um<e
Secondary symbols		
"."	Interruption point	that<+.
"b"	The beginning of a repeating unit	that<+b
"o"	The middle of a repeating unit	in<+b your<+o
Other symbols		
"<"	Separator between word and annotation	that<+b
"%"	Filled pause	%um

In this system, annotations will be organized in two levels: The primary symbols are used to represent the type of the disfluent word, and the secondary symbols are designed to mark the region a disfluent word belongs to. Primary and secondary symbols are ordered linearly from left to right, separated by the symbol <. The primary symbol can be omitted in the case of a complete restart, where the interruption point is immediately following the previous fluent utterance. The secondary symbol can be optional when the disfluency does not involve repetition and not immediately after the interruption point. Both primary and secondary symbols can be stacked, but primary symbols are always annotated to the left of the secondary symbol. A snapshot of the annotated transcript can be found in figure 25. The first row of the transcript records the speaker information, and the first two columns contain the time stamps of the starting and end time of the word segment.

```
(KENN ANTHONY_M_KENNEDY Justice)
19222800000 19224170000 that<+b
19224170000 19226450000 that<+o
19226450000 19227650000 to
19227650000 19229750000 me
19229750000 19231250000 is
19231250000 19232450000 one
19232450000 19233050000 of
19233050000 19234450000 the
19234450000 19237050000 hard
19237050000 19243050000 parts
19243050000 19244400000 of
19244400000 19247150000 this
19247150000 19253050000 case.
19253050000 19257940000 %uh
19257940000 19261240000 it's
19261240000 19263940000 not
19263940000 19271150000 quite
19271150000 19280890000 expressio
19280890000 19286740000 unius,
19286740000 19288610000 ex<-+b
19288610000 19290060000 ex<-+o
19290060000 19299550000 exclusialteri<=
19299550000 19300740000 but
19300740000 19303540000 it's
19303540000 19310810000 close.
19310810000 19317700000 {sil}
```

Figure 25: A screen shot of the annotated transcript from SCOTUS

4.3.2 The proposed variables

Following the same consideration adopted in previous sections (Fox et al., 1996; Clark & Wasow, 1998; Fox et al., 2010), variables to be considered in this detailed analysis will include features about both the repeated words or phrase themselves and the contextual features that capture syntactic, semantic and prosodic variations. Likewise, the odds of repeated word or phrase in non-repeating contexts against in repeating context, their lexical category, lexical frequency will be considered as the features of the repetition or repair, whereas the lexical category of the following words, distribution of the following words, length of the following speech segment, pause duration after repetition or repair, speech duration of the following speech segment, turn length and duration, the relative position of the repetition or repair in a turn and continuous speech segment, etc. will be considered to measure the contextual features. Unlike the previous large-scale analyses, turn and speech segments can be more accurately identified and segmented, as well as the part of speech and syntactic category of relevant words and phrases. Part of speech and syntactic category can be first annotated through automatic parser that is robust to conversational speech before manual annotation and correction. Each of these features, or information that is necessary for deriving the feature values, will be annotated as separate fields following the word segment in the transcripts.

These feature dimensions can be relatively easily extracted from the annotated transcription files. Within speaker comparison will be performed by comparing the speech from single justices across multiple debate sessions, and cross-speaker variation will be explored through comparing between justices along these dimensions. Contextual and discourse factors, including the theme of

the case, speech from other defendants and justices in the court debate, and other factors concerning the speech of each justices, such as speaking rate, silence to speech ratio, frequency of filled pauses, will also be considered in explaining the variation observed in repetitions.

4.3.3 Proposed quantitative analysis

The quantitative analyses will follow the same methodology used in the current study. The first step in the analysis is to get the distributional statistics about variations in repetitions and repairs in the feature space described above. From these distributional properties, the question to be answered is what are the possible structures, and their distribution, of repetitions and repairs in the corpus? Then what's the characteristics of each distinct structure? And finally what's the distribution for the speech from each individual speaker?

The distributional information described above will be converted to normalized numeric values for later quantitative data analysis for the purpose of pattern recognition. Since the first goal of this more detailed study is pattern recognition, the primary statistical statistical tool will be some clustering methods. Since the speech data contains a hierarchical structure, where speech samples can be grouped both by debate session and by speaker, hierarchical clustering algorithms will be considered first in identifying the structure of repetition patterns. Other methods, such as clustering based on similarity measurements (such as Multidimensional Scaling (MDS)) or distance measurements (such as Spectral Clustering) will be explored to establish the relations among the limited set of individual speakers. Dimensionality reduction techniques will be used for visualization purpose.

4.4 Disfluency, speech planning, and production models

This section serves as a discussion or summary of the findings in the previous section on linguistic contexts. An interpretation from the view of speech planning and language production will be offered. Disfluencies as cues to answering questions about the syntactic planning units during speech production have been discussed in Holmes (1988). With a more comprehensive description of the distribution of speech disfluencies in relation with the syntactic contexts, hypotheses on the planning units, such as those proposed in Ford and Holmes (1978), can be objectively tested. The information on the distributional properties of the lexical contexts of disfluencies can also help address problems faced with models of lexical planning, especially with the less restrictive lexical contexts presented in this study compared with experiments carried out in a lab. It is also hoped that from this more descriptive work with speech from a realistic setting, new insights can be brought to the design of processing experiment, with better control for previously uncontrolled covariates and aims to answer previously ignored questions about speech production.

4.5 Disfluency in population with neural degeneration

This section will serve as a pointer to the potential applications of the knowledge in speech disfluency in practical domain. I will try to apply what have been found in previous sections on the correlates of linguistic variables, as well as interpretations from the perspective of planning and cognitive correlates of speech production, to patients with known cognitive impariment, such as in the population with neural degenerative diseases. Detailed knowledge about aspects of distributions of dislfuencies can be especially useful in this domain, due to the lack of suitable speech data

and the high cost of data acquisition. In the proposed work, I will examine the speech produced during clinical interviews that are part of cognitive assessment, using the feature space discussed in this chapter. Specifically, the speech produced in the picture description task during the interview will be examined. Distributions of the linguistic variables among the clinical sample will be compared to the normative sample examined above. The speech samples from the clinical population will be obtained in collaboration with Penn FTD center. Since it has been known that speech disfluencies can be a salient feature of patients with FTD (Mack et al., 2015), but the diagnosis can be challenged by ambiguities in the speech presented for human judgement, the findings from the previous sections can be helpful in offering a more objective and robust line of reference.

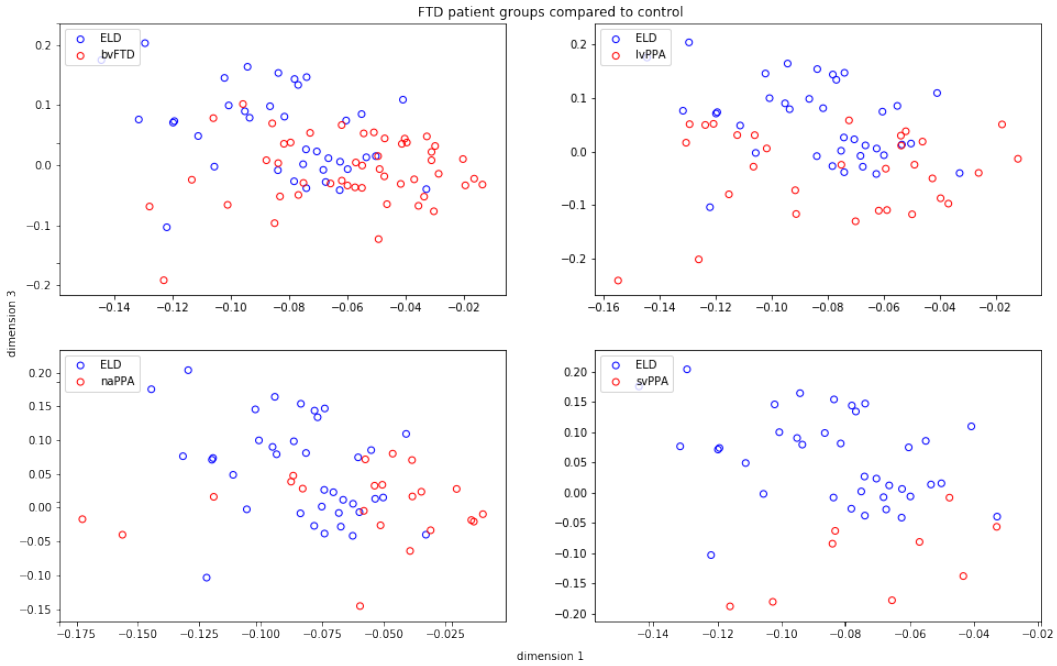


Figure 26: Four FTD patient groups compared to control in the derived space constructed using temporal and lexical features.

Figure 26 demonstrates how information on the temporal and lexical information can be helpful in discriminate among patients with variants of FTD. The original feature space was constructed using the temporal information, with the method described in Section 3.2.1, the term-frequency vector of the words in the immediate neighborhood of a silent pause longer than 250 ms, as well as the term-frequency vector of all the content words with frequency greater than 1 from the speech produced in a picture description task. Each dot represents an observed individual in the derived space. It can be observed that reasonable separation between the patient group and control group can be achieved in all FTD phenotypes in this 2D space. This naive demonstration nevertheless illustrates the potential application of a rich understanding of the linguistic contexts for disfluencies.

4.6 Conclusion

In this proposed chapter, I will explore the distributional properties of disfluencies with repsect to the linguistic contexts in which they occur. The examination will be conducted combining the

benefits of both methods based on large collection of speech corpus and detailed small sample analyses. An efficient annotation strategy for repair annotation will be developed and deployed in the proposed analyses. This chapter will also be the first of its kind in directly modeling the joint distribution of syntactic and semantic contexts of disfluencies. Implications of this joint distribution will be discussed from the perspective of speech production models, as well as its potential application in diagnosing variants of FTD diseases.

5 Cross-linguistic perspective: A case study of repetitions in Czech

Discussions on the cross-linguistic variation in repetition are often found in the literature on discourse analysis, where the terminology for repetition and repair is *recycling* and *replacement*. Although the forms of repetition have been shown to be dependent on the morphosyntactic structure of the language, the scope often respects the constituent boundary (Fox et al., 1996, 2010; Hayashi, 1994; Fincke, 1999). For example, in English, German and Hebrew, the scope of repetition generally involves from the function word immediately preceding the main content word of a constituent (Fox et al., 2010), while evidence of frequent partial verb repetition has been provided in Japanese, a language that is typically verb final and has tolerance of non-overt arguments in discourse (Fox et al., 1996). Japanese also contains morphological repetition and repair, where only the bound verbal suffix is repeated or replaced by another (Hayashi, 1994; Fox et al., 1996). The scope of repetition in Japanese is thus mostly within the constituents where repetition takes place, while repeating the full constituent is not impossible but rare. Similar observations have been made in Finnish (Kärkkäinen, Sorjonen, & Helasvuo, 2007). The examples below illustrate how repetitions and repairs may take place in Finnish and Japanese.

(2) tteyuuka koko denwa [kaket- kakete] kite sa
I:mean here telephone ca- call come FP
I mean, (they) ca- called us here,

In this example of Japanese, only the verb in the verb phrase is repeated in the repair, and proper inflection is added after the repetition. Therefore only the morphologically relevant segment in a phrase is repeated and repaired.

(3) mutta nyt [selvi-tä-än, -te-tä-än] nämä marka-t
but now manage-PASS-PERSCAUSE -PASS-PERS these mark-PL
But now let us manage, sort out these marks

In this example of Finnish, the speaker initially produced a passive intransitive verb, while wished to replace the verb with a transitive form. The strategy employed here is just to insert the transtive suffix *-te* and repeat the rest of suffixes that stay unchanged.

In the literature of discourse analysis, the scope of repetition and repair has been claimed to be related to the projectability of constituents, as proposed in Wouk (2005) citing examples from Indonesian. The syntax of repetition and self repair has been formalized in Uhmann (2001), citing evidence from German that the start of a repetition or repair has a preference of the functional head in the phrase structure. Under this view, the degree to which early parts of the constituents project

the syntactic structure and further indicate the completion of the clause determines the likelihood of the word or phrase being repeated or in the repair. For example, due to its right branching structure, the head of phrases in English project over the complements to the right, thus a wider scope of repetition is expected. On the other hand, in Japanese, this scope is rather limited due to its mostly left branching structure. Thus repairs need only be done with respect to the head. However, this claim does not justify the necessity of introducing the notion of scope of repetition into the picture: the null hypothesis should be that when it comes to repeating words in an utterance, a speaker would only repeat whatever is convenient: when there isn't small function words available at the left edge of a phrase, they would just repeat partial words, or possible filler words instead. Same logic also applies to repair, and is congruent with the assumption of preserving the continuity of delivery (Clark & Wasow, 1998). Since in the discourse studies literature, statistics of the distribution of repetitions and repairs is rarely found, and the sample size being worked with tends to be very small, the null hypothesis cannot be quantitatively rejected. Nevertheless, from examples cited in the qualitative literature, we have reason to expect that for a language with richer morphology, more flexible word order, and potentially different head directionality, the overall repetition frequency and type distribution would be very different from what have been known in English and related languages. Knowledge of the cross-linguistic pattern distributions of repetition and repair will not only help to enrich the current theory on the relation between repetition/repair and the syntax of a language, but also offer new insights to syntactic planning from a cross-linguistic perspective.

Czech is a good candidate for exploring the issues posed above. As a west Slavic language, Czech has relatively flexible word order and rich morphology. The case, gender, and number systems are almost exclusively expressed through a complex inflection system (Janda & Townsend, 2000). The trade-off of this complex inflection system is its limited inventory of function words such as pronouns and prepositions. Therefore what can be expected in terms of repetition in Czech is something like Finnish, a language with very complex morphology and relatively high freedom of word order. With the knowledge from Finnish and Japanese, it can be expected that repetitions of partial word and partial repairs of inflectional suffixes can be fairly common. One other possibility is that the possible variation of repetitions is more limited, while other types of disfluencies such as filled pauses would take place where repetition would otherwise occur. However, because unlike colloquial Finnish, where speakers tend to insert isolated pronouns, spoken Czech is even more restricted in the use of function word categories, a stronger pattern may also be expected in Czech.

To sum up, the questions to be addressed in this chapter can be formulated as the following. First, what is the patterns and their distributions of repetitions in Czech? And how to explain the observed variation with the morphosyntactic properties of the language, and interpret the observations in relation to other more widely studied languages? Unlike the research in discourse analysis, it is hoped that through mining the Czech Spontaneous Speech Corpus, the data would provide a robust response to the proposed questions.

In terms of the methodology, I propose to conduct an analysis of the repetitions largely following the procedure presented in chapter 3.4 and 4.3. The form variation and typology of repetitions in Czech will first be established. The syntactic variables, such as the location of repetitions in a phrase, the likelihood of observing a fluent or a repeated word of the same form across different phrase structures, and different utterance type or pause group type, will be explored. The semantic variables will include aspects of distributions of type of words following the repetitions, and how this variation is related to the type of repetitions. The goal of this chapter, in addition to offering

some first insight into the repetition phenomenon in a language that typologically more distinctive from our current knowledge, is mainly serves as bringing up a question that worth further studies. It is hoped that the proposed analyses will set off a predecessor for more fruitful studies on similar cross-linguistic problems.

6 Towards a more complete theory on disfluencies in spontaneous speech

In this chapter, I will discuss the implications from the results obtained from chapter 3 through 5. Unlike discussions in those chapters, where the sociolinguistic and linguistic variables are treated separately, I will consider jointly these variables in explaining how speech disfluencies happen in spontaneous speech. In the introduction session of her dissertation, Shriberg (1994) acknowledged that although an all-encompassing theory of disfluencies is the ultimate goal of disfluency research, the field was in an early stage of discovering the regularities in disfluency production. Although the past 25 years have seen tremendous development in the field, I will keep stressing the need of the continuous effort in this pattern recognition enterprise. What an ultimate theory on speech disfluencies is like may still be out of reach in the current date and time, but every piece of finer description of the phenomena will bring us a step closer to the goal. It will be finally argued that such a single theory that has significance in the theoretical work in sociolinguistics, psycholinguistics and practical applications should acknowledge, and encompass, the complexity of the phenomenon from all of the related perspectives.

7 Timeline

The overall plan is the following. I will finish the analyses of repetitions using large corpus data by May, and the annotation of SCOTUS over the summer. Data analysis of the annotated SCOTUS corpus and from clinical population will be finished by October. Data analysis on the Czech data will be done by December. Write-up of these results will be carried out along with the analyses. The theoretical implications in Chapter 4, and Chapter 6 will be finished by March. The final draft of the paper will be expected in April, and I will aim at a defense in May 2020. Table 12 outlines the more detailed plan.

Table 12: The proposed timeline for the dissertation

Time	Goal
May 2019	Finish the analyses for Chapter 3.4.3, 3.4.4, 3.4.5
June 2019	Finish the analyses for Chapter 4.2 Start annotating SCOTUS
July 2019	Annotating SCOTUS
August 2019	Finish annotating SCOTUS Start the analysis for Chapter 4.3
September 2019	Finish the analysis for Chapter 4.3 Start working on Chapter 4.5
October 2019	Continue working on Chapter 4.5 Start the analysis for Chapter 5
November 2019	Conclude the analysis for Chapter 4.5
December 2019	Conclude the analysis for Chapter 5
January 2020	Finish Chapter 3 and 4
February 2020	Finish Chapter 5
March 2020	Finish Chapter 6 and Chapter 1, 2
April 2020	Revise
May 2020	Defense

References

Acton, E. K. (2011). On gender differences in the distribution of um and uh. *University of Pennsylvania Working Papers in Linguistics*, *17*(2), 2.

Adell, J., Escudero, D., & Bonafonte, A. (2012). Production of filled pauses in concatenative speech synthesis based on the underlying fluent sentence. *Speech Communication*, *54*(3), 459–476.

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

Arnold, J. E., Kam, C. L. H., & Tanenhaus, M. K. (2007). If you say thee uh you are describing something hard: The on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(5), 914.

Arnold, J. E., Losongco, A., Wasow, T., & Ginstrom, R. (2000). Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, *76*(1), 28–55.

Ash, S., McMillan, C., Gunawardena, D., Avants, B., Morgan, B., Khan, A., ... Grossman, M. (2010). Speech errors in progressive non-fluent aphasia. *Brain and language*, *113*(1), 13–20.

Ash, S., Moore, P., Vesely, L., Gunawardena, D., McMillan, C., Anderson, C., ... Grossman, M. (2009). Non-fluent speech in frontotemporal lobar degeneration. *Journal of Neurolinguistics*, *22*(4), 370–383.

Bear, J., Dowding, J., Shriberg, E., & Price, P. (1993). *A system for labeling self-repairs in speech.* Feb.

Beattie, G. W., & Barnard, P. (1979). The temporal structure of natural telephone conversations (directory enquiry calls). *Linguistics*, *17*(3-4), 213–230.

Beattie, G. W., & Butterworth, B. L. (1979). Contextual probability and word frequency as determinants of pauses and errors in spontaneous speech. *Language and Speech*, *22*(3), 201–211.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustical Society of America*, *113*(2), 1001–1024.

Beňuš, Š., Gravano, A., & Hirschberg, J. (2011). Pragmatic aspects of temporal accommodation in turn-taking. *Journal of Pragmatics*, *43*(12), 3001–3027.

Beňuš, Š., Levitan, R., & Hirschberg, J. (2012). Entrainment in spontaneous speech: The case of filled pauses in Supreme Court hearings. In *2012 ieee 3rd international conference on cognitive infocommunications (coginfocom)* (pp. 793–797).

Blacfkmer, E. R., & Mitton, J. L. (1991). Theories of monitoring and the timing of repairs in spontaneous speech. *Cognition*, *39*(3), 173–194.

Blankenship, J., & Kay, C. (1964). Hesitation phenomena in English speech: A study in distribution. *Word*, *20*(3), 360–372.

Bohus, D., & Horvitz, E. (2014). Managing human-robot engagement with forecasts and... um... hesitations. In *Proceedings of the 16th international conference on multimodal interaction* (pp. 2–9).

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and speech*, *44*(2), 123–147.

Boschi, V., Catricala, E., Consonni, M., Chesi, C., Moro, A., & Cappa, S. F. (2017). Connected speech in neurodegenerative language disorders: A review. *Frontiers in psychology*, *8*, 269.

Brennan, S. E., & Williams, M. (1995). The feeling of another s knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of memory and language*, *34*(3), 383–398.

Broen, P. A., & Siegel, G. M. (1972). Variations in normal speech disfluencies. *Language and Speech*, *15*(3), 219–231.

Butcher, A. (1981). Aspects of the speech pause: Phonetic correlates and communication functions. *Arbeitsberichte Kiel*(15), 1–233.

Butzberger, J., Murveit, H., Shriberg, E., & Price, P. (1992). Spontaneous speech effects in large vocabulary speech recognition applications. In *Proceedings of the workshop on speech and natural language* (pp. 339–343).

Campione, E., & Véronis, J. (2002). A large-scale multilingual study of silent pause duration. In *Speech Prosody 2002, International Conference.*

Chuang, J., Manning, C. D., & Heer, J. (2012). Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the international working conference on advanced visual interfaces* (pp. 74–77).

Cieri, C., Miller, D., & Walker, K. (2004). The Fisher Corpus: A resource for the next generations of speech-to-text. In *Lrec* (Vol. 4, pp. 69–71).

Clark, H. H., & Tree, J. E. F. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*(1), 73–111.

Clark, H. H., & Wasow, T. (1998). Repeating words in spontaneous speech. *Cognitive psychology*, *37*(3), 201–242.

Consortium, L. D. (2009, 7). *Czech Broadcast Conversation MDE Transcripts.* `https://catalog.ldc.upenn.edu/LDC2009T20l`. (accessed April 18, 2019)

Cooper, W. E., & Paccia-Cooper, J. (1980). *Syntax and speech* (No. 3). Harvard University Press.

Corley, M., & Stewart, O. W. (2008). Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass*, *2*(4), 589–602.

Dahl, D. A., Bates, M., Brown, M., Fisher, W., Hunicke-Smith, K., Pallett, D., ... Shriberg, E. (1994). Expanding the scope of the ATIS task: The ATIS-3 corpus. In *Proceedings of the workshop on human language technology* (pp. 43–48).

Dickerson, W. B. (1972). *Hesitation phenomena in the spontaneous speech of non-native speakers of English.* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.

Duez, D. (1982). Silent and non-silent pauses in three speech styles. *Language and Speech*, *25*(1), 11–28.

Eklund, R. (2004). *Disfluency in Swedish human–human and human–machine travel booking dialogues* (Unpublished doctoral dissertation). Linköping University Electronic Press.

Ferreira, F. (1991). Effects of length and syntactic complexity on initiation times for prepared utterances. *Journal of Memory and Language*, *30*(2), 210–233.

Ferreira, F. (1993). Creation of prosody during sentence production. *Psychological review*, *100*(2), 233.

Ferreira, F. (2007). Prosody and performance in language production. *Language and Cognitive Processes*, *22*(8), 1151–1177.

Ferreira, F., & Bailey, K. G. (2004). Disfluencies and human language comprehension. *Trends in cognitive sciences*, *8*(5), 231–237.

Ferreira, V. S., & Pashler, H. (2002). Central bottleneck influences on the processing stages of word production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(6), 1187.

Fincke, S. (1999). The syntactic organization of repair in Bikol. *Cognition and function in language*, 252–267.

Ford, M., & Holmes, V. M. (1978). Planning units and syntax in sentence production. *Cognition*, *6*(1), 35–53.

Foster, J. (2010). cba to check the spelling investigating parser performance on discussion forum posts. In *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics* (pp. 381–384).

Fox, B. A., Hayashi, M., & Jasperson, R. (1996). Resources and repair: A cross-linguistic study of syntax and repair. *Studies in Interactional Sociolinguistics*, *13*, 185–237.

Fox, B. A., Maschler, Y., & Uhmann, S. (2010). A cross-linguistic study of self-repair: Evidence from English, German, and Hebrew. *Journal of Pragmatics*, *42*(9), 2487–2505.

Fruehwald, J. (2016). Filled pause choice as a sociolinguistic variable. *University of Pennsylvania Working Papers in Linguistics*, *22*(2), 6.

Gee, J. P., & Grosjean, F. (1983). Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive psychology*, *15*(4), 411–458.

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992* (Vol. 1, pp. 517–520).

Goldman-Eisler, F. (1958). Speech production and the predictability of words in context. *Quarterly Journal of Experimental Psychology*, *10*(2), 96–106.

Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech*. Academic Press.

Goldwater, S., Jurafsky, D., & 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.

Goodglass, H., Kaplan, E., & Barresi, B. (2000). *Boston Diagnostic Aphasia Examination Record Booklet*. Lippincott Williams & Wilkins.

Greene, J. O., & Cappella, J. N. (1986). Cognition and talk: The relationship of semantic units to temporal patterns of fluency in spontaneous speech. *Language and Speech*, *29*(2), 141–157.

Grossman, M., & Ash, S. (2004). Primary progressive aphasia: a review. *Neurocase*, *10*(1), 3–18.

Guitar, B. (2013). *Stuttering: An integrated approach to its nature and treatment*. Lippincott Williams & Wilkins.

Harmon, Z., & Kapatsinski, V. (2015). Studying the dynamics of lexical access using disfluencies. In *Papers presented at* (p. 41).

Hartsuiker, R. J., & Notebaert, L. (2009). Lexical access problems lead to disfluencies in speech. *Experimental psychology*.

Hayashi, M. (1994). A comparative study of self-repair in English and Japanese conversation. *Japanese/Korean Linguistics*, *4*, 77–93.

Heeman, P. A. (1997). Speech repairs, intonational boundaries and discourse markers: Modeling speakers' utterances in spoken dialog. *arXiv preprint cmp-lg/9712009*.

Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, *38*(4), 555–568.

Hieke, A. E. (1981). A content-processing view of hesitation phenomena. *Language and Speech*, *24*(2), 147–160.

Holmes, V. M. (1988). Hesitations and sentence planning. *Language and Cognitive Processes*, *3*(4), 323–361.

Hough, J. (2014). *Modelling incremental self-repair processing in dialogue.* (Unpublished doctoral dissertation). Queen Mary University of London.

Jacewicz, E., Fox, R. A., & Wei, L. (2010). Between-speaker and within-speaker variation in speech tempo of American English. *The Journal of the Acoustical Society of America*, *128*(2), 839–850.

Janda, L. A., & Townsend, C. E. (2000). *Czech*. Lincom Europa Munich.

Jescheniak, J. D., & Levelt, W. J. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 824.

Johnson, M., & Charniak, E. (2004). A TAG-based noisy channel model of speech repairs. In *Proceedings of the 42nd annual meeting on association for computational linguistics* (p. 33).

Johnson, T. R., & Goldman, J. (2009). *A good quarrel: America's top legal reporters share stories from inside the Supreme Court*. University of Michigan Press.

Johnson, W. (1961). Measurements of oral reading and speaking rate and disfluency of adult male and female stutterers and nonstutterers. *Journal of Speech & Hearing Disorders. Monograph Supplement*.

Kahn, J. G., Lease, M., Charniak, E., Johnson, M., & Ostendorf, M. (2005). Effective use of prosody in parsing conversational speech. In *Proceedings of the conference on human language technology and empirical methods in natural language processing* (pp. 233–240).

Kärkkäinen, E., Sorjonen, M.-L., & Helasvuo, M.-L. (2007). Discourse structure. *Language typology and syntactic description*, *2*, 301–371.

Kendall, T. S. (2009). *Speech rate, pause, and linguistic variation: An examination through the sociolinguistic archive and analysis project*. Duke University.

Kolár, J., Svec, J., Strassel, S., Walker, C., Kozlíková, D., & Psutka, J. (2005). Czech Spontaneous Speech Corpus with structural metadata. In *Ninth european conference on speech communication and technology.*

Kowtko, J. C., & Price, P. J. (1989). Data collection and analysis in the air travel planning domain. In *Proceedings of the workshop on speech and natural language* (pp. 119–125).

Krivokapić, J. (2007). Prosodic planning: Effects of phrasal length and complexity on pause duration. *Journal of Phonetics*, *35*(2), 162–179.

Laserna, C. M., Seih, Y.-T., & Pennebaker, J. W. (2014). Um... who like says you know: Filler word use as a function of age, gender, and personality. *Journal of Language and Social Psychology*, *33*(3), 328–338.

Levelt, W. J. (1983). Monitoring and self-repair in speech. *Cognition*, *14*(1), 41–104.

Levelt, W. J. (1989). *Speaking: From intention to articulation* (Vol. 1). MIT press.

Levelt, W. J., & Cutler, A. (1983). Prosodic marking in speech repair. *Journal of semantics*, *2*(2), 205–218.

Liberman, M. (2005, 3). *Young men talk like old women.* `http://itre.cis.upenn.edu/~myl/languagelog/archives/002629.html`. (accessed March 31, 2019)

Liberman, M. (2016, 3). *Political sound and silence.* `http://languagelog.ldc.upenn.edu/nll/?p=23990`. (accessed March 31, 2019)

Lickley, R. J. (1994). *Detecting disfluency in spontaneous speech* (Unpublished doctoral dissertation). University of Edinburgh.

Lickley, R. J. (1998). HCRC disfluency coding manual. *Human Communication Research Centre, University of Edinburgh*.

Lickley, R. J. (2015). Fluency and disfluency. *The handbook of speech production*, 445.

Lickley, R. J., & Bard, E. G. (1998). When can listeners detect disfluency in spontaneous speech? *Language and speech*, *41*(2), 203–226.

Liu, Y., Shriberg, E., Stolcke, A., Hillard, D., Ostendorf, M., & Harper, M. (2006). Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on audio, speech, and language processing*, *14*(5), 1526–1540.

Luce, P. A., & Charles-Luce, J. (1985). Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production. *The Journal of the Acoustical Society of America*, *78*(6), 1949–1957.

Mack, J. E., Chandler, S. D., Meltzer-Asscher, A., Rogalski, E., Weintraub, S., Mesulam, M.-M., & Thompson, C. K. (2015). What do pauses in narrative production reveal about the nature of word retrieval deficits in PPA? *Neuropsychologia*, *77*, 211–222.

Maclay, H., & Osgood, C. E. (1959). Hesitation phenomena in spontaneous English speech. *Word*, *15*(1), 19–44.

Mahl, G. F. (1956). Disturbances and silences in the patient's speech in psychotherapy. *The Journal of Abnormal and Social Psychology*, *53*(1), 1.

Martin, J. G. (1970). On judging pauses in spontaneous speech. *Journal of Verbal Learning & Verbal Behavior*.

McDougall, K., & Duckworth, M. (2017). Profiling fluency: An analysis of individual variation in disfluencies in adult males. *Speech Communication*, *95*, 16–27.

Moniz, H., Batista, F., Mata, A. I., & Trancoso, I. (2014). Speaking style effects in the production of disfluencies. *Speech Communication*, *65*, 20–35.

Nakatani, C. H., & Hirschberg, J. (1994). A corpus-based study of repair cues in spontaneous speech. *The Journal of the Acoustical Society of America*, *95*(3), 1603–1616.

Nevler, N., Ash, S., Jester, C., Irwin, D. J., Liberman, M., & Grossman, M. (2017). Automatic measurement of prosody in behavioral variant FTD. *Neurology*, *89*(7), 650–656.

Oomen, C. C., & Postma, A. (2001). Effects of time pressure on mechanisms of speech production and self-monitoring. *Journal of Psycholinguistic Research*, *30*(2), 163–184.

Ostendorf, M., & Hahn, S. (2013). A sequential repetition model for improved disfluency detection. In *Interspeech* (pp. 2624–2628).

Oviatt, S. (1995). Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language*, *9*(1), 19–36.

Oviatt, S., Darves, C., & Coulston, R. (2004). Toward adaptive conversational interfaces: Modeling speech convergence with animated personas. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *11*(3), 300–328.

OConnell, D., & Kowal, S. (2005). Uh and um revisited: Are they interjections for signaling delay? *Journal of Psycholinguistic Research*, *34*(6), 555–576.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: an ASR corpus based on public domain audio books. In *2015 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 5206–5210).

Plauché, M., & Shriberg, E. (1999). Data-driven subclassification of disfluent repetitions based on prosodic features. In *Proc. international congress of phonetic sciences* (Vol. 2, pp. 1513–1516).

Qian, X., & Liu, Y. (2013). Disfluency detection using multi-step stacked learning. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 820–825).

Rayson, P., Leech, G. N., & Hodges, M. (1997). Social differentiation in the use of English vocabulary: Some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics*, *2*(1), 133–152.

Roberts, P. M., Meltzer, A., & Wilding, J. (2009). Disfluencies in non-stuttering adults across sample lengths and topics. *Journal of Communication Disorders*, *42*(6), 414–427.

Rochester, S. R. (1973). The significance of pauses in spontaneous speech. *Journal of Psycholinguistic Research*, *2*(1), 51–81.

Schachter, S., Christenfeld, N., Ravina, B., & Bilous, F. (1991). Speech disfluency and the structure of knowledge. *Journal of Personality and Social Psychology*, *60*(3), 362.

Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, *53*(2), 361–382.

Schnadt, M. J., & Corley, M. (2006). The influence of lexical, conceptual and planning based factors on disfluency production. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 28).

Shriberg, E. (1994). *Preliminaries to a theory of speech disfluencies* (Unpublished doctoral dissertation). Citeseer.

Shriberg, E. (1995). Acoustic properties of disfluent repetitions. In *Proceedings of the international congress of phonetic sciences* (Vol. 4, pp. 384–387).

Shriberg, E. (2001). To errrris human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, *31*(1), 153–169.

Shriberg, E., & Lickley, R. (1993). Intonation of clause-internal filled pauses. *Phonetica*, *50*(3), 172–179.

Shriberg, E., & Stolcke, A. (1996). Word predictability after hesitations: a corpus-based study. In *ICSLP 96., Fourth International Conference on Spoken Language Processing.* (Vol. 3, pp. 1868–1871).

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63–70).

Siu, M.-h., & Ostendorf, M. (1996). Modeling disfluencies in conversational speech. In *Proceeding of fourth international conference on spoken language processing. icslp'96* (Vol. 1, pp. 386–389).

Skantze, G., & Hjalmarsson, A. (2010). Towards incremental speech generation in dialogue systems. In *Proceedings of the 11th annual meeting of the special interest group on discourse and dialogue* (pp. 1–8).

Skantze, G., Hjalmarsson, A., & Oertel, C. (2013). Exploring the effects of gaze and pauses in situated human-robot interaction. In *Proceedings of the sigdial 2013 conference* (pp. 163–172).

Smith, V. L., & Clark, H. H. (1993). On the course of answering questions. *Journal of memory and language*, *32*(1), 25–38.

Stolcke, A., & Shriberg, E. (1996). Statistical language modeling for speech disfluencies. In *1996 ieee international conference on acoustics, speech, and signal processing conference proceedings* (Vol. 1, pp. 405–408).

Swerts, M. (1998). Filled pauses as markers of discourse structure. *Journal of pragmatics*, *30*(4), 485–496.

Swerts, M., & Krahmer, E. (2005). Audiovisual prosody and feeling of knowing. *Journal of Memory and Language*, *53*(1), 81–94.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

Tannenbaum, P. H., Williams, F., & Hillier, C. S. (1965). Word predictability in the environments of hesitations. *Journal of verbal learning and verbal behavior*, *4*(2), 134–140.

Ten Bosch, L., Oostdijk, N., & Boves, L. (2005). On temporal aspects of turn taking in conversational dialogues. *Speech Communication*, *47*(1-2), 80–86.

Tily, H., Gahl, S., Arnon, I., Snider, N., Kothari, A., & Bresnan, J. (2009). Syntactic probabilities affect pronunciation variation in spontaneous speech. *Language and Cognition*, *1*(2), 147–165.

Tottie, G. (2011). Uh and um as sociolinguistic markers in British English. *International Journal of Corpus Linguistics*, *16*(2), 173–197.
Tottie, G. (2014). On the use of uh and um in American English. *Functions of Language*, *21*(1), 6–29.
Tsiamtsiouris, J., & Cairns, H. S. (2013). Effects of sentence-structure complexity on speech initiation time and disfluency. *Journal of Fluency Disorders*, *38*(1), 30–44.
Uhmann, S. (2001). Some arguments for the relevance of syntax to same-sentence self-repair in everyday German conversation. *Studies in interactional linguistics*, 373–404.
Walker, K., Ma, X., Graff, D., Strassel, S., Sessa, S., & Jones, K. (2015, 2). *RATS Speech Activity Detection LDC2015S02.* Hard Drive. Philadelphia: Linguistic Data Consortium.
Watanabe, M., Kashiwagi, Y., & Maekawa, K. (2015). The relationship between preceding clause type, subsequent clause length and duration of silent and filled pauses at clause boundaries in Japanese monologues. In *The 7th Workshop on Disfluency in Spontaneous Speech (DiSS 2015).*
Watson, D., & Gibson, E. (2004). The relationship between intonational phrasing and syntactic structure in language production. *Language and cognitive processes*, *19*(6), 713–755.
Wieling, M., Grieve, J., Bouma, G., Fruehwald, J., Coleman, J., & Liberman, M. (2016). Variation and change in the use of hesitation markers in Germanic languages. *Language Dynamics and Change*, *6*(2), 199–234.
Wouk, F. (2005). The syntax of repair in Indonesian. *Discourse Studies*, *7*(2), 237–258.
Yuan, J., & Liberman, M. (2008). Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America*, *123*(5), 3878.
Yuan, J., Xu, X., Lai, W., & Liberman, M. (2016). Pauses and pause fillers in Mandarin monologue speech: The effects of sex and proficiency. *Proceedings of Speech Prosody 2016*, 1167–1170.
Zellner, B. (1994). Pauses and the temporal structure of speech. In *Zellner, B.(1994). Pauses and the temporal structure of speech, in e. keller (ed.) Fundamentals of speech synthesis and speech recognition.(pp. 41-62). chichester: John Wiley.* (pp. 41–62). John Wiley.
Zvonik, E., & Cummins, F. (2003). The effect of surrounding phrase lengths on pause duration. In *Eighth European Conference on Speech Communication and Technology.*

Appendix

A. Topic list in Fisher Corpus

- **ENG01** Professional Sports on TV: Do either of you have a favorite TV sport? How many hours per week do you spend watching it and other sporting events on TV?
- **ENG02** Pets: Do either of you have a pet? If so, how much time each day do you spend with your pet? How important is your pet to you?
- **ENG03** Life Partners: What do each of you think is the most important thing to look for in a life partner?

- **ENG04** Minimum Wage: Do each of you feel the minimum wage increase - to $5.15 an hour - is sufficient?
- **ENG05** Comedy: How do you each draw the line between acceptable humor and humor that is in bad taste?
- **ENG06** Hypothetical Situations. Perjury: Do either of you think that you would commit perjury for a close friend or family member?
- **ENG07** Hypothetical Situations. One Million Dollars to leave the US.: Would either of you accept one million dollars to leave the US and never return? If you were willing to leave, where would you go, what would you do? What would you miss the most about the US? What would you not miss?
- **ENG08** Hypothetical Situations. Opening your own business: If each of you could open your own business, and money were not an issue, what type of business would you open? How would you go about doing this? Do you feel you would be a successful business owner?
- **ENG09** Hypothetical Situations. Time Travel.: If each of you had the opportunity to go back in time and change something that you had done, what would it be and why?
- **ENG10** Hypothetical Situations. An Anonymous Benefactor: If an unknown benefactor offered each of you a million dollars - with the only stipulation being that you could never speak to your best friend again - would you take the million dollars?
- **ENG11** US Public Schools.: In your opinions, is there currently something seriously wrong with the public school system in the US, and if so, what can be done to correct it?
- **ENG12** Affirmative Action.: Do either of you think affirmative action in hiring and promotion within the business community is a good policy?
- **ENG13** Movies.: Do each of you enjoy going to the movies in a theater, or would you rather rent a movie and stay home? What was the last movie that you saw? Was it good or bad and why?
- **ENG14** Computer games.: Do either of you play computer games? Do you play these games on the internet or on CD- ROM? What is your favorite game?
- **ENG15** Current Events.: How do both of you keep up with current events? Do you get most of your news from TV, radio, newspapers, or people you know?
- **ENG16** Hobbies.: What are your favorite hobbies? How much time do each of you spend pursuing your hobbies? Do you feel that every person needs at least one hobby?
- **ENG17** Smoking.: How do you both feel about the movement to ban smoking in all public places? Do either of you think Smoking Prevention Programs, Counter-smoking ads, Help Quit hotlines and so on, are a good idea?
- **ENG18** Terrorism.: Do you think most people would remain calm, or panic during a terrorist attack? How do you think each of you would react?

- **ENG19** Televised Criminal Trials.: Do either of you feel that criminal trials, especially those involving high-profile individuals, should be televised? Have you ever watched any high-profile trials on TV?
- **ENG20** Drug testing.: How do each of you feel about the practice of companies testing employees for drugs? Do you feel unannounced spot-checking for drugs to be an invasion of a person's privacy?
- **ENG21** Family Values.: Do either of you feel that the increase in the divorce rate in the US has altered your behavior? Has it changed your views on the institution of marriage?
- **ENG22** Censorship.: Do either of you think public or private schools have the right to forbid students to read certain books?
- **ENG23** Health and Fitness.: Do each of you exercise regularly to maintain your health or fitness level? If so, what do you do? If not, would you like to start?
- **ENG24** September 11.: What changes, if any, have either of you made in your life since the terrorist attacks of Sept 11, 2001?
- **ENG25** Strikes by Professional Athletes.: How do each of you feel about the recent strikes by professional athletes? Do you think that professional athletes deserve the high salaries they currently receive?
- **ENG26** Airport Security.: Do either of you think that heightened airport security lessens the chance of terrorist incidents in the air?
- **ENG27** Issues in the Middle East.: What does each of you think about the current unrest in the Middle East? Do you feel that peace will ever be attained in the area? Should the US remain involved in the peace process?
- **ENG28** Foreign Relations.: Do either of you consider any other countries to be a threat to US safety? If so, which countries and why?
- **ENG29** Education.: What do each of you think about computers in education? Do they improve or harm education?
- **ENG30** Family.: What does the word family mean to each of you?
- **ENG31** Corporate Conduct in the US.: What do each of you think the government can do to curb illegal business activity? Has the cascade of corporate scandals caused the mild recession and decline in the US stock market and economy? How have the scandals affected you?
- **ENG32** Outdoor Activities.: Do you like cold weather or warm weather activities the best? Do you like outside or inside activities better? Each of you should talk about your favorite activities.

- **ENG33** Friends.: Are either of you the type of person who has lots of friends and acquaintances or do you just have a few close friends? Each of you should talk about your best friend or friends.
- **ENG34** Food.: Which do each of you like better–eating at a restaurant or at home? Describe your perfect meal.
- **ENG35** Illness.: When the seasons change, many people get ill. Do either of you? What do you do to keep yourself well? There is a saying, "A cold lasts seven days if you don't go to the doctor and a week if you do." Do you both agree?
- **ENG36** Personal Habits.": According to each of you, which is worse: gossiping, smoking, drinking alcohol or caffeine excessively, overeating, or not exercising?
- **ENG37** Reality TV.: Do either of you watch reality shows on TV. If so, which one or ones? Why do you think that reality based television programming, shows like "Survivor" or "Who Wants to Marry a Millionaire" are so popular?
- **ENG38** Arms Inspections in Iraq.: What, if anything, do you both think the US should do about Iraq? Do you think that disarming Iraq should be a major priority for the US?
- **ENG39** Holidays.: Do either of you have a favorit holiday? Why? If either of you you could create a holiday, what would it be and how would you have people celebrate it?
- **ENG40** Bioterrorism.: What do you both think the US can do to prevent a bioterrorist attack?

B. Words considered in function word categories in section 3.4

prepositions in, on, of, at, for, with, about, from, to

pronouns i, you, he, she, they, my, your, his, her, their, him,them, there

articles the, a

demonstratives this, that, these, those

auxiliary verbs can, do, could, should, would, will, did, does, shall, is, are, be, been, being, was, were, done, may', might, must, ought', cant́, wasnt́', isnt́, dont́, didnt́, wont́, werent́, may not, might not, couldnt́, shouldnt́, wouldnt́, had, have, has, havent́, hasnt́', hadnt́, have not, has not, had not, will not, would not, could not, should not, arent́

conjunctions and, but, or, if, because, although, before, after, since, until, while, when, which, where

relative pronouns who, whom, whose, what, why

others here, every, any, all, many, most, much, more, not