Deception in Spoken Dialogue: Classification and Individual Differences

Sarah Ita Levitan

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2019

ABSTRACT

Deception in Spoken Dialogue: Classification and Individual Differences

Sarah Ita Levitan

Automatic deception detection is an important problem with far-reaching implications in many areas, including law enforcement, military and intelligence agencies, social services, and politics. Despite extensive efforts to develop automated deception detection technologies, there have been few objective successes. This is likely due to the many challenges involved, including the lack of large, cleanly recorded corpora; the difficulty of acquiring ground truth labels; and major differences in incentives for lying in the laboratory vs. lying in real life. Another well-recognized issue is that there are individual and cultural differences in deception production and detection, although little has been done to identify them. Human performance at deception detection is at the level of chance, making it an uncommon problem where machines can potentially outperform humans.

This thesis addresses these challenges associated with research of deceptive speech. We created the Columbia X-Cultural Deception (CXD) Corpus, a large-scale collection of deceptive and non-deceptive dialogues between native speakers of Standard American English and Mandarin Chinese. This corpus enabled a comprehensive study of deceptive speech on a large scale. In the first part of the thesis, we introduce the CXD corpus and present an empirical analysis of acoustic-prosodic and linguistic cues to deception. We also describe machine learning classification experiments to automatically identify deceptive speech using those features. Our best classifier achieves classification accuracy of almost 70%, well above human performance.

The second part of this thesis addresses individual differences in deceptive speech. We present a comprehensive analysis of individual differences in verbal cues to deception, and several methods for leveraging these speaker differences to improve automatic deception

classification. We identify many differences in cues to deception across gender, native language, and personality. Our comparison of approaches for leveraging these differences shows that speaker-dependent features that capture a speaker's deviation from their natural speaking style can improve deception classification performance. We also develop neural network models that accurately model speaker-specific patterns of deceptive speech.

The contributions of this work add substantially to our scientific understanding of deceptive speech, and have practical implications for human practitioners and automatic deception detection.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, thank you to my extraordinary advisor, Julia Hirschberg. It would be impossible to list all the ways Julia has helped me. I first met Julia as an undergraduate, when she was my mentor for a summer research project through the CRA-W DREU program. She encouraged me to apply to graduate school and accepted me as a PhD student. Julia is a model mentor, who goes above and beyond to make sure that her students thrive. She has been a constant source of encouragement, sage advice, and unwavering support. I have learned so much from her, both professionally and personally. It has been a true privilege to be her student.

Thank you to my dissertation committee members, Robert Leonard, Kathleen McKeown, Smaranda Muresan, and Andrew Rosenberg, for generously giving their time to review this thesis. Their valuable feedback and suggestions have improved this thesis.

It has been a pleasure to work with many collaborators and project students over the past few years. Thank you to: Guozhen An, Nishi Cestero, Ivy Chen, Bingyan Hu, Kai-Zhan Lee, Michelle Levine, Rivka Levitan, Yocheved Levitan, Min Ma, Angel Maredia, Gideon Mendels, Andrew Rosenberg, Kara Schechtman, James Shin, Mandi Wang, William Wang, and Jessica Xiang. Thank you also to the many Columbia, Barnard, and CUNY students that helped collect and annotate the CXD corpus.

Thank you to my fellow members of the Speech Lab, past and present, for providing a wonderful and supportive environment: Daniel Bauer, Nishi Cestero, Xi Chen, Erica Cooper, Bob Coyne, Rivka Levitan, Anna Prokofieva, Rose Sloan, Victor Soto, Svetlana Stoyanchev, Morgan Ulinski, Laura Willson, Shirley, Xia, and Zixiaofan (Brenda) Yang.

I have been fortunate to complete two summer internships during my PhD, at Interactions LLC and at Google Research. Both experiences have helped make me a better researcher, and I am grateful to my mentors: Taniya Mishra and Srinivas Bangalore at

Interactions, and Dan Ellis and Shawn Hershey at Google.

Thank you to Jessica Rosa and the student services team, to the CS administrative staff, and to CRF for all of their help over the years.

This work was funded through a grant from the Air Force Office of Scientific Research as well as the NSF Graduate Research Fellowship.

I am forever grateful to my family for their love and support throughout this journey. To my incredible parents and siblings who have supported me from day one: I could never have done this without you. In particular, I owe a tremendous debt of gratitude to my superstar mother, for helping me every step of the way, for selflessly giving her time and energy, and for teaching me to never give up. I can't thank you enough.

Thank you to my wonderful husband Yoni, for holding down the fort while I worked on this, and to Shoshana and Yaakov – you are my greatest joy. A special thank you to my dear grandmothers, Nana and Grandma Betty, for inspiring and encouraging me and for all the babysitting. Grandma Betty, thank you for carefully proofreading this thesis. I love you all more than words can express.

Chapter 1

Introduction

Detecting deception from different dimensions of human behavior is a major goal of law enforcement, military, and intelligence agencies, as well as commercial organizations. Studies show that humans are poor at detecting deception, performing at about chance level [Bond Jr and DePaulo, 2006]; therefore the development of automated methods for deception detection is of great importance. Researchers of psychology, criminology, and computational linguistics have explored the use of several modalities for deception detection, including biometric indicators (measured by the polygraph), facial expressions, gestures and postures, brain imaging, and linguistic information. Despite extensive effort to develop automated deception detection technologies, there have been few objective successes. The lack of large, cleanly recorded corpora; the difficulty of acquiring ground truth labels; and major differences in incentives for lying in the laboratory vs. lying in real life situations are all obstacles to this work. Another well-recognized issue is that there are individual and cultural differences in deception production and detection, although little has been done to identify them.

This thesis addresses these challenges for deception research. An important contribution of this thesis is the creation of the Columbia X-Cultural Deception (CXD) Corpus, a large corpus of within-subject deceptive and non-deceptive speech from native speakers of Standard American English (SAE) and Mandarin Chinese (MC). The corpus was created using an original experimental paradigm to collect cleanly-recorded sessions, where participants provided ground truth annotations in real-time, and were motivated by an ef-

fective monetary incentive for both detecting and producing successful deceptive behavior. This corpus enabled a comprehensive study of deceptive speech on a large scale. Using the CXD corpus, we identified acoustic-prosodic, lexical, and syntactic cues to deception, and trained machine learning classifiers to automatically identify deceptive speech using those cues. The development of strong performing deception classifiers, as well as the identified acoustic-prosodic and linguistic cues to deception, are key contributions of this thesis.

The second part of the thesis addresses individual differences in deceptive speech. We present a comprehensive analysis of differences in cues to deception across gender, native language (Standard American English and Mandarin Chinese), and personality traits (measured by the Five Factor model of personality). This work is the first large-scale analysis of gender, native language, and personality differences in acoustic-prosodic and linguistic cues to deception. We trained classification models to identify gender, native language, and personality traits from short samples of speech. These experiments were conducted for the purpose of providing speaker trait information for deception detection, but this work has implications beyond deception detection. For example, speaker trait identification can be very useful for speech analytics and personalization of human-machine interactions. Finally, we developed methods to leverage differences across speaker groups to improve deception classification performance. The methods introduced in this thesis for leveraging speaker differences in deception classification can be applied to other speech classification problems with variation across speakers.

This thesis is organized as follows. Part I introduces the Columbia X-Cultural Deception Corpus and presents an empirical analysis of acoustic-prosodic and linguistic cues to deception, as well as a series of deception classification experiments. Part II describes an empirical analysis of differences in cues to deception across gender, native language, and personality, and presents methods for leveraging these differences in deception classification. Part III discusses the conclusions and implications of this work.

Part I

Deception Detection from Text and Speech

Chapter 2

Motivation and Research Goals

Deception is the act of intentionally misleading others, in order to gain some advantage or avoid some penalty [Bok, 1999; Ford *et al.*, 1988; DePaulo *et al.*, 2003].

This definition of deception excludes falsehoods that result from of self-deception, pathological behavior, theater, or lies that are due to ignorance or error. When determining if a statement is deceptive, it is critical to consider the intention of the speaker. For example, the statement "it is raining outside" is not inherently truthful or deceptive. Suppose the speaker was outside a few minutes earlier when it was raining, and it has since stopped raining. If the speaker believes that it is still raining, we do not consider the statement to be deceptive. However, if the speaker is aware that it is sunny outside, and intends on misleading their interlocutor, the statement is deceptive.

Studies show that people lie frequently in daily life, with estimates as high as 2 lies per day [DePaulo *et al.*, 1996; Serota *et al.*, 2010]. These lies take place across various modalities in emails, phone calls, and face to face communication [Hancock, 2007]. Most of these daily lies fall under the category of low-stakes deception. These lies have little or no consequences for the deceiver, and are very difficult to detect. People often lie about their feelings, preferences, attitudes, and opinions, for various reasons. Sometimes people lie in order to make themselves seem better to others (e.g. "I was always a top student"), and other times they lie to avoid hurting others' feelings (e.g. "Great tie!"). These lies often go undetected, and there are minimal or no consequences if they are detected. Further, those that are lied to often want to believe the lie.

On the other hand, high-stakes deception takes place when there are serious consequences for the deceiver. In this less common category, there is a greater risk involved for the deceiver. Lying on a job resume, or calling in sick to work when you feel fine, has higher stakes – one may risk job termination if they are discovered. Lying to a judge about committing a crime or to a TSA agent about your travel plans, has even higher stakes, and can result in jail time for the deceiver.

We are interested in high-stakes deception. Automatic detection of high-stakes deception is a major goal of law enforcement, military, and intelligence agencies as well as commercial organizations. Theoretical models of deception state that there is greater cognitive load for the deceiver under high-stakes deception [Ekman and Friesen, 1969; Zuckerman *et al.*, 1981; Vrij *et al.*, 2008]. Creating a lie is a more difficult task than recalling the truth, and it requires great effort to keep all of the details of a lie consistent, while simultaneously sounding credible [Zuckerman *et al.*, 1981]. Ekman and Friesen [1969] proposed that there are *leakage cues* during high-stakes deception that betray a deceiver's true thoughts. Leakage cues can be expressed in several modalities, such as facial expressions, body posture, and hand gestures. In this work we focus on verbal cues to deception.

In Part I of this thesis, we present our work on deception detection from text and speech. The overarching goal of this work is two-fold: firstly, we aim to develop automated methods to detect deceptive language. But perhaps more importantly, we aim to increase our understanding of deceptive language, by carefully studying the characteristics of deceptive and truthful language.

In order to accomplish these goals, we first created a large-scale corpus of deceptive and non-deceptive speech – the Columbia X-Cultural (CXD) Corpus, described in detail in Chapter 4. We created this corpus in order to conduct cross-cultural research of deceptive speech using a cleanly recorded and well-annotated dataset, on a scale that had not been previously possible. The corpus uses a fake-resume paradigm with a monetary incentive in order to mimic high-stakes deception in a laboratory setting. In addition to the findings of the studies that are presented in this section, the creation of this corpus is an important contribution of this thesis.

Using this new corpus, we aimed to answer the following main research questions:

What are the acoustic-prosodic and linguistic characteristics of truthful and deceptive speech? We used statistical methods to analyze the features of deceptive and truthful speech, highlighting significant differences and placing our findings in the context of prior work. We also analyzed novel features that had not been previously considered in deception research. Chapter 5 presents the results of this analysis.

Can we use machine learning classification techniques to automatically distinguish between truthful and deceptive speech? In Chapter 6, we present multiple deception classification experiments using a variety of acoustic-prosodic and linguistic features. We also provide insights into best practices for deception classification based on our experimental results. In Chapter 7 we present a detailed error analysis to understand what kinds of speech segments are easier and more difficult for our trained classifiers, and how classifier judgments compared with human judgments of deception.

Finally, we explored entrainment in deception for the first time. Chapter 8 presents a detailed analysis of entrainment in the CXD corpus, and its relationship with deceptive speech.

Chapter 3

Related Work

Efforts to develop methods to detect deception date back to ancient times. As documented by Ford [2006], people suspected of lying in China (1000 B.C.) were given raw rice to put in their mouths and then spit out. Based on the theory that decreased salivation is associated with anxiety, they were found guilty of deception if the rice was still dry. Modern technology has produced more sophisticated deception detection techniques. These methods are based on the observation that there are discernible physiological characteristics present when one is lying. There are a range of methods that aim to measure these characteristics using a variety of modalities: facial expressions, biometric indicators, body posture and gestures, brain imaging, body odor, as well as linguistic information. Each of these methods has advanced our knowledge of deceptive behaviors, but most of these approaches have not resulted in robust deception detection technologies.

There are challenges associated with several of these approaches. Analysis of facial expressions is difficult to automate, requiring expensive video capture technology, labor-intensive human annotation, and subsequent alignment with transcribed and semantically interpreted language to identify mismatches between "micro-expressions" and language. Biometric indicators such as heart rate and respiratory patterns have been commonly measured by the polygraph, which has been shown to perform no better than chance [Eriksson and Lacerda, 2007]. The signals captured by polygraphs are also correlated with anxiety and fear, which can be experienced by an innocent person who is hooked up to a polygraph and interrogated, leading to false positives. Additionally, there are known countermeasure

techniques to avoid detection by a polygraph. More recent attempts to measure biometric indicators involve the use of thermal imaging technology [Rajoub and Zwiggelaar, 2014]. This is promising, but the cost can be prohibitive for common use.

There have been promising results using automatic capture of body gestures, such as head and hand movements, as cues to deception [Lu *et al.*, 2005; Meservy *et al.*, 2005; Tsechpenakis *et al.*, 2005]. Again, these methods require multiple, high-caliber cameras to capture movements reliably and align them with speech. The use of brain imaging techniques for deception detection is still in its infancy [Meijer and Verschuere, 2017] and requires the use of MRI techniques not practical for general use. Body odor as an indicator of deception is in early stages and it is too early to say whether this area of research will prove useful [Li, 2014].

Previous work on language cues to deception include text-based and speech-based studies. Language cues have the advantage of being inexpensive, non-invasive, and easy to collect. And importantly, prior studies examining linguistic cues to deception have yielded promising results. This thesis focuses on language-based cues to deception. In this chapter we begin by reviewing theoretical models of deception. We then review previous deception detection studies from text and speech. We conclude by discussing the gaps in the literature that this thesis aims to fill.

3.1 Deception Theory

The first influential theoretical paper on deception was published by Ekman and Friesen [1969]. They proposed two categories of cues: leakage cues and deception cues. Leakage cues betray a deceiver's true feelings, while deception cues indicate that deception is occurring, but do not convey the nature of the lie. For example, a micro-expression (a facial expression lasting for a fraction of a second) can be a leakage cue if a person attempts deceive someone that they are feeling happy and a flash of sadness appears on their face. A deception cue can be an inconsistency in one's story that alerts the listener that something is not right, but it does not explicitly convey the truth.

They hypothesize that leakage occurs because the deceiver feels guilty about their de-

ception, and subconsciously wants to be caught lying. They also describe factors that affect the presence of cues to deception as well as the success of the deceiver. For example, they emphasize the role that stakes play in cues to deception. Cues to deception are not salient in situations where stakes are low, such as when telling a white lie, or playing a game without reward. The success of the deceiver is affected by the psychology of the deceiver and their target. For example, they hypothesize that asymmetric deception, where the deceiver is highly motivated to deceive but the target is not focused on detecting deception, is more likely to succeed than symmetric deception, where the deceiver and target are focused on deception and deception detection, respectively. Ekman's work has been influential in practiced law enforcement, particularly in the area of identifying deception from facial expressions. He has created training courses to teach practitioners how to identify micro-expressions. More broadly, the idea that deception and leakage cues exist is the basis for much of the deception detection research, where the goal is to identify and interpret these cues.

Ekman's theoretical work is supported by the theory of cognitive deception, proposed by Zuckerman *et al.* [1981] and extended by Vrij *et al.* [2008]. Zuckerman *et al.* [1981] proposed that deceiving is a more cognitively complex task than truth-telling. Creating a lie is more difficult than simply recalling the truth. The liar must construct a story with details that are consistent with each other and also consistent with the knowledge of the listener. Based on this theory, an increase in cognitive load when lying can result in cues to deception such as increased response latencies, more speech disfluencies and hesitations, and a reduction in complexity of language. Another hypothesis is that increased cognitive load when lying results in a decrease in hand and leg movements.

Vrij *et al.* [2008] extended this theory. Instead of assuming that the act of deception increases cognitive load enough to have an observable effect on deception cues, they proposed that imposing cognitive load on a potential deceiver is a method to highlight cognitive differences between lying and truth-telling. For example, an interviewer can ask an interviewee to tell a story in reverse order. Because a deceiver is using more cognitive resources to create and maintain a lie, he will have fewer resources remaining to perform a cognitively complex task than a truth-teller. They validate this theory with laboratory experiments, showing

that interviewers performed better at deception detection when they imposed cognitive load on interviewees.

One of the most widely known theories of deception is Interpersonal Deception Theory (IDT), developed by Buller and Burgoon [1996] from the perspective of communication theory. IDT highlights the role of interactivity in deceptive behavior, and states that interactive deception is fundamentally different from noninteractive: in an interactive communication the deceiver is constantly updating strategy to reflect feedback from the receiver, while noninteractive communication is static and has no explicit receiver of the deception. They hypothesize that the degree of interaction in a given communication interface will affect the communication process and outcomes such as trust, honesty, and credibility. In general, the more interactivity in a communication, the greater the trust and perceived honesty. According to IDT, the deceiver's motivation for deception is an important moderator of deception cues. They differentiate between three forms of motivation for deception: instrumental, relational, and identity. They hypothesize that deceivers who are motivated instrumentally would have the greatest fear of getting caught, and would therefore exhibit more cues to deception (which they term arousal cues) than someone motivated by relational or identity goals.

DePaulo *et al.* [2003] describe a self-presentational perspective for understanding deception. They argue that the vast majority of lies are told, not for material gain or escape from punishment, but for psychological benefits. People lie to make themselves appear more sophisticated or more virtuous, or to protect themselves or others from disapproval. Although truth-tellers often have these motivations, they attempt to achieve these goals within the framework of honesty, while liars use deception to achieve these goals. Cues to deception in everyday life tend to be weak, and this self-presentational theory proposes that the strength of these cues is moderated by the self-presentational processes involved in communicating truths and lies. Because of the discrepancy between a liar's story and their true beliefs, deceptive self-presentations will be less convincing than truthful ones, and they will have a greater sense of deliberateness. Deceivers will also appear less forthcoming, because they fear being questioned on details, and also because they are less familiar with their stories than truth-tellers. Because of moral misgivings and discomfort from lying, deceivers will be

less pleasant and more tense. They argue that a meta-analysis of many deception studies provides evidence for this self-presentational theory of deception.

3.2 Deception Detection from Language

Language has been a fruitful area of deception research. Language, both oral and written, is the primary way that humans communicate, and it is a natural modality to study deceptive communication. Many of the deception theories outlined above have implications for deceptive and truthful language. Compared to other modalities for examining cues to deception, there are several advantages to exploring language: it is relatively easy, inexpensive, and non-invasive to collect, and several studies have found that there are salient cues to deception in language. Here we review some prominent studies on cues to deception from text and from speech.

3.2.1 Text-based Deception Detection

Several practitioners and researchers have examined text-based cues to deception. Deceptive texts can have many different forms. It can be transcribed speech (e.g. from a witness testimony in court), formal writing (e.g. a letter, newspaper article), or informal writing (e.g. a social media message). Deceptive texts can be found in multiple domains on various topics, and researchers have studied many different kinds of texts.

Early text-based deception detection methods include Statement Analysis [Adams, 1996] and SCAN (Scientific Content Aalysis) [Smith, 2001]. These are text-analysis techniques that combine lexical and syntactic features, such as word tense and part of speech distributions to determine whether a text is deceptive or truthful. The intuition behind these approaches is that there are often many ways to phrase a particular message, and the specific choices that a speaker makes can contain deception cues. This is rooted in the theory of "leakage cues" proposed by Ekman and Friesen [1969]. These two approaches, along with other text-based signals identified by Reid and associates [Buckley, 2000], have been popular efforts among law enforcement and military personnel, despite the lack of rigorous validation for these approaches. The methods have been developed based on case-studies and

intuition rather than empirical evidence. Although these methods were developed without scientific validation, they have been foundational in providing a set of features to be tested empirically by others. For example, Bachenko *et al.* [2008] partially automated some features used in Statement Analysis, and demonstrated successful evaluation of this approach on small amounts of written text, including criminal narratives and police interrogations.

An especially useful resource for text-based deception detection is the Linguistic Inquiry and Word Count (LIWC), developed by Pennebaker and King [1999]. LIWC groups words into psychologically motivated categories, and this tool has been used in a wide range of deception studies. Newman *et al.* [2003] showed that LIWC dimensions were useful for distinguishing between truthful and deceptive accounts in multiple domains: opinions on abortion, feelings about friends, and a mock crime scenario. Ott *et al.* [2011] used LIWC features as well as other linguistic features to detect deception in a crowd-sourced dataset of fake hotel reviews. Linguistic features such as n-grams and language complexity have been analyzed as cues to deception [Yancheva and Rudzicz, 2013; Pérez-Rosas and Mihalcea, 2015]. Syntactic features, such as part of speech tags, have also been found to be useful for structured data [Ott *et al.*, 2011; Feng *et al.*, 2012]. An important domain for text-based deception detection is online communication. Hancock [2007] researched deceptive text in online forums and online dating profiles [Hancock *et al.*, 2007b]. Zhou *et al.* [2004] used a variety of linguistic cues to identify deception in online text messages. Recently, there have been efforts to identify deception [Shu *et al.*, 2017] and satire [Rubin *et al.*, 2016] in the news media.

3.2.2 Speech-based Deception Detection

Relatively little work has been done on spoken cues to deception, although speech technologies have the advantage of being cheap and easily portable. Early methods to detect deception from speech centered around Voice Stress Analysis (VSA). This technology has been marketed in the past as a "lie-detector" but is now viewed as psuedo-science. The premise underlying the technology is that speech production is different when a person is experiencing stress. During normal speech, an inaudible low-frequency micro-tremor is produced, and during a stressful, condition the natural micro-tremor production is suppressed.

This hypothesis has been discredited – the existence of micro-tremors has not been validated, the connection between stress and deception is not clear, and VSA technology has not been effective at detecting deception [Horvath, 1982].

A few studies of deception have included audio analysis. Ekman *et al.* [1991] found a significant increase in pitch for deceptive speech over truthful speech. Streeter *et al.* [1977] reported similar results, with stronger findings for more highly motivated subjects. A meta-analysis DePaulo *et al.* [2003] identified cues to deception that were significant across many studies. Some of these cues were acoustic-prosodic, including duration, vocal tension, and pitch.

There have been few large scale computational approaches developed for detection of deception from speech. This is likely due to the lack of large, cleanly recorded corpora for deception. Hirschberg *et al.* [2005] created the first large scale corpus of deceptive speech, the Columbia-SRI-Colorado (CSC) corpus, comprising about 7 hours of subject speech. They empirically studied more sophisticated acoustic, prosodic, and lexico-syntactic features and found that acoustic-prosodic features are promising indicators of deception. More recently, Amiriparian *et al.* [2016] used emotion labels inferred from speech to detect deception, with some success.

3.3 Conclusions

Although there have been many studies of deception in text and some in speech, there is much that remains to be done. This thesis presents novel work on deception that helps fill gaps in prior deception research. Several of the previous studies were done on a very small scale, with only a handful of speakers analyzed. In this work we created a new corpus of deceptive and non-deceptive speech, comprised of over 120 hours of speech from 340 subjects. For comparison, the previously largest corpus of cleanly recorded deceptive speech – the CSC corpus – contained about 7 hours of subject speech from 32 speakers. This new corpus enabled research on a much larger scale. In addition, our corpus uniquely contains speech from both the deceiver and the target of the deception, playing the roles of an interviewee and interviewer respectively. This enabled a study of entrainment and

deception. Finally, many of the previous methods to detect deception used simple rule-based algorithms. Our large dataset allowed us to explore more sophisticated modeling approaches.

Chapter 4

Data and Features

This chapter describes the corpus, features, and units of analysis used in this thesis. Some of this work was published in Levitan *et al.* [2015a,b].

4.1 Columbia X-Cultural Deception Corpus

The Columbia X-Cultural Deception (CXD) Corpus is a collection of within-subject deceptive and non-deceptive speech from native speakers of Standard American English (SAE) and Mandarin Chinese (MC), all speaking in English. The corpus contains dialogues between 340 subjects, comprising 122 hours of subject speech. A variant of a fake resume paradigm was used to create the corpus. Previously unacquainted pairs of subjects played a lying game with each other. Each subject filled out a 24-item biographical questionnaire, where they were instructed to create false answers for a random half of the questions, following guidelines to ensure that their false answers differed significantly from the truth. Each subject also completed the NEO-FFI personality inventory [Costa and McCrae, 1989], and provided a 3-4 minute baseline sample of speech by answering open-ended questions truthfully. During the baseline session, an experimenter asked the subject open-ended questions (e.g."What do you like best/worst about living in New York City?"). Subjects were instructed to be truthful in answering.

The lying game took place in a sound-proof booth, with the two subjects seated across from each other, separated by a curtain to ensure no visual contact. For the first half of

the game, one subject assumed the role of the interviewer, while the other answered the biographical questions, lying for one half and being truthful for the other half (based on their questionnaire's configuration). For the second half of the game, everything was the same except the subjects' roles were reversed.

As interviewers, their goal was to try to identify when the interviewee was lying and when they were telling the truth. As interviewees, their goal was to try to convince their interviewer that everything they said was true. For motivation, they were told that their compensation depended on their ability to deceive while being interviewed, and to judge truth and lie correctly while interviewing. As interviewer, they received $1 each time they correctly identified an interviewee's answer as either lie or truth and lost $1 for each incorrect judgment. As interviewee, they earned $1 each time their lie was judged to be true, and lost $1 each time their lie was correctly judged to be a lie by the interviewer.

Figure 4.1: Setup of experiment in sound-proof booth.

During the game, the interviewer was allowed to ask the 24 questions in any order he or she chose; the interviewer was also encouraged to ask follow-up questions to aid them in determining the truth of the interviewee's answers. The speech from the game was recorded to digital audiotape, on separate channels for each speaker, using a Crown head-mounted close-talking microphone. For each question, the interviewer recorded their true/false judgment, along with a confidence score from 1-5. While answering the questions, the interviewee pressed the T or F key, providing a local veracity label for each utterance. After the game was completed, both subjects completed a brief questionnaire, reporting

on how well they thought they performed at deceiving their partner and at judging their partner's lies.

The advantages of this paradigm over other possible choices are:

- It allows subjects to choose the content of their own lies so the lies will be more genuine.
- It collects data on deception perception as well as production.
- It provides financial motivation for the interviewer and the interviewee, tailored to the interests of each role.
- It provides additional self-presentational motivation by pairing subjects with other subjects in indirect competition.

As explored later in Part II of this thesis, the CXD corpus is also ideal for the study of individual differences because of the cross-cultural nature of the data, as well as the demographic information and personality scores that were collected for each subject.

The entire corpus was orthographically transcribed using the Amazon Mechanical Turk (AMT)[1] crowd-sourcing platform, and the transcripts were force-aligned with the audio recordings using the Kaldi Speech Recognition Toolkit [Povey *et al.*, 2011]. Our collaborators at CUNY organized the transcription task and did the alignment. After the crowd-sourced transcription and automatic alignment were completed, there was substantial hand-correction done by Columbia and CUNY students. For example, we obtained three transcripts for each audio segment from three different crowd workers, and the three transcripts were combined using Rover techniques [Fiscus, 1997]. The rover combination produced a rover output score, measuring the agreement between the initial three transcripts. For clips with a score lower than 70%, transcripts were manually corrected. The transcripts of 9.7% of the clips need to be hand-corrected.

4.2 Units of Analysis

Throughout the thesis, we refer to the following units of analysis:

[1] https://www.mturk.com/mturk/

An **inter-pausal unit (IPU)** is defined as a pause-free segment of speech from a single speaker, with a pause length threshold of 50 ms. This threshold has been used in other speech research, derived from average stopgap length in speech corpora. Automatic IPU segmentation was done using Praat [Boersma and others, 2002] and was subsequently hand-corrected.

A **turn** is defined as a maximal sequence of IPUs from a single speaker without any interlocutor speech that is not a backchannel (a simple acknowledgment that is not an attempt to take the turn). Turn boundaries were identified by processing the IPUs of both the interviewer and interviewee. Non-backchannel overlaps between speakers were resolved by computing the average distance between IPUs within turns for each speaker. Using that metric, we determine whether the overlapped IPU should be concatenated with the previous IPU, the next IPU, or become an independent turn.

We also defined topical units of analysis, considering the 24 biographical questions as conversational topics. Consider the following dialogue:
Interviewer: What is the most you have ever spent on a pair of shoes?
Interviewee: It was a little more than five hundred dollars.
Interviewer: What did they look like and where did you wear them?
Interviewee: They were very nice Jimmy Choo shoes, blue with a three and a half inch heel, and I wore them to my sister's wedding.

A **question response** is an interviewee turn that is a direct answer to an interviewer question from the list of 24 biographical questions. In the above example, the question response is, "It was a little more than five hundred dollars."

A **question chunk** is a set of interviewee turns that are answers to an interviewer biographical question and its related follow-up questions. In the above example, the question chunk is, "It was a little more than five hundred dollars. They were very nice Jimmy Choo shoes, blue with a three and a half inch heel, and I wore them to my sister's wedding."

We developed a question identification system in order to annotate these topical segments. The details of the system are described in Maredia *et al.* [2017]. It uses word embeddings to match semantically similar variations of questions to the target list of biographical questions. This was necessary because interviewers asked those questions using

different wording from the original list of questions. The question identification system obtained an F1-score of 95 on a set of hand-labeled turns. After identifying the interviewer turns that corresponded to biographical questions, we annotated question responses as the first interviewee turn following each biographical question. We annotated the set of interviewee turns between two interviewer questions, $q1$ and $q2$, as a question chunk corresponding to $q1$. We evaluated this segmentation method on a hand-annotated test set of 17 interview dialogues (about 10% of the corpus) consisting of 2,671 interviewee turns, 408 interviewer biographical questions, and 977 follow up questions. This approach resulted in 77.8% accuracy, with errors mostly due to turns that were unrelated to any question.

A summary of the total number of each unit of analysis is shown in Table 4.1.

Unit	*Interviewer*	*Interviewee*	*Total*
IPU	81,536	111,428	192,964
Turn	41,768	43,673	85,459
Question Response	8,092	8,092	16,184
Question Chunk	8,092	8,092	16,184

Table 4.1: Number of interviewer and interviewee segmentation units: IPUs, turns, and question segments.

4.3 Ground Truth Annotation

One of the greatest challenges in researching deceptive speech is obtaining high-quality ground truth annotations of deception. Unlike other paralinguistic information that can be labeled by human annotators (e.g. emotion), deception labels must be provided by the deceiver. Following Enos *et al.* [2007], we distinguish between two forms of deception annotation: **global** and **local**. Global deception refers to the veracity of a salient discourse topic, while local deception refers to the veracity of utterances that are spoken in support of a topic. In the CXD corpus, the discourse topics are the 24 biographical questions that are used during interviews, and each interviewee response to a question is globally true or false. Interviewees often provided additional information to support their global

response, sometimes at the prompting of the interviewer in the form of follow-up questions, and sometimes voluntarily without prompting. The distinction between local and global deception is important because speakers do not always lie at a local level in support of a global lie.

Consider the following dialogue:
Interviewer: What is your mother's job?
Interviewee: My mother is a doctor (F). She has always worked very late hours and I felt neglected as a child (T).

Is the interviewee response true or false? Globally, the response to the question is deceptive. However, it contains local instances of both truth and deception. Participants were instructed to only tell the truth while answering a question truthfully – even in response to follow up questions. However, when answering a question deceptively, they were allowed to answer follow up questions with truthful responses.

4.3.1 Global Deception

Each participant filled out a biographical questionnaire at the beginning of the session, and the form indicated which questions should be answered truthfully and which should be answered deceptively. This was automatically generated to ensure that a random half of the questions would be answered deceptively. A sample biographical questionnaire is shown in Appendix A.4. After the game, the questionnaires were logged in a spreadsheet, allowing for easy retrieval of the global deception label of each interviewee response.

4.3.2 Local Deception

During the game, interviewees labeled their responses with local deception annotations by pressing a "T" or "F" key for each utterance as they spoke. A script running during the game captured the keypresses, which were then automatically aligned with the recording from the game. This approach allowed for automatic annotation of local deception on a large scale. However, there are several challenges involved when relying on participants to self-label their utterances. Some subjects neglected to label every utterance, resulting in speech segments without an associated local deception label. There were also several

instances of speech segments that contained both "T" and "F" keypresses. We developed three approaches for resolving these issues and evaluated the performance of the approaches on a subset of manually annotated data.

1. **High precision** This approach labels each segment with the keypresses that fall between the start and end time of the segment, and discards any segments that are either missing any labels or contain conflicting labels.

2. **High recall** This approach uses every segment in the corpus. It does this by resolving ambiguous segments in possibly noisy ways. If a segment is missing labels, it finds the keypress that appears closest to the segment, either before or after the segment. Segments that contain both true and false segments are labeled as deceptive.

3. **Mixed** This approach proposes a middle ground. It attempts to resolve ambiguity, but limits the distance of adjacent labels that can be used. This approach labels almost 90% of the segments.

Table 4.2 shows the percentage of data labeled and the accuracy of labels, for each labeling approach. The numbers shown are for turn segmentation units, but the trends are comparable for IPUs.

Approach	*Accuracy*	*% labeled*
High precision	98.4	58.9
High recall	92.4	100
Mixed	97.2	89.1

Table 4.2: Results of three veracity labeling approaches, evaluated using turn segmentation units.

We computed accuracy using a subset of turns from 20 sessions that were hand-labeled with TF labels. We hand-labeled the sessions by listening to the audio recording and using the adjacent keypresses and context along with the biographical questionnaire form to resolve ambiguous turns. Using the high precision approach results in the smallest amount of data, but the labels are clean. The high recall approach allows us to use all of the data,

albeit with noisy labels. And the mixed approach gives us most of the data, with somewhat noisy labels.

To determine which labeling approach was optimal for deception classification, we trained deception classifiers using each of the three labeling approaches, and then evaluated them on a gold standard test set with hand annotated veracity labels. The classifiers were trained using a combination of acoustic-prosodic and lexical features. We found that the classifier trained using the labels from the *mixed* approach yielded the best deception classification performance. The *high precision* approach was the second best performing, and the *high recall* approach was the worst, despite having the most training data. Thus, we used the *mixed* labeling approach for all turn and IPU classification experiments presented in this thesis.

In order to maximize the amount of data, we manually inspected the remaining approximately 10% unlabeled segments from the *mixed* approach and determined the veracity label using a combination of the global deception label and the context from the dialogue.

4.4 Features

Here we describe the features used for analysis and classification of deception as discussed in this thesis.

4.4.1 Acoustic-Prosodic Features

We examined two sets of acoustic-prosodic features.

- Interspeech 2009 Emotion Challenge feature set (IS09)
- Praat acoustic-prosodic features set (Praat-15)

The IS09 feature set [Schuller *et al.*, 2009] contains 384 features extracted using openSMILE [Eyben *et al.*, 2010], designed for the task of emotion recognition. Because emotion and deception are related (e.g. emotion features can predict deception [Amiriparian *et al.*, 2016]), we hypothesized that the IS09 emotion feature set will be useful for deception detection. The features were computed from various functionals over low-level descriptor (LLD)

contours, including prosodic, spectral, and voice quality features. The LLDs and functionals used are shown in Table 4.3. There are 16 LLDs: (1) zero-crossing-rate (ZCR) from the time signal, (2) root mean square (RMS) frame energy, (3) pitch frequency (F0), (4) Noise-to-Harmonics ratio (NHR), and (5-16) 12 mel-frequency cepstral coefficients (MFCC). In addition, delta coefficients for each LLD were computed, for a total of 32 LLDs. A set of 12 functionals are applied to each of the LLDs, for a total of $32 \cdot 12 = 284$ features.

LLD (16 · 2)	*Functionals (12)*
(Δ) ZCR	mean
(Δ) RMS energy	standard deviation
(Δ) F0	kurtusis, skewness
(Δ) NHR	extremes: value, relative position, range
(Δ) MFCC 1-12	linear regression: offset, slope, MSE

Table 4.3: IS09 features: low level descriptors and functionals.

Praat-15 is a set of 15 acoustic-prosodic features commonly used in speech analysis. Some are included in IS09, but these were extracted using Praat [Boersma and others, 2002], an open-source speech processing toolkit, and we focus on them for some of the analysis in this thesis. The 15 features are: (1-6) pitch {minimum, maximum, mean, median, standard deviation, mean absolute slope}, (7-10) intensity {minimum, maximum, mean, standard deviation}, (11) jitter, (12) shimmer, (13) noise-to-harmonics ratio (NHR), (14) speaking rate, and (15) duration. Several of these features have been proposed in the literature on deception as possible indicators of deception [DePaulo *et al.*, 2003].

Pitch refers to the fundamental frequency (f0) of the speech signal, or the frequency of vocal fold vibrations. It measures how high or low the frequency of a person's voice sounds. We computed the minimum, maximum, mean, median, and standard deviation of pitch values over a speech segment. We also computed the mean absolute slope for pitch, which is the average absolute slope across all turning points in a pitch contour.

Intensity (or energy) refers to the perceived loudness of a sound, and is measured by the amplitude of vocal fold vibrations. The greater the amplitude, the more energy is carried

by the wave, and the sound will have increased intensity. We computed the minimum, maximum, mean and standard deviation of intensity values over a speech segment.

Jitter, shimmer, and NHR are three measures of voice quality, variation in vocal fold behavior which affect listeners' perception of the harshness, creakiness, or breathiness of the voice. Jitter and shimmer are measures of f0 disturbance: jitter describes variation in frequency across cycles, and shimmer describes variation in amplitude. NHR measures the ratio between periodic and non-periodic components in a segment of voiced speech.

There are several ways to calculate speaking rate; in this work we estimated speaking rate by calculating the ratio of voiced to total frames.

Duration is calculated as $endtime - starttime$ for each segment, measured in seconds.

All acoustic-prosodic features were z-score normalized by speaker ($z = (x\text{-}\mu)/\sigma$; x = value, μ = speaker mean, σ = speaker standard deviation).

4.4.2 Lexical Features

We examined four sets of lexical features from the crowd-sourced transcriptions of the CXD corpus.

- Linguistic Inquiry and Word Count (LIWC)
- N-grams
- Word embeddings
- Linguistic Deception Indicators (LDI)

LIWC 2015[Pennebaker *et al.*, 2015b] is a text analysis program that computes word counts for 93 semantic classes. LIWC relies on an internal dictionary that maps words to psychologically motivated categories. When analyzing a target text, the program looks up the target words in the dictionary and computes frequencies for each of the 93 dimensions. The categories include standard linguistic dimensions (e.g. percentage of words that are pronouns, articles), markers of psychological processes (e.g. affect, social, cognitive words), punctuation categories (e.g. periods, commas), and formality measures (e.g. fillers, swear

words). LIWC dimensions have been used in many studies to predict outcomes including personality [Pennebaker and King, 1999], deception [Newman *et al.*, 2003], and health [Pennebaker *et al.*, 1997]. We extracted a total of 93 features using LIWC 2015. A full description of these features is found in [Pennebaker *et al.*, 2015a].

N-grams are contiguous sequences of n tokens from text, and are commonly used in NLP applications to represent text. We extracted unigrams, bigrams, and trigrams from each speech segment, in order to examine differences in word usage between deceptive and truthful speech. Although it is standard practice for other applications, we did not remove stopwords from the corpus because we were interested in studying function word usage in deceptive and truthful speech. We extracted unigram, bigram, and trigram features, and used TF-IDF to weight the terms. Terms that appeared fewer than three times in the corpus were excluded.

Word embeddings are a distributed representation of words, where words are mapped to vectors of real numbers. We used GloVe [Pennington *et al.*, 2014] pre-trained word embeddings. GloVe is an unsupervised learning algorithm that uses a log-bilinear regression model based on global word co-occurrence counts in a training corpus. We used a model trained on 2 billion tweets to produce 200 dimension word vectors. Unlike n-gram features, word embeddings have been shown to capture semantic relationships between words, and are therefore very useful features for downstream NLP tasks.

Linguistic Deception Indicators (LDI) are a set of 28 linguistic features which we adopted from previous deception studies such as [Enos, 2009; Bachenko *et al.*, 2008; DePaulo *et al.*, 2003]. Included in this list are binary and numeric features capturing hedge words, filled pauses, laughter, complexity, contractions, and denials. We include Dictionary of Affect Language (DAL) [Whissell *et al.*, 1986] scores that measure the emotional meaning of texts, and a specificity score which measures level of detail [Li and Nenkova, 2015]. The full list of LDI features is shown in Table 4.4. Some of the features were computed using lexicons of hedge words and cue phrases. The lists of these word categories are found in Appendix C. The hedge lexicon was developed by Ulinski *et al.* [2018]. Laughter labels were manually annotated during IPU segmentation correction.

Name	*Description*
hasAbsolutelyReally	Contains either the word absolutely or the word really
hasContraction	Has an apostraphe
hasI	Contains I
hasWe	Contains the word we
hasYes	Contains the word yes
hasNAposT	contains n't
hasNo	Contains the word no
hasNot	contains the word not
isJustYes	Only contains the word yes and no other words
isJustNo	Only contains the word no and no other words
noYesOrNo	Does not contain the word yes or no
specificDenial	Contains "I didn't" or "I did not"
thirdPersonPronouns	Contains third person pronouns
hasFalseStart	Contains a word that is cut off in middle
hasFilledPause	Contains a filled pause
numFilledPauses	Number of filled pauses
hasCuePhrase	Contains a cue phrase
numCuePhrases	Number of cue phrases
hasHedgePhrase	Contains a hedge phrase
numHedgePhrases	Number of hedge phrases
hasLaugh	Contains laughter
numLaugh	Number of laughter instances
complexity	# syllables / # words
DAL-wc	# words that appear in the DAL dictionary
DAL-pleasant	DAL pleasantness score
DAL-activate	DAL activation score
DAL-imagery	DAL imagery score
specScores	Specificity score

Table 4.4: LDI features: linguistic deception indicators.

4.4.3 Syntactic Features

Syntactic features are a set of features that we developed based on previous studies of syntax in deceptive speech. They include the following feature sets:

- Complexity
- Part-of-speech tags (POS)
- Part-of-speech tags and words (POS+word)
- Production rules, unlexicalized (PR-unlex)
- Production rules, lexicalized (PR-lex)
- Grandparent annotated production rules, unlexicalized (Grand-PR-unlex)
- Grandparent annotated production rules, lexicalized (Grand-PR-lex)

Complexity features were extracted using a system for automatic syntactic complexity analysis, described in [Lu, 2010]. There are 23 complexity features in total. These include nine features representing the number of words (W), sentences (S), verb phrases (VP), clasues (C), t-units (T), dependency clauses (DC), complex t-units (CT), coordinate phrases (CP), and complex nominals (CN). In addition, there are 14 measures of syntactic complexity shown in Table 4.5. These features are based on measures that are used to evaluate second language proficiency.

Measure	*Code*	*Definition*
Mean length of clause	MLC	# words / # clauses
Mean length of sentence	MLS	# words / # sentences
Mean length of T-unit	MLT	# words / # T-units
Sentence complexity ratio	C/S	# clauses / # sentences
T-unit complexity ratio	C/T	# clauses / # T-units
Complex T-unit ratio	CT/T	# complex T-units / # T-units
Dependency clause ratio	DC/C	# dep. clauses / # clauses
Dependency clauses per T-unit	DC/T	# dep. clauses / # T-units
Coordinate phrases per clause	CP/C	# coordinate phrases / # clauses
Coordinate phrases per T-unit	CP/T	# coordinate phrases / # T-units
Sentence coordination ratio	T/S	# T-units / # sentences
Complex nominals per clause	CN/C	# complex nominals / # clauses
Complex nominals per T-unit	CN/T	# complex nominals / # T-units
Verb phrases per T-unit	VP/T	# verb phrases / # T-units

Table 4.5: Syntactic complexity features.

The remaining six syntactic feature sets were obtained using the Stanford parser [Chen and Manning, 2014].

The part-of-speech tags (POS) feature set consists of n-grams that use POS tags as tokens instead of words. The part-of-speech and word (POS+word) feature set is n-grams where the tokens are POS tags concatenated with their corresponding words. The POS tag set used is the Penn Treebank tag set. A list of the tags and their descriptions is found in Appendix B.

We also extracted four sets of deep syntactic features derived from the dependency parse trees. These features were adapted from Feng *et al.* [2012]. Unlexicalized production rules (PR-unlex) are all production rules in the parse tree, except for those with terminal nodes. Lexicalized production rules (PR-lex) are all production rules, including those with terminal nodes. Grandparent annotated unlexicalized production rules (Grand-PR-unlex) are unlex-

icalized production rules combined with the grandparent node, and grandparent annotated lexicalized production rules (Grand-PR-lex) are lexicalized production rules combined with the grandparent node. All of these features are represented as n-grams, where each token is a production rule.

Chapter 5

Feature Analysis

The CXD corpus allowed us to analyze deceptive speech on a scale that had not been previously possible. This chapter takes a close look at the features that are representative of truthful and deceptive speech. This work aims to answer the following question: *What are the differences in acoustic-prosodic and linguistic features between truthful and deceptive interviewee responses?*

Many previous studies have reported classification performance with particular feature sets, and some include ablation studies or feature ranking experiments to highlight which features contribute the most to deception classification, but few studies include a careful analysis of the characteristics of truthful and deceptive speech. Such analysis is critical for furthering our scientific understanding of deceptive language.

5.1 Method

In order to analyze the differences between deceptive and truthful speech, we considered features extracted from two segmentation units: 1) question responses; and 2) question chunks. We chose these segmentations because they allow us to study some linguistic properties that require contextual information (unlike the shorter IPU and turn segmentations). All features were z-normalized by speaker, so that features represent distance from a speaker's mean, measured in standard deviations. We then calculated a series of paired t-tests comparing the mean feature values of two groups of segments: truthful and

deceptive.

All tests for significance correct for family-wise Type I error by controlling the false discovery rate (FDR) at $\alpha = 0.05$. The k^{th} smallest p value is considered significant if it is less than $\frac{k*\alpha}{n}$.

In all the tables in this chapter, D indicates that a feature was significantly increased in deceptive speech, and T indicates a significant indicator of truth. We consider a result to approach significance if its uncorrected p value is $<= 0.05$ and indicate this with parentheses (e.g. "(D)") in the tables.

Some of this work was published in Levitan *et al.* [2018a,b] and was done in collaboration with my co-authors.

5.2 Acoustic-Prosodic Analysis

In this section we present the results of our analysis of acoustic-prosodic characteristics of truthful and deceptive interviewee responses. The following 8 acoustic-prosodic features were examined: pitch max, pitch mean, intensity max, intensity mean, speaking rate, jitter, shimmer, and noise-to-harmonics ratio (NHR). These features are described in detail in Chapter 4, Section 4.4.1.

Table 5.1 shows the t-test results for the question response segmentation analysis. Question responses consist of the set of first interviewee turns in response to the 24 biographical questions.

Feature	*t*	*df*	*p*	*Sig.*
Pitch Max	5.32	8053	1.10E-07	D
Pitch Mean	0.62	8048	0.53	
Intensity Max	7.58	8066	3.90E-14	D
Intensity Mean	1.52	8070	0.13	
Speaking Rate	-1.87	8082	0.062	
Jitter	-1.08	7691	0.28	
Shimmer	-2.16	7659	0.031	(T)
NHR	0.46	8022	0.64	

Table 5.1: Differences in mean acoustic-prosodic features in truthful and deceptive interviewee question responses. D=increased in deceptive speech, T=increased in truthful speech.

This table shows that on average, deceptive interviewee responses were characterized by an increase in pitch max, as well as an increase in intensity max, compared with truthful responses. This suggests that speakers on average tended to speak with a higher pitch and louder volume when lying than when telling the truth. There was also a trend of increased shimmer in truthful speech, but this was not statistically significant after correcting for multiple comparisons.

Table 5.2 shows the same analysis comparing acoustic-prosodic features in truthful and deceptive speech, but this time for the question chunk segmentation. A question chunk is a set of interviewee turns that are answers to an interviewer biographical question and its related follow-up questions. The acoustic-prosodic features for question chunks were computed by averaging the turn-level features within the question chunk.

Feature	*t*	*df*	*p*	*Sig.*
Pitch Max	6.88	8064	6.40E-12	D
Pitch Mean	-1.75	8071	0.081	
Intensity Max	9.33	8076	1.40E-20	D
Intensity Mean	3.56	8064	0.00037	D
Speaking Rate	-0.77	8017	0.44	
Jitter	-1.73	8013	0.083	
Shimmer	-1.84	7982	0.066	
NHR	-0.07	8038	0.94	

Table 5.2: Differences in mean acoustic-prosodic features in truthful and deceptive interviewee question chunks. D=increased in deceptive speech, T=increased in truthful speech.

This table shows that the difference in truthful and deceptive speech were consistent in both the initial interviewee response and in the entire question chunk (which includes responses to follow up questions). In both segmentations, deceptive responses are characterized by an increase in pitch max and intensity max. There was also increased intensity mean in deceptive question chunks.

These results align well with previous studies of the acoustic-prosodic characteristics of deceptive speech. Several studies of deceptive speech reported an increase in voice pitch during deception [Ekman *et al.*, 1976; Streeter *et al.*, 1977]. In their meta-analysis of deceptive speech research, DePaulo *et al.* [2003] reported a significant effect of increased pitch in deceptive speech. Fewer studies have included an analysis of speech intensity. Chittaranjan and Hung [2010] found that the distribution of both pitch and energy values were higher for deceivers. However, DePaulo *et al.* [2003] reported no significant effect of energy in deceptive speech. They also note that cues to deception across multiple studies are generally quite weak. This is due to several factors, including differences in experiment design, and importantly, inter-speaker variability. Although we observe some significant findings when analyzing the aggregated behavior of all speakers, it is important to note that these trends are not true for all speakers. There has been little work done to understand this variability – e.g. why do some speakers raise their pitch while lying and some lower

their pitch? We explore these differences across speakers in detail in Part II of this thesis.

5.3 Linguistic Analysis

This section summarizes the results of our analysis of linguistic characteristics of truthful and deceptive interviewee responses. The following feature sets were examined: Linguistic Deception Indicators (LDI), Linguistic Inquiry and Word Count (LIWC), and syntactic complexity. These features are described in detail in Chapter 4, Section 4.4.2.

5.3.1 Linguistic Deception Indicators

We analyzed the set of 28 LDI features. The features were z-score normalized per speaker, so that each feature represented the speaker's distance from their mean feature value, measured in standard deviations. Paired t-tests were computed between the feature values in truthful and deceptive segments. This approach was applied to features extracted from both question response segments and question chunk segments. Table 5.3 shows the t-test results for the LDI features extracted from the question response segmentation.

Feature	*t*	*df*	*p*	*Sig.*
hasAbsolutelyReally	3.38	7660	0.00072	D
hasContraction	-0.46	8069	0.65	
hasI	3.54	8036	0.0004	D
hasWe	3.54	7236	0.0004	D
hasYes	5.69	7857	1.30E-08	D
hasNAposT	0.2	8057	0.84	
hasNo	-12.04	7938	4.20E-33	T
hasNot	-3.01	8012	0.0027	T
isJustYes	1.05	8016	0.29	
isJustNo	-9.73	7522	3.00E-22	T
noYesOrNo	5.88	8082	4.20E-09	D
specificDenial	-0.8	8053	0.42	
thirdPersonPronouns	4.34	7661	0.000014	D
hasFalseStart	2.75	7800	0.006	D
hasFilledPause	6.76	8051	1.50E-11	D
numFilledPauses	7.4	7430	1.50E-13	D
hasCuePhrase	-2.91	8060	0.0036	T
numCuePhrases	3.63	7269	0.00029	D
hasHedgePhrase	3.38	7972	0.00074	D
numHedgePhrases	4.33	7253	0.000015	D
hasLaugh	1.79	7960	0.073	
complexity	-0.73	8072	0.47	
numLaugh	1.37	7968	0.17	
DAL.wc	-3.06	8056	0.0022	T
DAL.pleasant	7.42	8051	1.30E-13	D
DAL.activate	-3.54	8062	0.0004	T
DAL.imagery	3.93	8011	0.000087	D
specScores	5.73	7897	1.00E-08	D

Table 5.3: Differences in mean LDI numeric features in truthful and deceptive interviewee question responses. D=increased in deceptive speech, T=increased in truthful speech.

Of the 28 LDI features, 21 were significantly different in truthful and deceptive interviewee responses. Deceptive interviewee responses had higher DAL imagery scores ($DAL-imagery$), which indicate words that are used to create vivid descriptions. They also had higher specificity scores ($specScores$), indicating that deceptive responses contained more detailed language than truthful responses. This is somewhat counterintuitive – deceptive responses describe events that did not occur, so one might assume that the language would be less descriptive and detailed. However, Malone *et al.* [1997] asked liars about the origins of their lies, and found that most said that they used their own experiences but altered critical details. Thus, they were able to create vivid and detailed stories that were very similar to truthful events that occurred. Additionally, it is conceivable that deceivers would attempt to conceal their deception by overcompensating with very detailed descriptions.

Consistent with DePaulo *et al.* [2003], deceptive responses had significantly more filled pauses ($hasFilledPause$, $numFilledPauses$) than truthful responses. Deceivers are hypothesized to experience an increase in cognitive load [Vrij *et al.*, 1996], and this can result in difficulties in speech planning, which can be signaled by filled pauses. Although Benus *et al.* [2006] found that, in general, the use of pauses correlates more with truthful than with deceptive speech, filled pauses such as "um" were increased in deceptive speech in the CXD corpus. Deceptive responses also had more false starts than truthful responses ($hasFalseStart$), which supports the theory that deceptive responses contain more disfluencies.

Hedge words and phrases ($hasHedgePhrase$, $numHedgePhrases$), which speakers use to distance themselves from a proposition, were more frequent in deceptive speech. This is consistent with Statement Analysis Adams [1996], which posits that hedge words are used in deceptive statements to intentionally create vagueness that obscures facts.

Deceivers used more cue phrases ($numCuePhrases$) when lying than when telling the truth. This feature captures 34 discourse markers such as "ok," "also," and "basically," and this is consistent with previous work that suggest that deceptive speech should contain more cue phrases [Adams, 1996; Enos, 2009]. Consistent with DePaulo *et al.* [2003] and Hancock *et al.* [2004], deceptive responses had a higher rate of third person pronouns ($thirdPersonPronouns$). However, the binary feature $hasCuePhrase$ was in-

creased in truthful responses. This is interesting because deceptive responses had on average more cue phrases per response than truthful responses. Although $hasCuePhrase$ and $numCuePhrases$ are strongly related, it seems that truthful responses are more likely to contain a cue phrase, and that deceptive responses contain on average more cue phrases than truthful responses. Examining binary as well as numeric features can add additional insight into the linguistic characteristics of deception.

DAL pleasantness scores ($DAL-pleasant$), which rate words on a scale from unpleasant to pleasant, were higher in deceptive responses. Previous studies have produced mixed results regarding emotional content of deceptive speech. Consistent with our findings, Enos [2009] report an increase of positive emotion words in deceptive speech. Burgoon *et al.* [2003] report an increase in all emotion words, both positive and negative, in deceptive speech. However, Newman *et al.* [2003] found that negative emotion words were increased in deceptive speech. Ott *et al.* [2011] point out that the goal of the deceiver affects the emotional content of his lies. In the context of fake hotel reviews, they found that positive emotion words were more frequent in deceptive reviews, where the goal is clearly to create a fake positive review.

Deceptive responses had higher values for the feature $hasAbsolutelyReally$, which is true if a response contains the word "absolutely" or "really." These words typically convey certainty, so this finding seemingly contradicts the increase in hedge words in deceptive speech. However, upon closer analysis, we found that many of these responses included negations such as "not really," which is a hedge phrase and does not convey certainty.

Although there was an increase in third person pronouns in deceptive speech, there were also greater frequencies of first person pronouns in deceptive speech, including "I" and "we" ($hasI$, $hasWe$). Ott *et al.* [2011] found that deceptive text had more pronouns overall, similar to imaginative writing (rather than informative writing).

Truthful responses had higher DAL activation scores ($DAL - activate$) than deceptive responses, and also included more words per response than were found in the DAL dictionary ($DAL - wc$). This suggests that truth-tellers used language that was more active, and also more commonly found in the DAL dictionary, so the DAL scores are more reliable for truthful responses than deceptive responses.

Three features capturing negation, *hasNo*, *hasNot*, and *isJustNo* all had greater frequencies in truthful responses. Other studies have reported the opposite: deceptive responses tend to contain more negation [Newman *et al.*, 2003]. As with emotion, this is likely a domain dependent phenomenon. For example, Fornaciari and Poesio [2013] studied criminal testimony in Italian court cases, and found an increase in negative adverbs such as "no" and "not" in deceptive statements. In that domain, deceptive statements involve denial of committing crimes and require heavy use of negation. In the CXD corpus, subjects are asked a variety of biographical questions, and some questions are more likely to be true when negation is used. For example, "Have you ever watched a person or a pet die?" "Have your parents divorced?" and "Have you ever gotten into trouble with the police?" were all more likely to be true in the negative in the CXD corpus. Thus, this trend of negation words in truthful responses might be an effect of the domain rather than a reliable indicator of truthful speech.

The same analysis was applied to features extracted from question chunks. Table 5.4 shows the t-test results for the LDI features extracted from the question response chunks.

Feature	*t*	*df*	*p*	*Sig.*
hasAbsolutelyReally	8.24	7838	2.00E-16	D
hasContraction	6.26	8074	4.00E-10	D
hasI	8.01	8064	1.30E-15	D
hasWe	7.89	7688	3.60E-15	D
hasYes	4.43	7991	9.40E-06	D
hasNAposT	4.7	8027	2.70E-06	D
hasNo	-7.86	8075	4.50E-15	T
hasNot	2.75	8018	0.0061	D
isJustYes	-0.36	8077	0.72	
isJustNo	-6.99	7234	3.00E-12	T
noYesOrNo	4.8	8041	1.60E-06	D
specificDenial	1.92	7935	0.055	
thirdPersonPronouns	9.72	7911	3.20E-22	D
hasFalseStart	6.27	7946	3.70E-10	D
hasFilledPause	8.55	8041	1.50E-17	D
numFilledPauses	8.28	7730	1.40E-16	D
hasCuePhrase	1.6	8077	0.11	
numCuePhrases	9.26	7776	2.70E-20	D
hasHedgePhrase	8.17	8077	3.40E-16	D
numHedgePhrases	8.21	7871	2.60E-16	D
hasLaugh	1.39	8039	0.16	
complexity	-0.36	8019	0.72	
numLaugh	-0.63	7847	0.53	
DAL.wc	0.53	8069	0.6	
DAL.pleasant	7.66	7986	2.10E-14	D
DAL.activate	-8.73	8071	3.10E-18	T
DAL.imagery	6.72	8077	2.00E-11	D
specScores	12.77	8075	5.50E-37	D

Table 5.4: Differences in mean LDI features in truthful and deceptive interviewee question chunks. D=increased in deceptive speech, T=increased in truthful speech.

The results for features extracted from question chunks were almost identical to the results from the question responses. However, there were some notable differences. The *DAL.wc* feature was not significantly higher in truthful chunks than deceptive chunks. It seems that this effect is only found in the first turn of each interviewee responses, but not in the full set of interviewee turns.

The use of contractions was not an indicator of deception or truth in question responses, but *hasContraction* and *hasNAposT* were increased in deceptive chunks. In their training course on interrogation and interviewing techniques, Inbau *et al.* [2011] posit that contractions are a sign of truthful speech, since it is assumed to be more natural to say something like "I didn't do it" than "I did not do it." The opposite trend was found for the CXD corpus. Perhaps in an effort to sound casual during deception, people used contractions more frequently. We note that contractions are largely used by native English speakers and not L2 speakers, and we explore differences in cues to deception between native and non-native speakers in Section II of this thesis.

Another difference between the analysis of question responses and question chunks is that for question responses, the *hasNot* feature was increased in truthful responses. For question chunks, this feature was increased in deceptive chunks. Using this form of negation was associated with deception in an interviewee's initial response, but this effect was not seen when analyzing the question chunk segmentation. Similarly, *hasCuePhrase* was increased in truthful question responses, but this feature was not significantly different between truthful and deceptive question chunks. Use of a cue phrase in one's first response was associated with truth, but there was no such effect in the full chunk of interviewee responses.

These findings suggest that when studying deceptive responses, there is a difference between a speaker's initial response (i.e. their first turn) and how they respond to follow up question. Using the word "not" in one's first response was an indicator of truth, while in a follow up question it was an indicator of deception. It is important to keep this in mind when analyzing cues to deception.

5.3.2 LIWC

We analyzed the set of LIWC features. To avoid noise, LIWC features that did not appear in at least 10% of question response segments were eliminated. This reduced the analysis to 42 of the 93 LIWC dimensions for question responses, and 77 LIWC dimensions for question chunks. For example, seven punctuation categories (parentheses, exclamation mark, colon, semi colon, quotation mark, period, and comma) were excluded from this analysis because they do not appear in the corpus. This is due to the orthographic transcription scheme used, which did not include punctuation transcription. The features represent normalized frequencies of words in each semantic category in a given text segment. Paired t-tests were used to compare the mean frequencies of these semantic categories in truthful and deceptive samples.

Table 5.5 shows the t-test results for the LIWC features extracted from the question response segmentation.

Feature	*t*	*df*	*p*	*Sig.*
adj	-0.17	7181	0.87	
adverb	1.03	7352	0.3	
affect	1.75	7416	0.08	
affiliation	1.52	7261	0.13	
allPunc	-1.09	7770	0.27	
analytic	7.72	8077	1.30E-14	D
apostro	-1.54	7612	0.12	
article	1.33	7844	0.18	
assent	5.59	7933	2.30E-08	D
authentic	1.53	8071	0.13	
auxverb	0.84	8046	0.4	
bio	1.72	6405	0.086	
clout	9.2	8067	4.60E-20	D
cogproc	-3.59	7655	0.00034	T
compare	1.45	6148	0.15	

conj	3.2	6984	0.0014	D
dic	-1.8	8015	0.071	
differ	-1.68	6905	0.094	
drives	2.5	7988	0.013	D
family	4.85	7341	1.20E-06	D
focuspast	4.27	7466	2.00E-05	D
focuspresent	-1.26	8059	0.21	
function.	-9.72	8068	3.20E-22	T
i	1.72	8060	0.085	
informal	6.31	8024	3.00E-10	D
insight	-0.83	5928	0.4	
ipron	0.13	7087	0.9	
negate	-15.2	7694	1.90E-51	T
netspeak	2.19	6500	0.029	(D)
nonflu	3.48	7794	0.00051	D
number	1.88	7971	0.06	
posemo	2.27	6928	0.023	(D)
ppron	3.71	8078	0.00021	D
prep	3.49	7580	0.00049	D
pronoun	3.36	8083	0.00077	D
relativ	1.04	8037	0.3	
sixltr	2.05	8073	0.04	(D)
social	4.26	7993	0.000021	D
space	2.84	7469	0.0045	D
tentat	-0.66	6164	0.51	
time	-0.31	7740	0.76	
Tone	2.3	7320	0.022	(D)
verb	1.86	8045	0.063	
WC	6.58	7960	4.90E-11	D
work	2.04	7548	0.041	(D)

WPS	6.52	7963	7.20E-11	D

Table 5.5: Differences in mean LIWC features in truthful and deceptive interviewee question responses. D=increased in deceptive speech, T=increased in truthful speech.

16 LIWC features were significantly higher on average in deceptive speech and three were significantly higher in truthful responses. There were also five features that approach significance and were increased in deceptive responses. Here we highlight some interesting findings.

Some of the LIWC results align well with the LDI results reported above. $nonflue$, which captures nonfluencies (e.g. "er," "hm," "um") was higher in deceptive responses, as was $informal$, which captures a range of informal language (e.g. swear words,"ok," nonfluencies, and fillers like "I mean"). Total *pronoun* use was higher in deceptive responses, and so was the use of *ppron* or personal pronouns such as "I," "we," and "her." Emotional *tone* trended towards more positive in deceptive responses, which complements our previous finding that DAL pleasantness scores were higher in deceptive speech. Words that signal *assent*, such as "agree," "yes," "okay" were more frequently used in deceptive responses. This is consistent with the above finding that $hasYes$ was associated with deceptive answers.

Frequencies of *clout* words, which show confidence and expertise were higher on average in deceptive speech. This is somewhat counter to the previous finding that hedge words were more frequent in deceptive speech. However, these findings are not mutually exclusive it is possible to express confidence while also using hedge words. For example, many deceptive responses contained the phrase "you know," and the word "know" is in our hedge phrase lexicon, but use of the word "you" appears in the LIWC lexicon for clout.

It is interesting that the $focuspast$ category, which includes words in the past tense, was more frequently used in deceptive speech. Verb tense is very important in statement analysis, and changes in verb tense are studied carefully. Statement analysis does not have general rules about verb tense and deception, rather it considers the verb tense in the context of the situation [Adams, 1996]. For example, use of past tense when referring to a missing person is suspicious, and in fact helped lead to the conviction of Susan Smith in the murder of her own children, when the FBI observed her say about them that they *needed*

her. Enos [2009] did not find a significant difference in use of past tense between deceptive and truthful responses in the CSC corpus.

Social words, which include references to family and friends, were increased in deceptive responses. Prepositions and conjunctions were more frequently used in deceptive responses, Deceptive responses had increased word count (WC) and words per sentence (WPS) on average. They also used more *analytic* language, or words that convey analytical thinking.

Consistent with the analysis of LDI features, truthful responses used more words from the *negate* dimension, which express negation. They also used more $function$ words, as well as more words from the *cogproc* category, which are associated with cognitive processes such as "cause" and "know."

This LIWC analysis was also conducted for the question chunk segmentation. Table 5.6 shows the t-test results for the LIWC features extracted from the question chunk segmentation.

Feature	*t*	*df*	*p*	*Sig.*
achieve	3.82	7555	0.00013	D
adj	3.72	8081	0.0002	D
adverb	5.37	8034	8.10E-08	D
affect	-0.52	7952	0.6	
affiliation	3.65	8022	0.00026	D
allPunc	-1.57	8053	0.12	
analytic	3.49	8084	0.00048	D
anger	2.46	7253	0.014	D
apostro	-2.03	8046	0.042	(T)
article	5.21	8061	2.00E-07	D
assent	-1.13	8052	0.26	
authentic	2.05	8084	0.04	(D)
auxverb	0.56	8042	0.58	
bio	2.84	7983	0.0045	D
body	3.5	7575	0.00046	D
cause	4.95	7725	7.50E-07	D

certain	-2.18	7794	0.03	T
clout	8.4	8062	5.20E-17	D
cogproc	-0.7	7989	0.49	
compare	5.1	7989	3.50E-07	D
conj	11.86	8010	3.60E-32	D
dash	3.61	7227	0.00031	D
dic	-3.09	8075	0.002	T
differ	-0.17	7929	0.86	
discrep	3.42	7504	0.00062	D
drives	6.61	8083	4.20E-11	D
family	3.62	7936	0.0003	D
feel	5.91	6593	3.60E-09	D
female	2.75	7970	0.006	D
focusfuture	3.93	7125	0.000086	D
focuspast	8.05	8006	9.40E-16	D
focuspresent	-0.32	8046	0.75	
friend	5	7267	5.80E-07	D
function.	-2.5	7924	0.012	T
health	2.29	7765	0.022	D
hear	1.34	7036	0.18	
home	0.97	7583	0.33	
i	1.73	8047	0.084	
informal	-3.21	7998	0.0014	T
insight	2.35	7957	0.019	D
interrog	5.63	7646	1.90E-08	D
ipron	6.8	8055	1.10E-11	D
leisure	2.08	7966	0.038	(D)
male	6.7	7744	2.30E-11	D
money	1.34	8024	0.18	
motion	7.92	7789	2.70E-15	D

negate	-15.49	7421	2.80E-53	T
negemo	1.71	7941	0.088	
netspeak	-2.8	7762	0.0052	T
nonflu	0.35	8001	0.73	
number	1.02	8083	0.31	
percept	5.09	7916	3.70E-07	D
posemo	-0.43	7945	0.67	
power	6.06	8014	1.40E-09	D
ppron	5.95	8032	2.80E-09	D
prep	9.46	8048	3.80E-21	D
pronoun	8.35	7969	7.70E-17	D
quant	2.26	7945	0.024	D
relativ	4.33	8072	0.000015	D
reward	6.97	7515	3.40E-12	D
risk	2.42	6482	0.016	D
sad	3.85	6585	0.00012	D
see	4.13	7270	0.000036	D
shehe	6.26	7827	4.10E-10	D
sixltr	1.25	7999	0.21	
social	6.8	8083	1.10E-11	D
space	4.59	8084	4.60E-06	D
tentat	1.27	7944	0.2	
they	3.66	7457	0.00025	D
time	0.64	8023	0.52	
tone	1.67	8050	0.095	
verb	5.54	8010	3.20E-08	D
WC	16.26	7932	1.70E-58	D
we	5.63	7180	1.80E-08	D
work	3.1	8072	0.0019	D
WPS	16.08	7928	2.60E-57	D

you	6.15	7043	8.40E-10	D

Table 5.6: Differences in mean LIWC features in truthful and deceptive interviewee question chunks.D=increased in deceptive speech, T=increased in truthful speech.

49 LIWC features were significantly higher on average in deceptive speech and six were significantly higher in truthful responses. There were also two features that approached significance and were increased in deceptive responses, and one that approached significance and was increased in truthful responses.

Overall, there were several additional significant differences in question chunks than in question responses. This is likely due to the fact that some LIWC categories did not appear often enough in the first turn, but when we aggregated all responses to follow up questions into question chunks and analyzed LIWC features with greater context, certain patterns were able to emerge.

For example, in question chunks, frequencies of *adjectives*, *adverbs*, and *articles* were increased in deceptive chunks. Frequencies of *verbs*, *we* and *you* frequencies were also increased in deceptive chunks, but this trend was not apparent in question responses. The LIWC language summary variables of *authenticity* and *adjectives* were indicators of deception: in an effort to sound more truthful and authentic, interviewees may have provided a level of detail that is uncharacteristic of truthful speech. This is consistent with the previous finding that deceptive responses had higher specificity scores, or more detailed language, than truthful responses.

Another trend seen only in chunks is that deceptive chunks had higher frequencies of *dash*, which are used in this corpus to indicate false starts. This form of disfluencies was only significantly increased in deceptive chunks, but not in responses. *Interrogatives* were also increased in deceptive chunks. In the context of the interviewer-interviewee paradigm, these are interviewee questions to the interviewer. Perhaps this was a technique used to stall so that interviewees had more time to develop an answer (e.g. "Can you repeat the question?") or to deflect the interviewer's attention from their deception and put the interviewer on the spot.

In addition to an increase in $focuspast$ words in deceptive chunks, there was an increase

in $focusfuture$ words, or verbs in future tense. This again signals the important role that verb tense plays in deception; however it is probably a phenomenon that is context dependent.

As with LDI features, it seems that there are some differences in indicators of truth and deception depending on whether we consider the initial interviewee response, or the full context of responses to follow up questions. These differences are important to keep in mind when applying these findings to new situations.

This LIWC analysis provides some insight into the characteristics of deceptive and truthful language. However, we note that there are areas to improve. LIWC relies on a dictionary that maps words to semantic categories, and there are many words in our corpus that do not appear in the dictionary. For example, our corpus includes many named entities, some misspelled words (due to transcription errors), and instances of laughter that are not represented in the LIWC dictionary. In addition, the dictionary uses regular expressions to match words, and we found some mistakes in these regular expressions. For example, one of the dictionary entries for the *religion* category is "monk*," which matched the word "monkey" in our corpus. A turn containing the term "Saint Louis" was given a high religion score because "Saint" is a religious word. Therefore, it is important to be careful about taking LIWC results at face value, and to carefully analyze the text to ensure that the LIWC scores are capturing what they should. (See Franklin [2015] for a more detailed discussion of LIWC concerns.)

In this work we compared average LIWC scores in truthful and deceptive speech, and reported which categories had significant differences. We did not state that certain LIWC categories imply deception, rather, we noted these differences and used them to hypothesize about the characteristics of deceptive speech in this domain. Further analysis of LIWC variables in multiple domains is necessary to make strong claims about the general nature of deceptive speech.

5.3.3 Syntactic Complexity

This section summarizes the results of the analysis of syntactic complexity features in truthful and deceptive interviewee responses. 23 syntactic complexity features were examined.

These features are described in detail in Chapter 4.

Table 5.7 shows the t-test results for the question response segmentation analysis and Table 5.8 shows the t-test results for the question chunk analysis.

Feature	*t*	*df*	*p*	*Sig.*
W	5.73	7361	1.00E-08	D
VP	6.09	7366	1.20E-09	D
C	6.37	7392	2.00E-10	D
T	5.54	8076	3.00E-08	D
DC	5.08	7422	3.90E-07	D
CT	3.81	8001	0.00014	D
CP	3.21	7296	0.0013	D
CN	5.49	7406	4.10E-08	D
MLS	5.68	7380	1.40E-08	D
MLT	6.19	7273	6.20E-10	D
MLC	4.58	8028	4.80E-06	D
C.S	6.31	7416	2.90E-10	D
VP.T	6.53	7271	6.90E-11	D
C.T	6.75	7347	1.60E-11	D
DC.C	3.53	8025	0.00042	D
DC.T	5.38	7330	7.80E-08	D
T.S	5.49	8080	4.20E-08	D
CT.T	4.07	7993	4.60E-05	D
CP.T	3.05	7384	0.0023	D
CP.C	1.43	7937	0.15	
CN.T	5.59	7376	2.40E-08	D
CN.C	3.69	8022	0.00023	D

Table 5.7: Differences in mean complexity features in truthful and deceptive interviewee question responses. D=increased in deceptive speech, T=increased in truthful speech.

Feature	*t*	*df*	*p*	*Sig.*
W	10.21	7732	2.50E-24	D
S	6.04	7961	1.70E-09	D
VP	10.03	7774	1.60E-23	D
C	9.8	7787	1.60E-22	D
T	7.41	7945	1.30E-13	D
DC	9.39	7628	7.70E-21	D
CT	8.71	7779	3.80E-18	D
CP	8.43	7615	4.10E-17	D
CN	9.31	7776	1.60E-20	D
MLS	10.24	7431	1.90E-24	D
MLT	9.36	7680	1.00E-20	D
MLC	6.24	8034	4.60E-10	D
C.S	10.52	7452	1.10E-25	D
VP.T	9.76	7645	2.30E-22	D
C.T	9.67	7636	5.50E-22	D
DC.C	8.43	8076	4.00E-17	D
DC.T	7.75	7791	1.10E-14	D
T.S	9.56	8080	1.60E-21	D
CT.T	7.43	8069	1.20E-13	D
CP.T	6.19	7656	6.40E-10	D
CP.C	2.49	7910	0.013	D
CN.T	7.42	7953	1.30E-13	D
CN.C	5.19	8080	2.20E-07	D

Table 5.8: Differences in mean complexity features in truthful and deceptive interviewee question chunks. D=increased in deceptive speech, T=increased in truthful speech.

These tables show that all measures of syntactic complexity that were examined are significantly different in truthful and deceptive chunks, and all except *CP.C* are significantly different in truthful and deceptive responses. Interestingly, these measures are all

significantly increased in deceptive responses, suggesting that deceptive responses are characterized by increased complexity, compared with truthful responses. This is somewhat counter-intuitive, and conflicts with the theory that lying increases cognitive load [Vrij *et al.*, 1996]. However, since the deception in the CXD corpus is partially premeditated (i.e. subjects had time to prepare their responses before the interview, but they did not know what follow up questions would be asked), it is possible that they did not experience an increase in cognitive load.

5.4 Discussion

This chapter aimed to answer the question: *What are the differences in acoustic-prosodic and linguistic features between truthful and deceptive interviewee responses?* We carefully analyzed the characteristics of truthful and deceptive speech in the CXD corpus. Two segmentation units, question responses and question chunks, were analyzed in order to study the differences in cues to deception and truth between the immediate response to a question and the responses to related follow-up questions. Using paired t-tests to compare feature means between truthful and deceptive speech segments, we studied acoustic-prosodic, lexical, and syntactic feature sets.

Acoustic-Prosodic Indicators Pitch max and intensity max were significantly increased in deceptive interviewee responses and chunks, and intensity mean was increased in deceptive interviewee chunks. Speakers on average tended to speak higher in their pitch range and with louder volume when they were lying. Increased pitch has been previously identified as a cue to deception [Ekman *et al.*, 1976; Streeter *et al.*, 1977], but few previous studies analyzed the relationship between intensity and deception.

Lexical Indicators 28 Linguistic Deception Indicator (LDI) features and 93 LIWC dimensions were analyzed in truthful and deceptive interviewee speech segments. Several patterns of deceptive language were apparent from this analysis, and many of the findings confirmed prior studies of deceptive language. For example, disfluencies such as false starts and filled pauses were increased in deceptive speech, as were hedge words and phrases; these findings confirmed previous work that identified these cues to deception [DePaulo

et al., 2003; Vrij *et al.*, 1996; Adams, 1996]. There was increased pronoun usage (I, we, third person pronouns) in deceptive speech, which is characteristic of imaginative writing Ott *et al.* [2011]. Other findings contradicted previous studies. For example, features capturing negation had greater frequencies in truthful responses, while previous studies found that there was more negation in deceptive statements. This is likely a domain-dependent phenomenon. In some situations, negation is likely to be deceptive. For example, in the context of an interrogation about a crime, a guilty suspect who is lying will deny their guilt with negation.

There were some differences in cues to deception between question responses and question chunks. This is an important distinction, and suggests that there are some cues present in a speaker's immediate response to a question, while others are only captured over a longer dialogue segment. In practice, it is possible that practitioners can benefit from treating cues to deception differently depending on where they appear in a dialogue.

Syntactic Indicators 23 syntactic complexity features were analyzed; all of these measures were indicators of deceptive responses. This suggests that syntactic complexity measures are useful features for automatic deception detection. Contrary to theories that deceptive language is simplistic, all complexity measures were increased in deceptive responses. It is possible that since interviewees in the CXD corpus were given time before the interview to prepare their lies, they did not experience an increase in cognitive load and therefore syntactic complexity was not reduced in deceptive speech.

Most of the findings presented in this chapter were consistent with prior work, but some contradicted previous findings. It is difficult to identify global deception indicators, since there are many important differences in experimental paradigms that affect the nature of deception. Are the lies premeditated or spontaneous? Is there incentive provided for successful deception? Are the deceptive responses constrained to a particular structure (e.g. yes/no responses) or domain (e.g. opinion about death penalty)? In what modality does the deception take place (e.g. face-to-face, text, oral)?

In this work we identified acoustic-prosodic, lexical, and syntactic characteristics of deceptive and truthful speech in the context of the CXD corpus. Interviewees responded to a set of 24 biographical questions with premeditated lies or truths, but also spontaneous

responses to follow up questions. They were provided financial incentive to lie well, and the target of the deception was the interviewer. The deception modality was audio only. This paradigm mimics a real-world scenario where an individual might be questioned about their background over the phone.

This systematic analysis of over 150 speech- and text-based features in a large-scale corpus of deceptive speech furthers our scientific understanding of deceptive language and is an important contribution of this thesis.

There are several ways to extend this work. One area that can be improved is the quality of the lexical features. Several of features (e.g. LIWC, hedge words) are identified using lexicons, and this approach often introduces noise. For example, there is ambiguity in hedge word identification, where contextual information is necessary to determine whether a word is a hedge or not. Contextual cues can be leveraged in a rule-based or machine learning classifier [Ulinski *et al.*, 2018] to improve the quality of the features.

Additionally, this analysis identifies trends across all speakers in the corpus, but there are some speakers that do not exhibit these trends. It is important to consider not only the patterns of behavior in the aggregate, but also of individuals and of sub-groups. In Part II of this thesis we analyze the same features, considering subgroups of speakers that have the same gender, native language, or personality type.

Chapter 6

Deception Classification

In Chapter 5, we demonstrated that there were significant differences between deceptive and truthful interviewee responses – in prosody, lexical content, psycholinguistic dimensions, and syntactic complexity measures. Motivated by these differences, we used acoustic-prosodic and linguistic features to train machine learning classifiers to automatically distinguish between deceptive and truthful speech. This chapter presents the results of a series of classification experiments to answer the following questions:

- What segmentation unit is best for deception classification?
- What is the best classification approach for automatic deception detection?
- Which features are useful for deception classification?

In order to shed light on the optimal segmentation size for deception classification, we compared the results of classifiers trained on four different segmentation units described in Chapter 4: IPUs, turns, question responses, and question chunks. An *inter-pausal unit (IPU)* is defined as a pause-free segment of speech from a single speaker, with a pause length threshold of 50 ms. A *turn* is defined as a maximal sequence of IPUs from a single speaker without any interlocutor speech that is not a backchannel. A *question response* is an interviewee turn that is a direct answer to an interviewer question from the list of 24 biographical questions. A *question chunk* is a set of interviewee turns that are answers to an interviewer biographical question and its related follow-up questions.

IPUs and turns have local deception annotations, while question responses and chunks have global deception annotations. Global deception refers to the veracity of a multi-utterance response to a set of questions related to a salient discourse topic. Local deception refers to the veracity of utterances that are spoken in support of a topic. In the CXD corpus, the discourse topics are the 24 biographical questions that are used during interviews, and each interviewee response to a question is globally true or false. Comparing deception classification results across the four segmentation units helps us understand the role of context in deception classification, as well as the trade-offs between global and local deception annotations.

We compared the performance of several supervised learning approaches in order to study which method performed best for deception classification. We selected four classification models that are commonly used in speech and text classification problems, described below. (In addition to these four classifiers, we also explored neural network classifiers. Those experiments are described later, in Chapter 14.)

- Random Forest (RF)

  Random Forest is an ensemble method, where multiple decision trees are generated, each trained on a random subset of features, and classification is done by majority voting. We used forests of 100 trees for our experiments.

- Logistic Regression (LR)

  Logistic Regression is a linear model for classification, which uses a logistic (sigmoid) function to model the probability of a binary dependent variable. We used L2 regularization to reduce overfitting.

- Support Vector Machine (SVM)

  A Support Vector Machine determines an optimal hyperplane to separate classes. We used an SVM with a linear kernel.

- Naive Bayes (NB)

  A Naive Bayes classifier applies Bayes' theorem which assumes independence of features. They have been shown to work well for document classification problems. We

used an implementation of Gaussian Naive Bayes, which assumes the likelihood of the features is Gaussian.

We used the scikit-learn implementation of all classification models (`http://scikit-learn.org`). The local deception labels for IPUs and turns were not balanced: 57% of IPUs were labeled as 'T' and 42% of IPUs were labeled as 'F'; 60% of turns were labeled as 'T' and 40% of turns were labeled as 'F'. On the other hand, the global labels for question responses and chunks *were* balanced, since participants were instructed to lie for exactly 12 of the 24 questions. In order to overcome the skewed distribution of local deception labels, we randomly sub-sampled the truthful class so all of the data was balanced for these experiments. This enabled easy comparison of results across segmentations. Thus, the random baseline for all four segmentations is 50% accuracy; that is, a classifier that always predicts the same class will correctly label 50% of the test samples. Another baseline that we compare our results to is human performance. Because the subjects playing the role of interviewer provided deception judgments during the interview, we can measure human performance as the average percentage of correct judgments made by interviewers, which was 56.75% This human baseline performance is for the task of deception classification of question chunks only. This is because interviewers marked their judgments after asking each question and corresponding follow up questions. We did not collect human judgments of deception for any other segmentation; thus this human baseline can only be directly compared with classification of question chunks.

For all experiments, we evaluated the models using 10-fold cross validation, with unique speakers in each fold. The speakers per fold were the same for all segmentations to ensure consistency.

We trained classifiers using acoustic-prosodic, lexical, and syntactic feature sets described in Chapter 4. We first trained classifiers on each individual feature set, and then on feature combinations. We also conducted feature ranking analysis to understand which features were most useful for deception classification. All features were z-score normalized per speaker.

6.1 Individual Feature Classification

We began by training classifiers on individual feature sets in order to assess which single feature sets were most discriminative between truthful and deceptive speech. We compared the performance of multiple classifiers trained on each feature set. We repeated these experiments on each of the four segmentation units. In the tables below, we show the classification results. The classifiers were evaluated using accuracy metric, as well as the F1-score for truth (F1-T), F1-score for deception (F1-F), and the average F1-score for truth and deception (F1-F). The "CLF" column in the tables below represents the classifier that performed best for a particular feature set. In the analysis of the results and in choosing the best classifiers, we focused on the average F1 metric which captures the balance of precision and recall for both truthful and deceptive classes, and is a robust measure of the classifier performance.

The individual feature sets that we assessed were: Praat, IS09, LIWC, LDI, Complexity, and N-gram features.

Table 6.1 shows the results for classification of IPUs.

Feature	*Acc*	*F1-T*	*F1-F*	*F1-Avg*	*CLF*
Praat	51.23	50.10	52.07	51.09	LR
IS09	52.08	51.70	52.35	52.03	LR
LIWC	**53.32**	51.98	54.58	**53.28**	LR
LDI	52.59	49.97	**54.93**	52.45	LR
Complexity	51.12	50.34	51.85	51.09	LR
N-gram	53.30	**54.00**	52.56	**53.28**	LR

Table 6.1: IPU classification with all individual feature sets.

IPU classification results ranged from 51.09 F1 (Praat, complexity) to 53.28 F1 (LIWC, n-gram). All results were better than the random baseline (50% accuracy), however they were only marginally better. Overall, the text-based features did slightly better than the speech-based features. Complexity features performed poorly. This is likely because IPU segments were short and did not have enough context to capture useful syntactic complex-

ity features. The logistic regression classifier was the best performing model for all IPU classification tasks.

Table 6.2 shows the results of turn classification experiments.

Feature	*Acc*	*F1-T*	*F1-F*	*F1-Avg*	*CLF*
Praat	52.00	56.79	45.82	51.30	LR
IS09	52.10	52.94	51.17	52.05	LR
LIWC	54.39	56.60	51.93	54.26	SVM
LDI	52.96	54.46	51.32	52.89	LR
Complexity	52.15	57.77	44.69	51.23	LR
N-gram	**55.87**	**58.42**	**52.95**	**55.69**	LR

Table 6.2: Turn classification with all individual feature sets.

Turn classification results ranged from 51.23 F1 (complexity) to 55.69 F1 (n-grams). As with IPUs, complexity features performed poorly for turn classification. Although many turns are longer than IPUs and provide more context, there are also many turns that consist of a single IPU or even a single word, which makes it difficult to capture meaningful syntactic structures. The best performing feature set was n-gram features, and these features benefited from the additional context in turns over IPUs. The logistic regression classifier was the best model for all feature sets except LIWC, which performed best with the SVM classifier.

Table 6.3 shows the results of classification experiments for question response segmentation.

Feature	*Acc*	*F1-T*	*F1-F*	*F1-Avg*	*CLF*
Praat	52.76	55.77	49.15	52.46	LR
IS09	54.84	52.89	56.59	54.74	RF
LIWC	58.86	55.82	61.50	58.66	SVM
LDI	57.75	54.68	60.40	57.54	LR
Complexity	53.97	55.66	52.06	53.86	LR
N-gram	**60.54**	**58.56**	**62.33**	**60.44**	LR

Table 6.3: Question response classification with all individual feature sets.

Results for question responses ranged from 52.46 F1 (Praat) to 60.44 F1 (n-grams). Praat features were again the worst performing feature set, and n-gram features were again the best performing feature set, achieving an F1-score about 10% better than the random baseline. LIWC features also performed strongly (58.66 F1). The logistic regression classifier performed best for all features except LIWC (SVM was best) and IS09 (RF was best).

Table 6.4 shows the classification results for question chunk segmentation.

Feature	*Acc*	*F1-T*	*F1-F*	*F1-Avg*	*CLF*
Praat	55.69	58.66	51.90	55.28	RF
IS09	56.15	54.65	57.47	56.06	RF
LIWC	59.62	58.53	60.64	59.59	SVM
LDI	58.89	58.89	58.87	58.88	LR
Complexity	57.53	57.19	57.83	57.51	SVM
N-gram	**60.96**	**59.89**	**61.95**	**60.92**	LR

Table 6.4: Question chunk classification with all individual feature sets.

Question chunk classification results ranged from 55.28 F1 (Praat) to 60.92 F1 (n-grams). For question chunks, logistic regression was no longer the preferred classifier for most feature sets. Instead, we see that acoustic-prosodic feature sets (Praat, IS09) performed best with the random forest classifier, while LIWC and complexity features performed best with the SVM classifier, and LR was the best performing classifier for LDI and n-gram features. It

is interesting that different classification algorithms performed best with different feature sets depending on the segmentation unit.

In summary, all classifiers incrementally improved as segmentation unit sizes increased, from IPUs to turns to question responses to question chunks. This was true for all feature sets. It is particularly interesting that classification of question responses consistently performed better than turn classification, since question responses are simply a subset of turns that are direct answers to biographical questions. Despite the fact that the set of question responses is only about 20% of the full set of turns, we obtained much better performance from reducing the data size. It seems that it is easier to classify turns with global deception labels than local deception labels. The remaining 80% of turns include answers to follow up questions, but they also include statements that are off-topic or perhaps do not have a clearly defined deception label. These results are consistent with the work of Enos *et al.* [2007], who found that classification of so-called "critical segments," segments that are relevant to salient deception topics, yielded better performance than classification of local deception in their full corpus.

Praat features were the lowest performing feature set for all segmentations. This is likely because these are only 15 summary statistics of acoustic-prosodic features. IS09 features are a much larger and complex acoustic-prosodic feature set, and it seems that these features better capture the prosodic differences between truthful and deceptive speech. Complexity features were highly sensitive to segmentation unit, performing barely above baseline for IPUs and turns, but achieving 57.51 F1 for question chunks. Text-based features generally performed better than acoustic-prosodic features, and standard n-grams were surprisingly the best-performing feature set, outperforming our customized deception features (LDI) and psychologically motivated features (LIWC).

We also explored additional syntactic features for the topic based segmentation - question responses and question chunks. IPUs and turns were excluded from these classification experiments because the syntactic features captured by the complexity feature set were noisy for those shorter segmentation units. The additional syntactic features that we explored are:

- POS (Part-of-speech)

- Word+POS
- PR-lex (Production rules, lexicalized)
- PR-unlex (Production rules, unlexicalized)
- G-PR-lex (Grandparent-annotated production rules, lexicalized)
- G-PR-unlex (Grandparent-annotated production rules, unlexicalized)

These features are described in detail in Chapter 4, Section 4.4.2. A list of the POS tags and their descriptions is found in Appendix B.

For classification purposes, we represented each of these feature sets as a bag of words (n-gram) model, but instead of words as tokens, we used the feature (e.g. POS tag, or production rule) as tokens. The logistic regression classifier performed best for all of these syntactic features, for both question response and question chunk segmentations.

Table 6.5 shows the classification results using these syntactic features for question response segmentation.

Feature	*Acc*	*F1-T*	*F1-F*	*F1-Avg*
POS	57.42	52.75	61.24	57.00
Word+POS	**60.53**	**58.94**	**61.98**	**60.46**
PR-Lex	59.74	57.51	61.75	59.63
PR-Unlex	56.55	52.66	59.84	56.25
GPR-Lex	59.22	56.92	61.27	59.09
GPR-Unlex	55.94	51.96	59.30	55.63

Table 6.5: Question response classification with individual syntactic feature sets.

Results for question response segmentation ranged from 55.63 F1 (GPR-Unlex) to 60.46 F1 (Word+POS). Combining word tokens with their part of speech tags was useful for deception detection. However, we note that the performance of this feature set was almost the same as using n-grams alone (60.44 F1), so it is unclear whether there is much to be gained from adding part of speech tag information. Lexicalized production rules (PR-Lex) also performed well, with an F1-score of 59.63.

Table 6.6 shows the classification results using syntactic features for question chunk segmentation.

Feature	*Acc*	*F1-T*	*F1-F*	*F1-Avg*
POS	57.27	56.22	58.25	57.24
Word+POS	**60.93**	**60.68**	**61.17**	**60.92**
PR-Lex	59.90	58.90	60.82	59.86
PR-Unlex	57.30	56.34	58.16	57.25
GPR-Lex	58.99	58.29	59.62	58.96
GPR-Unlex	56.60	55.88	57.23	56.55

Table 6.6: Question chunk classification with individual syntactic feature sets.

The results for question chunk segmentation are similar to those from the question response segmentation, ranging from 56.55 F1 (GPR-Unlex) to 60.93 F1 (Word+POS). Again, the best results come from combining words with their part of speech tags, and this yields the same performance as training a model with word-only n-grams (60.93 F1). In general, lexicalized production rules performed better than unlexicalized production rules.

These results provide insight into which classifiers and features are useful for deception detection, depending on the segmentation unit being classified. In the next section, we explore classifiers trained on combinations of features to leverage the strengths of multiple features. In addition, we explore feature ranking to understand which features contribute the most to classification.

6.2 Feature Combinations

This section presents the results of classification experiments using combinations of features explored in the previous section. We grouped features into three main feature sets:

- Acoustic-prosodic
- Lexical
- Syntactic

Acoustic-prosodic features consist of Praat and IS09 feature sets, and lexical features consist of LIWC, LDI, and n-gram feature sets. The syntactic feature set is only the complexity features for IPUs and turns, and is the combination of POS, word+POS, and production rule features for question responses and question chunks.

In this section we present the results of classifiers trained with each of these feature sets, as well as combinations of these three feature sets. To combine feature sets, we concatenated the feature vectors of each feature. We did this for each of the four segmentation units.

Because concatenating feature vectors resulted in a very large number of features, we used feature selection to reduce the feature space and eliminate features that were not helpful to classification. Feature selection was done using the SelectKBest function in scikit-learn. We used a score function which scores features using the ANOVA F-value between the class label and each feature. We used grid search to optimize k, the number of top-ranked features that were selected. The tables below show the results of these experiments, evaluated by accuracy (Acc), F1-score for the truthful class (F1-T), F1-score for the deceptive class (F1-F), and average F1-score (F1-Avg). The CLF column indicates which classifier performed best for each feature set, and the k column indicates the number of features that were selected for that feature set.

Table 6.7 shows the results using feature combinations for IPU classification.

Feature	*Acc*	*F1-T*	*F1-F*	*F1-Avg*	*CLF*	*k*
Acoustic	52.90	52.27	53.50	52.89	LR	200
Lexical	56.01	56.16	55.85	56.00	LR	3000
Syntactic	51.12	50.34	51.85	51.09	LR	all
Acoustic+Lexical	56.25	56.33	56.17	56.25	LR	3000
Acoustic+Syntactic	52.72	52.26	53.16	52.71	LR	300
Lexical+Syntactic	56.00	56.33	55.66	55.99	LR	3000
All	**56.29**	**56.37**	**56.20**	**56.29**	LR	3000

Table 6.7: IPU classification with combined feature sets.

Classification results for IPUs using feature combinations ranged from 51.09 F1 (syn-

tactic) to 56.29 F1 (all features). The syntactic feature set for IPUs consisted of only complexity features, and as we observed previously, these features do not capture useful differences between truthful and deceptive IPUs. This is likely because IPUs are too short to have meaningful syntactic complexity measures. The best performance of 56.29 F1 was obtained by combining all three sets of features and using the top 3000. However, this performance was not much better than the performance obtained using the best single feature set, lexical – 56 F1. It seems that lexical features (LIWC + LDI + n-grams) were the most useful for IPU classification. Combining these three categories of lexical features yielded better results than training with any of those feature sets individually. As shown in Table 6.1 above, the best performance for IPU classification with a single feature set was 53.28 F1 (n-grams). Consistent with our findings from single feature classification, the best classifier for IPU segmentation with combined features was logistic regression, for all feature combinations.

We show the results for turn classification with feature combinations in Table 6.8.

Feature	*Acc*	*F1-T*	*F1-F*	*F1-Avg*	*CLF*	*k*
Acoustic	52.98	53.85	52.06	52.96	LR	200
Lexical	58.03	60.19	55.60	57.90	LR	3000
Syntactic	52.15	57.77	44.69	51.23	LR	all
Acoustic+Lexical	**59.77**	**65.78**	51.21	**58.49**	NB	3000
Acoustic+Syntactic	53.03	53.65	52.40	53.02	LR	300
Lexical+Syntactic	57.86	59.83	55.67	57.75	LR	3000
All	57.86	58.53	**57.17**	57.85	LR	3000

Table 6.8: Turn classification with combined feature sets.

Classification results for turns using feature combinations ranged from 51.23 F1 (syntactic) to 58.49 F1 (acoustic+lexical). As with IPUs, the syntactic feature set for turns consisted of only complexity features, and as we observed previously, these features do not capture useful differences between truthful and deceptive turns.

The best performance of 58.49 F1 was obtained using a combination of acoustic+lexical

feature sets. The acoustic+lexical NB classifier also achieved the highest accuracy (59.77%) and the highest F1 for the truthful class (65.6 F1-T). However, as seen from the breakdown of F1 for T and F classes, the classifier has a high F1 of 65.78 for truthful turns, and a low F1 of 51.21 for truthful turns. In contrast, the LR classifier trained on all features had a slightly lower average F1 (57.86), but a more evenly balanced F1 across T and F classes. Depending on the application, one might prefer to optimize a classifier for F1 of a particular class. For example, in a high-stakes scenario where it is critical to avoid a false positive (e.g. incriminating an innocent person), we would prefer a model with a very high F1 for deception, even at the cost of F1 for the truthful class. All of our experiments were optimized for average F1 across both classes, but this objective can be modified depending on the application.

Although using acoustic+lexical features yielded the best average F1-score, training with lexical features alone yielded an F1 of 57.90, which was very close to the best performance. As with IPUs, it seems that lexical features (LIWC + LDI + n-grams) were the most useful for turn classification. Combining these three categories of lexical features yielded better results than training with any of those feature sets individually. As shown in Table 6.2 above, the best performance for turn classification with a single feature set was 55.69 F1 (n-grams).

Consistent with our findings from single feature classification, the logistic regression classifier was preferred for all feature combinations except for acoustic+lexical features combined.

We show the results for question response classification with feature combinations in Table 6.9.

Feature	*Acc*	*F1-T*	*F1-F*	*F1-Avg*	*CLF*	*k*
Acoustic	56.40	56.45	56.34	56.40	SVM	200
Lexical	64.43	64.56	64.27	64.42	SVM	3000
Syntactic	**66.05**	**66.29**	**65.80**	**66.04**	SVM	5000
Acoustic+Lexical	63.47	63.99	62.93	63.46	SVM	3000
Acoustic+Syntactic	64.31	65.25	63.32	64.28	SVM	5000
Lexical+Syntactic	65.77	66.27	65.24	65.76	SVM	5000
All	63.69	64.61	62.71	63.66	SVM	5000

Table 6.9: Question response classification with combined feature sets.

Classification results for question responses using feature combinations ranged from 56.40 F1 (acoustic) to 66.04 F1 (syntactic). Syntactic features for question responses include complexity features as well as POS, word+POS, and production rule n-gram features. This combined feature set yielded strong classification performance. The total size of the syntactic feature set without feature selection was close to 30,000, and the best performance was obtained using the top 5,000 ranked features.

This combined syntactic feature set yielded better results than training with any of those feature sets individually – as shown in Table 6.5 above, the best performance for question response classification with a single syntactic feature set was 60.46 F1 (word+POS).

Combining the syntactic features with other feature sets did not improve performance. Some of the syntactic features capture lexical content (e.g. word+POS features, lexicalized production rules), and this explains why combining syntactic with lexical features does not improve performance. It is surprising that combining acoustic features with syntactic features did not improve over syntactic features alone, since they provide a new dimension. However, training with acoustic features on their own resulted in an F1-score of 56.4, so it seems that acoustic features were not as useful in discriminating between truthful and deceptive responses.

In contrast to IPUs and turns, where logistic regression was the preferred classifier, here we found that SVMs resulted in the best performance for question response classification with feature combinations.

Finally, we show the results for question chunk classification with feature combinations in Table 6.10.

Feature	*Acc*	*F1-T*	*F1-F*	*F1-Avg*	*CLF*	*k*
Acoustic	58.10	57.48	58.69	58.09	SVM	200
Lexical	64.96	64.65	65.25	64.95	NB	3000
Syntactic	69.34	69.38	69.29	69.34	NB	5000
Acoustic+Lexical	66.31	65.93	66.66	66.30	NB	3000
Acoustic+Syntactic	69.24	69.23	69.24	69.23	NB	5000
Lexical+Syntactic	**69.81**	**69.95**	**69.66**	**69.80**	NB	5000
All	69.43	69.53	69.32	69.43	NB	5000

Table 6.10: Question chunk classification with combined feature sets.

Classification results for question chunks using feature combinations ranged from 58.09 F1 (acoustic) to 69.8 F1 (lexical+syntactic). Syntactic features for question chunks include complexity features as well as POS, word+POS, and production rule n-gram features. This feature set, combined with lexical features (LIWC, LDI, and n-grams) yielded strong classification performance. The lexical+syntactic classification results were only marginally better than the results using syntactic features alone, likely because lexical content is also captured by some of the syntactic features. Combining these features yielded better results than training with any of those feature sets individually. As shown in Table 6.6 above, the best performance for question chunk classification with a single feature set was 60.92 F1 (word+POS). Interestingly, Naive Bayes (NB) classifiers resulted in the best performance for question response classification with feature combinations. It seems that classifier selection should be made based on the segmentation unit as well as the features being classified.

In summary, acoustic-prosodic, lexical, and syntactic features are predictive of deceptive language. We achieved performance well above a random baseline of 50% accuracy for each segmentation: +6.3% for IPUs, +9.8% for turns, +16% for question responses, and +19.8 for question chunks. Because interviewers recorded their judgments for each of the 24 biographical questions, we also have a human baseline for question chunk classification

of 56.75% accuracy. The best performance for automatic question chunk classification was 69.8% accuracy, an absolute increase 13.05% and a relative increase of 23%. Thus, we can achieve "super-human" performance at deception detection using natural language processing and machine learning techniques.

We found that combining feature sets always improved performance over using individual features, and using reduced feature sets with top ranked features further improved performance. We also observed that different classification algorithms performed better for different feature sets and segmentation units, suggesting that there are many important factors to consider when modeling deception.

6.3 Feature Ranking

Having demonstrated that these acoustic-prosodic, lexical, and syntactic features are highly effective at deception classification, we were interested in analyzing which features contributed most to classification. In particular, we observed that feature selection was an important step to improve classifier performance, since the full set of features totaled to over 30,000 features.

For each of three main feature groups – acoustic-prosodic, lexical, and syntactic, we show the top features selected by feature selection. We also show the top features for the set of all features combined. Feature selection was done using the SelectKBest function in scikit-learn. We used a score function which scores features using the ANOVA F-value between the class label and each feature. Below we show the top 20 features and their F-values for each group of features. Because question chunks consistently yielded the best performance, we show the feature ranking for the question chunk segmentation.

The top 20 ranked acoustic-prosodic features are shown in Figure 6.1.

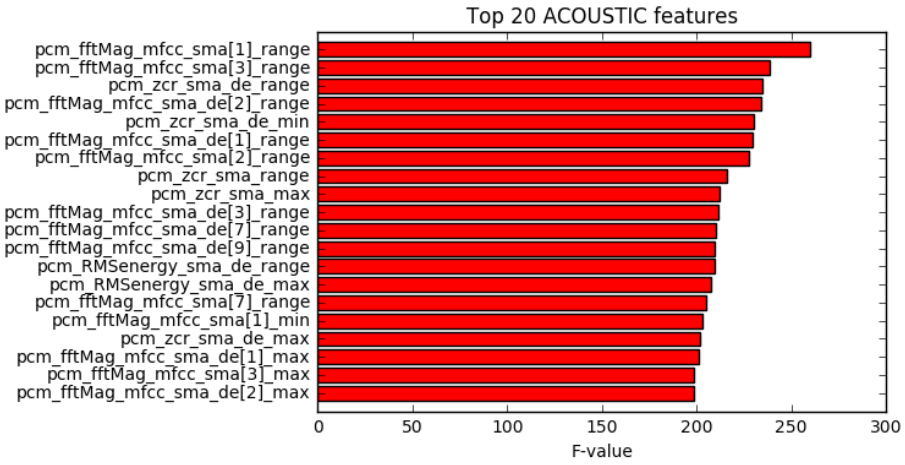

Figure 6.1: Top 20 acoustic features for deception classification, ranked by ANOVA F-values.

All 20 top acoustic features came from the IS09 feature set, and none came from the Praat feature set. 13 were MFCC features, five were functionals computed over the zero-crossing rate (ZCR) from the time signal, and two features were functionals computed over RMS energy.

Figure 6.2 shows the top 20 ranked features from the lexical feature set.

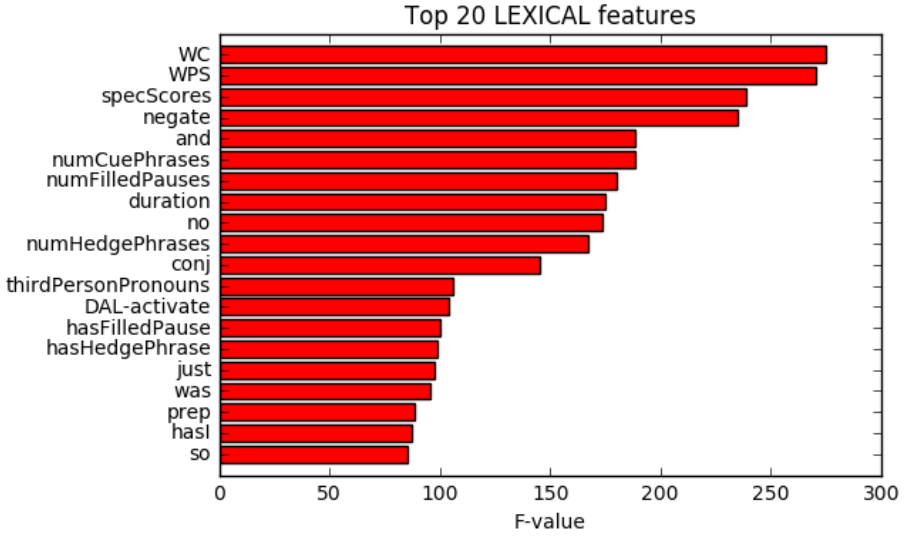

Figure 6.2: Top 20 lexical features for deception classification, ranked by ANOVA F-values.

Six of the top 20 lexical features came from the LIWC feature set, nine came from

the LDI feature set, and five came from the n-gram feature set. Several of these features were found to be significant indicators of deception or truth, as shown in Chapter 5. For example, hedge phrases and filled pauses were more frequently used in deceptive interviewee responses. Specificity scores were also increased in deceptive speech. N-gram features were not previously analyzed, and here we see that some unigrams appear in the top lexical feature set: *and, no, just, was, so.* Some of these were captured in LDI or LIWC features. Conjunctions (e.g. and) were more common in deceptive responses, as were past tense verbs (e.g. was). Thus, these ranked features are largely consistent with the findings of our prior analysis. Unlike the top acoustic-prosodic features, which were dominated by IS09 features, it seems that there was a more equal distribution of top features from all three lexical feature sets.

Figure 6.3 shows the top 20 ranked syntactic features.

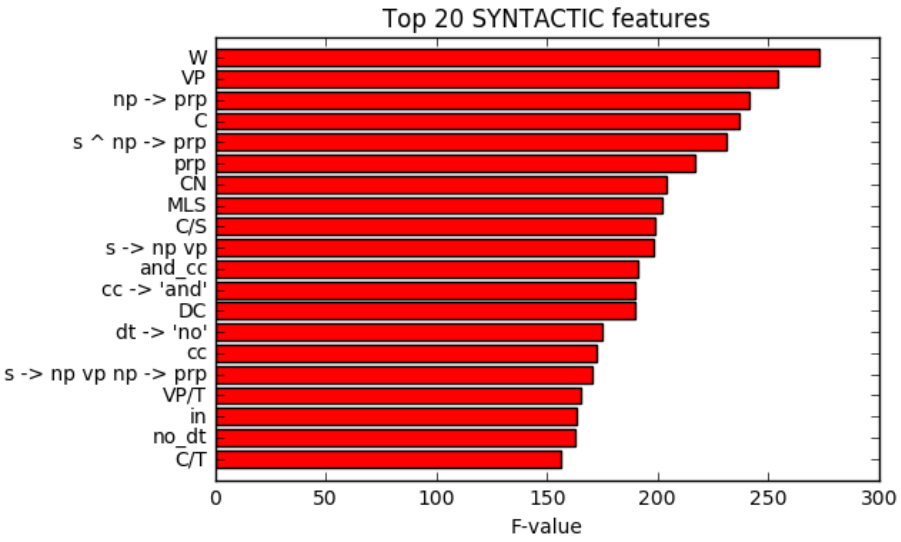

Figure 6.3: Top 20 syntactic features for deception classification, ranked by ANOVA F-values.

Nine of the top 20 syntactic features were from the complexity feature set, five from the POS and word+POS features, and six from the production rules feature sets. Some of these features capture similar cues from other feature sets. For example, lexicalized production rules "DT− >no" and "CC− >and" and word+POS features "no-dt" and "and-cc" are the

same, and they are essentially the same as the n-grams "no" and "and", which were top lexical features. Prepositions also appear in a few of the top syntactic features, and these were found to occur more frequently in deceptive responses.

Figure 6.4 shows the top 20 ranked features from the combined set of acoustic, lexical, and syntactic features.

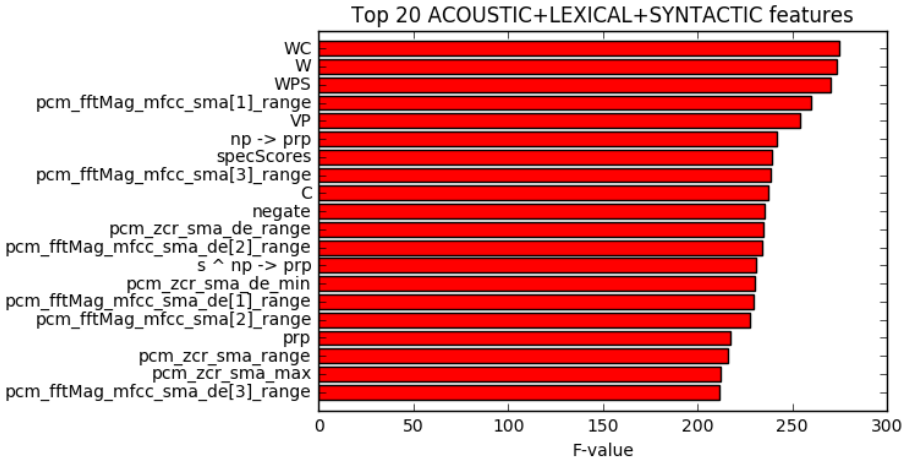

Figure 6.4: Top 20 acoustic+lexical+syntactic features for deception classification, ranked by ANOVA F-values.

From all of the feature sets, WC (word count from LIWC) and W (word count from complexity, and WPS (words per sentence) are the top three ranked features. All three of these features capture the same trend – deceptive statements had more words and words per sentence than truthful statements. We see from this figure that the top 20 features are a mix of acoustic-prosodic, lexical, and syntactic features. Feature selection was an important step in improving classification performance, and it is also helpful to examine the top selected features to understand which features were effective at distinguishing between truthful and deceptive responses. We found that the top selected features were generally consistent with our statistical analysis of cues to deception and truth.

6.4 Discussion

This chapter presented a comprehensive set of machine learning experiments to classify deceptive and truthful interviewee responses. The results of these experiments enable us to address our original questions about automatic deception classification:

What segmentation unit is best for deception classification? Our experiments were conducted using four segmentation units: IPUs, turns, question responses, and question chunks. Two of the segmentations, IPUs and turns, were labeled with *local* deception annotations using the participant keypress logs. The other two deception annotations, question responses and turns, were labeled with *global* deception annotations using the participant responses from the biographical questionnaire. The experimental results consistently showed improved performance as segmentation duration increased. IPU classification had the lowest performance, followed by turn classification, then question response classification, and the best performance was achieved using question chunk classification. This was true across all feature sets and classification methods.

Despite the fact that the shorter segmentation units had the largest amount of training instances, the results clearly indicate that the best deception classification performance is achieved using question chunks. Although there are many more instances of IPU and turn segmentations, they include ambiguous segments that do not have a clearly defined veracity label. Contextual information is often necessary for disambiguation. For example, an IPU that consists of a filled pause or laughter is not clearly truthful or deceptive. It is only in the context of the preceding and following IPUs that a veracity label can be determined. Question chunk classification has the advantage of a large amount of contextual information.

Further, comparing performance across segmentation units allowed us to evaluate the benefits of local vs. global deception annotations. The experimental results suggest that the globally annotated data was more useful than locally annotated data. Overall, it seems that data *quality* is much more important than data *quantity*.

What is the best classification approach for automatic deception detection? We compared the performance of four classification algorithms: Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), and Naive Bayes (NB). The classification results indicate that there is no single best approach for deception classification,

but that different classifiers are suited for different tasks.

Logistic regression (LR) was the best classification algorithm for most IPU and turn classification tasks. LR models the probability of a dependent variable (in this case true or deceptive) using a logistic function. It is often used as a baseline before trying other more complex models since it is efficient, interpretable, and works well off-the-shelf without parameter tuning. IPU and turn classification were the lowest performing tasks across all classifiers and feature sets, suggesting that the truthful and deceptive classes were not easily separable using the various acoustic-prosodic and linguistic features. It seems that the simple LR classification algorithm worked best for the more difficult tasks. However, it did not perform strongly – the best LR IPU classification performance was 53.28 F1 for single feature sets (n-grams) and 56.29 F1 for feature combinations (all features). The best LR turn classification performance was 55.69 F1 for single feature sets (n-grams) and 58.49 F1 for feature combinations (all features).

Logistic regression was also the best performing classification algorithm for classification of n-gram features across all segmentations. For question responses and question chunks, LR was the best classifier for all individual n-gram feature sets, including word n-grams, POS n-grams, and n-grams of various forms of syntactic production rules. N-grams are high dimensional, sparse features and highly correlated, and logistic regression can handle features with these characteristics.

After combining multiple feature sets and applying feature selection, SVM was the best performing classifier for the question response segmentation. However, Naive Bayes was the best performing classifier for the question chunk segmentation, after feature selection. NB applies Bayes' theorem with the naive assumption of independence between features. Thus, it performs poorly at text classification without feature selection, since there are highly correlated features. NB does not perform well for deception classification with acoustic features which are also highly correlated with each other. The best classifier for acoustic-prosodic features extracted from question chunks was random forest, which can handle correlated features. Despite the strong performance of NB for the question chunk segmentation, it was not the best classifier for the question response segmentation. The quality of the features, in particular the syntactic n-grams, was higher for question chunks than question responses,

since dependency parses were more accurate when there was more context available. It seems that the NB classifier benefited from these higher quality features.

Overall, there is no single classification algorithm that is "best" for deception detection. Rather, best practices for deception classification vary significantly depending on the segmentation units classified and the feature sets used. The classifier that seems most versatile across feature sets and segmentation units is logistic regression. However, the best performance was obtained by using a NB classifier trained on selected lexical and syntactic features for question chunk segmentation.

Which features are useful for deception classification? We trained classifiers using two acoustic-prosodic feature sets (Praat, IS09), three lexical feature sets (LDI, LIWC, n-gram), and one syntactic feature sets for IPUs and turns (complexity), and seven syntactic feature sets for question responses and question chunks (complexity, POS, word+POS, PR-lex, PR-unlex, G-PR-lex, and G-PR-unlex). Classifiers were trained using individual feature sets and combinations of multiple feature sets. The experimental results show that text-based features (lexical and syntactic) generally performed better than acoustic-prosodic features. The best performing single feature set across all segmentations was word n-grams, ranging from 53.28 F1 (IPU) to 60.92 F1 (question chunk).

Although there were some trends across all segmentations, some features performed very differently depending on the segmentation unit. For example, complexity features on their own performed barely above chance for IPUs and turns (51 F1), but were more useful for question chunks (57.51 F1). The complexity features were computed from dependency parses, which were much more accurate for longer segmentations.

Feature combinations yielded the best deception classification performance, and feature selection was an important pre-processing step to reduce the feature dimensions. The optimal feature combinations varied across segmentations: all features for IPUs (56.29 F1), acoustic+lexical for turns (58.49 F1), syntactic features for question responses (66.04 F1), and lexical+syntactic for question chunks (69.8 F1).

In addition to comparing classification results for different feature sets, we conducted feature ranking analysis to understand which specific features were most discriminative between truthful and deceptive speech. This analysis is complementary to the feature anal-

ysis reported previously in Chapter 5. This provides insight into which acoustic-prosodic, lexical, and syntactic features were most useful for classification.

In summary, we compared the performance of 4 classification algorithms, trained with various acoustic-prosodic, lexical, and syntactic feature sets. We trained classifiers for 4 segmentation units: IPUs, turns, question responses, and question chunks, and reported optimal classifiers and feature sets for the different segmentation units. We explored classification with various feature combinations and used feature selection to improve performance. Finally, we presented feature ranking results to understand which features contributed most to classification. Our best classifier was a Naive Bayes classifier trained with a combination of lexical and syntactic features extracted from question chunks, and achieved an accuracy of 69.8% – well above human performance of 56.75% accuracy. In addition to the contribution of these strong performing deception classifiers, this work contributes to our scientific understanding of deceptive language, and provides useful insights for future experiments with automatic language-based deception detection.

Chapter 7

Error Analysis

Having trained automatic deception detection classifiers, in this chapter we take a closer look at the classification performance and compare it with human performance. We aimed to answer the following questions:

- Does classifier and human deception detection performance vary across speakers? I.e. Are there speakers that are "easier" or "harder" for humans or machines to detect when they are lying?
- Are classifier judgments of deception related to human judgments?
- Are there particular groups of *speakers* (e.g. by gender, native language, or personality) that are easier or harder to classify? Does this differ for human and machine judges?
- Are there particular groups of *segments* that are easier or harder to classify? Does this differ for human and machine judges?

In order to answer these questions, we analyzed the predictions made by the best classification model: a Naive Bayes classifier trained on a combination of 5000 selected lexical and syntactic features, for the question chunk segmentation. This model achieved an F1-score of 69.8.

7.1 Deception Detection per Speaker

Does classifier and human deception detection performance vary across speakers? To answer this question, we grouped the question chunk segments by speaker, and computed the average F1-score of the classifier predictions for each speaker individually. We also computed the average F1-score of human predictions for each speaker. There are 340 unique speakers in the corpus, and each speaker had a maximum of 24 question chunk segments (some speakers had slightly fewer segments, because of missing features or missing data from the interview). Table 7.1 shows summary statistics of average F1-score per speaker, computed from both classifier and human predictions. Q_1, Q_2, and Q_3 represent the first, second, and third quartiles.

Judge	*Mean*	*Std*	*Min*	Q_1	Q_2	Q_3	*Max*
CLF	68.48	11.17	31.25	60.79	69.42	77.22	91.66
Human	55.33	12.50	22.57	46.67	54.17	64.26	100.00

Table 7.1: Summary statistics for speaker-level F1-scores, for both classifier and human judgments. (CLF=classifier)

Figure 7.1 shows the distribution of average F1-scores per speaker, comparing F1-scores from human and classifier predictions. F1-scores from human judges are shown in red, classifier F1-scores are shown in blue, and the overlapping region is purple.

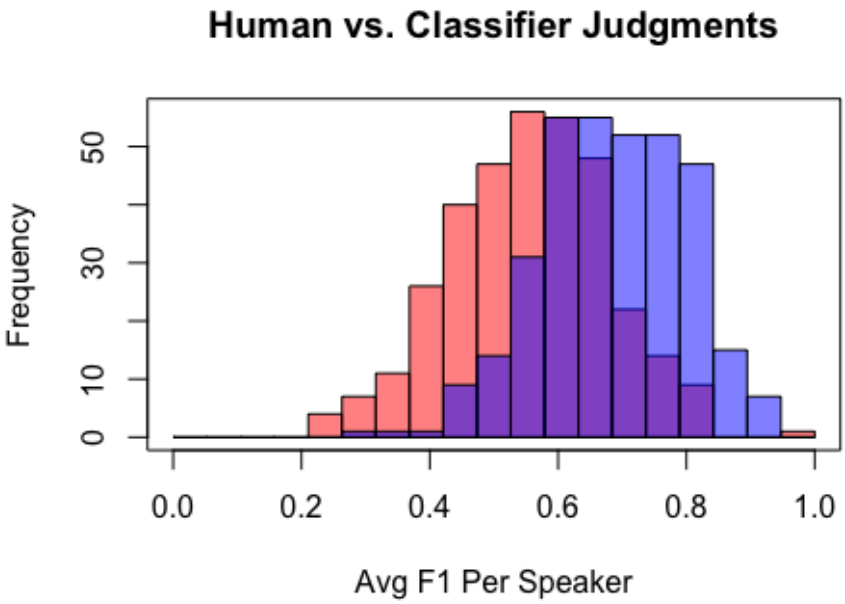

Figure 7.1: Histogram of speaker-level F1-scores for classifier and human judgments. F1-scores from human judges are shown in red, classifier F1-scores are shown in blue, and the overlapping region is purple.

As shown from both the summary statistics and the histogram, there is a wide range of both human and classifier F1-scores across speakers. The standard deviation is high for both humans and machines. In addition, classifier performance is consistently higher than human judge performance.

7.2 Human vs. Machine Performance

Are classifier judgments of deception related to human judgments? We explored this question at both the segment level and the speaker level. At the segment level, we aimed to discover whether the deception classifier and the human judges made similar deception judgments across all interviewee responses (i.e. were *segments* that were easy/hard for humans to judge also easy/hard for the classifier to judge?). And at the speaker level, we aimed to discover whether the classifier and humans performed similarly for each speaker

(i.e. were *speakers* that were easy/hard for humans to judge also easy/hard for the classifier to judge?).

At the segment level, we performed two analyses to answer this question. First, we used the Pearson's chi-squared test to compare classifier judgments and human judgments across all question responses, and found that they were not independent ($\chi^2(1, N = 7772) = 94.65, p \approx 0$). That is, human judgments and classifier judgments were strongly related. Next, we examined whether classifier and human *performance* at deception detection was related at the segment level. To do this, we computed a "correct" or "incorrect" label for each segment, for both human predictions and classifier predictions. We then compared classifier performance and human performance using the chi-squared test, and found that these were also strongly related ($\chi^2(1, N = 7772) = 32.17, p \approx 0$). Thus, classifier and human judgments of deception, as well as classifier and human performance at deception detection, were strongly related at the segment level. This is true despite the fact that the human judgments were made by many different interviewers.

At the speaker level, we computed the average F1 of the classifier for all segments per speaker CLF_{F1}, as well as the average F1 of the human interviewer for all segments per speaker $human_{F1}$. These measures represent how difficult or easy it was for a classifier or human judge to detect deception for a particular speaker. We used three analysis methods to study the relationship between human and classifier performance at the speaker level.

We computed the Pearson's correlation between $human_{F1}$ and CLF_{F1}, and found that there was no correlation between these measures ($r(340) = -0.02, p = 0.73$). Thus, although human and machine deception judgments were correlated at the segment level, human and classifier performance were not correlated at the speaker level.

We also computed the Kendall's rank correlation coefficient τ between $human_{F1}$ and CLF_{F1}. This statistic measures the ordinal association, i.e. the relationship between rankings, between two variables. We observed no significant correlation between speaker ranking by $human_{F1}$ and speaker ranking by CLF_{F1} ($\tau(340) = 0.01, p = 0.76$).

Finally, we partitioned each speaker into one of three bins – high, average, or low – using quantiles of the F1-score to partition the speakers. Speakers who were classified with an F1-score in the top 75% were placed in the "high" bin, representing speakers that were

classified with high performance, while speakers who were classified with an F1-score in the bottom 25% were placed in the "low" bin. The remaining speakers were placed in the "average" bin. We computed the bins using $human_{F1}$ and CLF_{F1} and used Pearson's Chi-squared test to evaluate whether the distributions were independent. The results show no significant interaction between the two distributions ($\chi^2(4, N = 340) = 0.85, p = 0.93$).

In summary, all three analysis methods suggest that classifier performance per speaker is not related to human performance. The average F1 values per speaker were not correlated between human and speaker judgments, and the relative rankings of the speakers by human and machine judgments were also not correlated. Finally, the distribution of high, med, and low F1-scores from human and classifier predictions were independent. Thus, although human judgments of deception were strongly related to classifier judgments across all segments, humans and the classifier did not perform similarly at the speaker level.

7.3 Classifier and Human Performance Across Speaker Traits

Are there particular groups of *speakers* (e.g. by gender, native language, or personality) that are easier or harder to classify? Does this differ for human and machine judges? To answer these questions, we compared the CLF_{F1} and $human_{F1}$ measures across groups of speakers. Paired t-tests comparing both CLF_{F1} and $human_{F1}$ between male and female interviewees, and between native Chinese and native English speakers, yielded no significant differences. It seems that the classifier and human judges did not perform significantly better or worse for speakers of a particular gender or native language.

We used an ANOVA to compare $human_{F1}$ and CLF_{F1} across personality factors, using the high, average, and low personality bins described in Chapter 12. We observed a significant effect of the personality factor of Conscientiousness on CLF_F1 ($F(2, 337) = 3.99, p = 0.02$). A Tukey post-hoc test revealed that the difference came from the comparison of CLF_{F1} between speakers that were in the Low and Average Conscientiousness bins. The mean CLF_{F1} for speakers in the Average bin was 66.7, while the mean for speakers in the Low bin was 70.32 ($p = 0.014$). Thus, the classifier performed significantly better for

speakers who were low on the Conscientiousness scale.

We did not observe any significant effect for personality when comparing $human_{F1}$. In summary, interviewee gender, native language, and personality did not generally have a significant effect on classifier or human performance. However, we did observe an effect of Conscientiousness on classifier performance, but not on human judge performance.

7.4 Classifier and Human Performance Across Segment Characteristics

Are there particular groups of *segments* that are easier or harder to classify? Does this differ for human and machine judges?

We considered segment characteristics of duration, the biographical question that was used to elicit the response segment, and the biographical question type.

Duration

Duration is end_time - start_time of a segment, and for question chunks this represents the duration of a dialogue about a particular biographical question (since it includes interviewer follow up questions).

We computed paired t-tests to compare segment duration between predicted true and predicted false segments, from both classifier and human predictions. The results showed that duration was significantly different between segments that were believed to be true or judged to be false, by humans and the classifier. The mean duration of segments that were judged as false by humans was 37.07, while the mean duration of segments that were judged as true by humans was 31.39 ($t(7772) = 6.19, p \approx 0$). The difference was even more stark for classifier judgments: the mean duration of segments that were judged as false by the classifier was 48.78, while the mean duration of segments that were judged as true by the classifier was 18.95 ($t(7772) = 35.48, p \approx 0$). Both humans and the machine learning classifier tended to judge longer question chunk segments as deceptive, and shorter segments as truthful. This is an intuitive result for human judges, since the interviewers decided how many follow up questions to ask for each question.

It is likely that when an interviewer was skeptical about the interviewee's initial response,

they would ask more follow up questions. To test this hypothesis, we compared the number of follow up questions in question chunks that were believed or not believed by humans and the classifier. The results confirmed this hypothesis: the mean number of follow up questions in chunks that were believed by interviewers was 5.06, while the mean number of follow up questions in question chunks judged as deceptive was 5.74 ($t(7772) = 5.19, p \approx 0$). The same was true for classifier judgments: the mean number of follow up questions in chunks that were judged as true by the classifier was 3.69, while the mean number of follow up questions in question chunks judged as deceptive was 7.02 ($t(7772) = 27.48, p \approx 0$). These findings are intuitive, since the number of follow up questions per question chunk was strongly correlated with the chunk duration ($r(7772) = 0.83, p \approx 0$).

We also computed paired t-tests to analyze the difference between classifier and human *performance* across segment duration and number of follow up question per segment. We found that segment duration was not significantly different between segments that were correctly or incorrectly judged by humans ($t(7772) = 0.43, p = 0.67$), nor was the number of follow up questions ($t(7772) = 0.09, p = 0.93$). Similarly, we found no difference in duration between segments that were correctly or incorrectly judged by the classifier ($t(7772) = 1.09, p = 0.28$) and no difference in number of follow up questions ($t(7772) = 1.31, p = 0.18$).

Although the classifier and human *judgments* of deception were strongly related to the duration and number of follow up questions per segment, human and classifier *performance* were not.

Biographical Question

Next, we examined whether human and classifier judgments of deception varied across responses to different questions. There were 24 biographical questions used in each interview session (see Appendix A.4 for a sample questionnaire). We aimed to discover whether certain questions were easier or more difficult to classify, for humans or for the classifier. We computed CLF_{F1} and $human_{F1}$ aggregated by question number, from 1-24. That is, we computed the F1 score individually for all segments that were responses to a particular biographical question. Figure 7.2 shows the classifier F1 per question, and Figure 7.3 shows the human F1 per question.

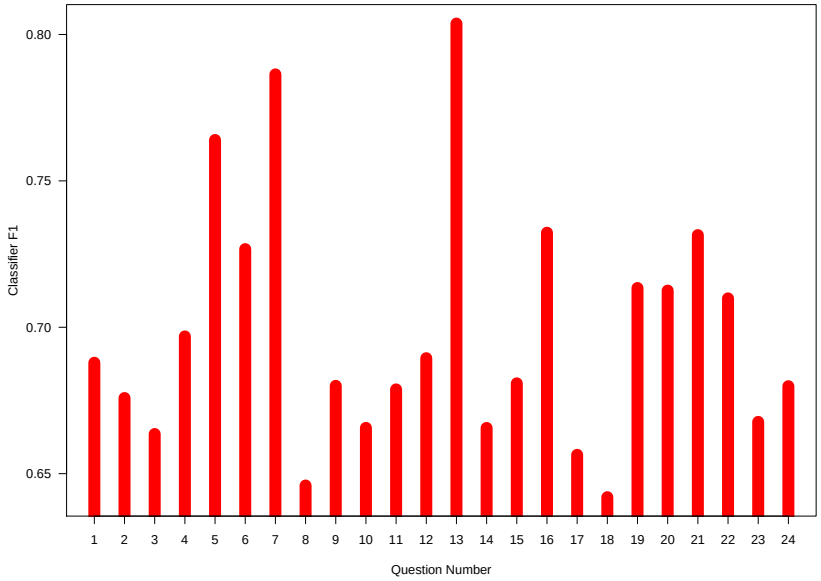

Figure 7.2: Classifier F1 per Question Number

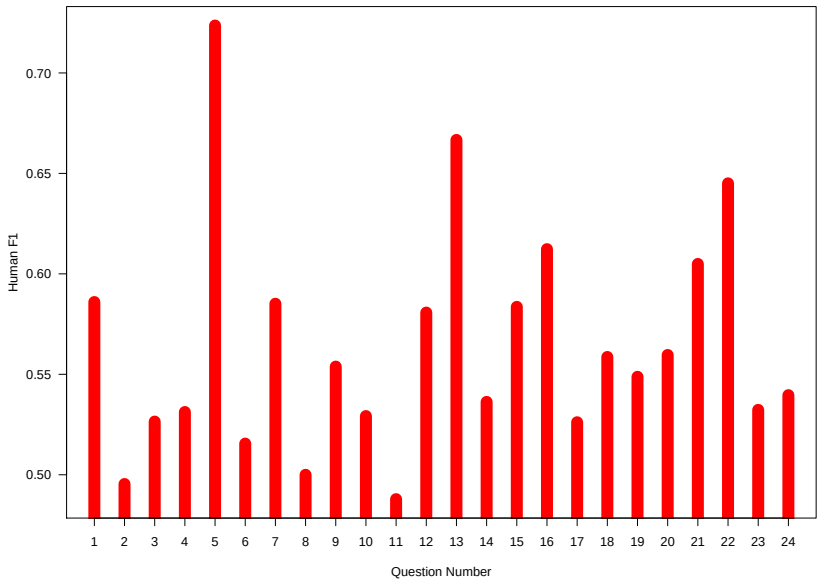

Figure 7.3: Human F1 per Question Number

The figures show that there was significant variation in both classifier and human perfor-

mance across questions. Further, there are some similar trends in performance by question for the classifier and humans. We computed the Pearson's correlation between CLF_{F1} and $human_{F1}$ and found that they were significantly correlated ($r(24) = 0.69, p = 0.0002$). Thus, questions that were easier or more difficult for human judges were similarly easier or more difficult for the classifier. For example, both humans and the classifier performed strongly for question number 5, "Have your parents divorced?" Human F1 was 72.36 for this question (+16 from the mean of 56.37 across all questions). Classifier F1 was 76.39 for this question (+6.58 from the mean of 69.82 across all questions). Both humans and the classifier also performed strongly for question 13, "Have you ever gotten into trouble with the police?" and question 16, "What is the most you have ever spent on a pair of shoes?"

To understand why these questions might have been easier for both humans and the classifier to identify deception, we examined the interviewee responses to these questions. In addition to writing a lie in response to half of the questions, subjects were also instructed to record the truthful response to each question, enabling us to study the ground truth for each question. We identified important trends in the data that help explain this pattern. For question 5, we found that about 80% of *truthful* responses stated that their parents had not divorced, while only 20% of truthful responses stated that their parents had divorced. Thus, human interviewers could have used this intuition about general divorce trends to judge responses to question 5. Similarly, the classifier could learn this distribution for this question. If the classifier always judged "yes" responses to this question as false, and "no" responses to this question as true, it would obtain a very high accuracy of about 80%.

We found a very similar trend for question 13: about 80% of interviewees truthfully stated that they had never gotten into trouble with the police, while only about 20% truthfully stated that they did. Again, humans using their intuition about the probability that a college student had previously had trouble with the police, or the classifier learning the distribution, could result in high performance.

For question 13, participants were asked to provide the amount of money that they spent on their most expensive pair of shoes. We analyzed the amount reported for truthful and for deceptive responses, and found that the median amount for truthful responses was $150, while the median amount for deceptive responses was $350. Human interviewers could use

their domain knowledge of shoe prices and college student behavior, and a machine learning classifier could learn this pattern from training data.

On the other hand, question 8, "Have you ever stayed overnight in a hospital as a patient?" was particularly difficult for humans and the classifier. Humans achieved an F1 of 50 for this question (F1 of 50, -6.37 from mean), and classifier performance was 64.59 F1 (-5.22 from mean). We analyzed the truthful responses to question 8 and found that 61% of interviewees had never stayed overnight in the hospital as a patient, while 39% had. This difference is much smaller than the 80%-20% difference for the "easy" questions, and helps explain why it would be harder to detect deception for this question.

There were also some differences between human and classifier performance per question. Question 6, "Have you ever broken a bone?" had one of the lowest F1 scores for human judges (51.55, -4.82 from mean), but was above average for the classifier (72.61, +2.85 from mean). An analysis of the truthful responses to question 6 showed that 75% of interviewees had never broken a bone, while 25% had. It is possible that human intuition about this question was incorrect, while the classifier was able to learn this distribution and perform better.

Question Type

Having established that there is variation in human and classifier performance across responses to different questions, in this next analysis we studied whether there are also differences across categories of questions. One way of categorizing the questions is by differentiating between questions that require a yes-no response (e.g. "Have you ever tweeted?") and those that are open-ended (e.g. "What is your major?"). 13 of the 24 questions are yes-no questions (questions 5-11,13,18,19,22,23,24), and 11 are open-ended questions (questions 1-4,12,14,15,16,17,20,21). Another way to categorize the questions is into sensitive and non-sensitive questions. We followed the criteria in Tourangeau and Yan [2007] to define sensitive questions in our corpus. These questions are related to money (number 16), parental or romantic relationships (5,14,15), mortality (23), socially undesirable behaviors or experiences (12,13,24). In total, there are 8 sensitive questions and 16 non-sensitive questions.

To explore whether human and classifier judgments were different across question type,

we used the Pearson's chi-squared test to compare judgments between sensitive/non-sensitive questions, and between yes-no/open-ended questions. We found that there was no effect of question type on human judgments of deception (sensitive/non-sensitive:$\chi^2(1, N = 7772) = 2.07, p = 0.15$; yes-no/open-ended: $\chi^2(1, N = 7772) = 0.19, p = 0.66$). That is, human judgments of deception were not significantly different across question type. We also compared human *performance* at deception detection across pair type by assigning a "correct" or "incorrect" label to each segment, indicating whether the human interviewer had correctly or incorrectly judged that segment as true or false. We found a significant effect for sensitive/non-sensitive questions type ($\chi^2(1, N = 7772) = 17.18, p \approx 0$). To examine this effect, we computed $human_{f1}$ across all sensitive questions and all non-sensitive questions, and found that it was higher for sensitive questions (59.67 F1) than for non-sensitive questions (54.70 F1). No effect for yes-no/open-ended question type was observed ($\chi^2(1, N = 7772) = 0.67, p = 0.41$).

Next, we repeated this analysis for classifier judgments of deception. In contrast to our findings for human judgments, we found a strong effect of question type on classifier judgments of deception, for both sensitive/non-sensitive questions ($\chi^2(1, N = 7772) = 39.757, p \approx 0$) and yes-no/open-ended questions ($\chi^2(1, N = 7772) = 114.5, p \approx 0$). The classifier was more likely to predict that a segment was true if it was in response to a non-sensitive question, and also if it was in response to a yes-no question.

However, also in contrast to our findings for human judgments, we observed no effect of question type on classifier *performance* (sensitive/non-sensitive:$\chi^2(1, N = 7772) = 2.64, p = 0.1$; yes-no/open-ended: $\chi^2(1, N = 7772) = 1.84, p = 0.18$).

It seems that question type had different effects on human and classifier judgments of deception. Human judgments were not different across question type, but human performance was higher for sensitive questions. On the other hand, classifier judgments were different across question type (more "trusting" of responses to non-sensitive questions and open-ended questions), but classifier performance did not vary across question type. This suggests that the classifier was influenced in some way by the different patterns of responses to different question types, but this did not affect the classifier performance. Humans were better at detecting deception in responses to sensitive questions, and perhaps this finding

can be useful for practitioners – it is possible that asking sensitive questions rather than neutral questions can aid in deception detection.

7.5 Discussion

This chapter took a close look at the deception judgments made by the best performing deception classifier, to understand the strengths and weaknesses of the model compared with human judges. This work addressed the following questions:

Does classifier and human deception detection performance vary across speakers? That is, are there some speakers that are easier or more difficult for humans or machines to detect when they are lying? We computed the average F1-score per speaker of classifier and human judgments, and found that there was a wide range of both human and classifier judgments across speakers. Human performance ranged from 22.57 to 100 F1, with a median of 54.17, and classifier performance ranged from 31.25 to 91.66 F1, with a median of 69.42. Classifier performance was consistently superior to human judge performance at deception detection, but they both had significant variation across speakers.

Are classifier judgments of deception related to human judgments? With this analysis, we aimed to discover whether the deception classifier and human judges performed better/worse for the same speakers, or for the same interviewee responses. Using multiple statistical analyses, we found that classifier performance was strongly correlated with human performance per segment, but not per speaker. That is, there were particular kinds of interviewee responses (across all speakers) that were easy or difficult for both humans and classifiers to judge. However, humans and classifiers did not perform similarly for particular speakers.

Are there particular groups of speakers or segments that are easier or harder to classify, for humans or machines? We studied classifier and human judgments across speaker gender, native language, and personality traits, to understand whether classifier or human judgments of deception vary depending on these speaker characteristics. The analysis showed no significant effect of gender or native language on human or machine judgments of deception. There was also no effect of interviewee personality on human judgments of

deception. However, there was a significant effect of Conscientiousness on classifier judgments of deception – the classifier performed significantly better for speakers who scored low on Conscientiousness. These individuals are characterized as careless, inefficient, and not dependable. It seems that automatic deception detection is easier for speakers who have these characteristics.

We also analyzed classifier and human judgments across groups of interviewee responses, considering segment characteristics of duration, the biographical question that elicited the response, and the biographical question type (open-ended vs. yes-no, sensitive vs. non-sensitive). Segment duration was a significant factor in human and classifier judgments – longer segments (in duration, and in number of turns per chunk) were judged as more deceptive. However, human and classifier performance was not affected by duration. That is, there was no significant difference in duration between segments that were correctly or incorrectly classified (by humans or the classifier). The biographical question that was used to elicit an interviewee response played an important role in human and classifier judgments. Performance at deception detection varied greatly across questions, and classifier and human performance across questions were strongly correlated. We identified specific questions that were easier/harder for both humans and machines, likely because the distribution of truthful answers for some questions was skewed. Question type also played a role in deception judgments. Humans performed better at deception detection for sensitive questions, and the classifier was more likely to predict that a segment was true if it was in response to a yes-no question or a non-sensitive question.

This chapter highlights the importance of carefully analyzing classifier predictions to understand the factors that affect those predictions. These experiments show that the classifier acts similarly to human judges in some ways, and very differently in other ways. It is important to note that we analyzed a single deception classifier and compared it with aggregated judgments of multiple human interviewers. One classifier was used to classify all interviewee responses in the corpus, whereas the human judge was different for each session. This makes it difficult to draw conclusions from the analysis – each utterance in the corpus was labeled by a single interviewer, and there are many factors that affect each interviewer's judgments. In future work, we plan to use crowd-sourcing to collect multiple

judgments of deception for each interviewee segments, which will allow for a more complete analysis of human deception judgments compared with the classifier.

We found that there were substantial differences in classifier performance across responses to different biographical questions. It seems that using domain knowledge can be very useful for both human and machine deception detection. However, there are trade-offs involved in leveraging domain-specific information for deception classification. If the goal is to develop a general purpose deception classifier that can detect deception independent of the domain, then using domain-specific information should be avoided. A possible way to achieve this is to train a classifier using multiple data sources from different domains. On the other hand, if optimal deception detection performance for a particular domain is the objective, domain knowledge can be leveraged to achieve this goal. For example, in our classification experiments the classifier was blind to the questions that were asked to elicit the interviewee responses. It only had access to features from the interviewee in isolation. Based on these results, it is likely that giving the classifier the question number as a feature would be useful and further improve performance of question chunk classification.

Chapter 8

Entrainment in Deceptive Dialogue

In this chapter we present an analysis of entrainment in deceptive dialogues. Entrainment is the phenomenon of interlocutors becoming similar to each other in dialogue. It has been found to occur in multiple dimensions of spoken language, including acoustic-prosodic [Levitan *et al.*, 2012], linguistic style [Danescu-Niculescu-Mizil *et al.*, 2011], and syntactic structure [Reitter and Moore, 2006]. Importantly, entrainment has been associated with positive conversation outcomes, such as likability [Chartrand and Bargh, 1999], naturalness, and task success [Nenkova *et al.*, 2008]. Prior studies of entrainment have examined (apparently) truthful dialogues, mostly goal-oriented. For example, [Levitan *et al.*, 2012] studied acoustic-prosodic entrainment in a corpus of spontaneous dialogue between partners playing collaborative computer games. Lee *et al.* [2010] measured acoustic-prosodic entrainment in dialogues between married couples discussing problems in their relationship.

In this work, we studied entrainment in deceptive dialogue. Deceptive dialogue is fundamentally different from truthful dialogue in terms of conversational goals. Interpersonal Deception Theory (IDT) [Buller and Burgoon, 1996] models deception as an interactive process between a deceiver and his conversational partner, where both interlocutors make strategic adjustments during their communication. The goal of the deceiver is to convince his partner that his lies are in fact true. Because of this important difference between truthful and deceptive speech, we were interested in examining the relationship between dialogue coordination and deception. The closest previous work to ours is that of Yu *et al.* [2015], which examined nonverbal entrainment (e.g. synchrony of facial expressions and head move-

ments) in deceptive and truthful dialogue, and found that synchrony features were useful for automatic discrimination of deception from truth. In another relevant study, Hancock *et al.* [2007a] identified correlations between linguistic category usage of deceivers and their partners, and observed greater correlations during deceptive than truthful speech.

This work focuses on entrainment in acoustic-prosodic and lexical features. Entrainment in this features has not been previously studied in deceptive dialogues. We aimed to answer the following questions:

1. Do interlocutors entrain in acoustic-prosodic and lexical dimensions in deceptive dialogues?

2. Is entrainment related to deception outcomes? (a) Is entrainment correlated with the ability to deceive or detect deception? (b) Is there a difference in entrainment behavior between truthful and deceptive speech?

The CXD corpus is particularly useful for a study of entrainment. Most deception corpora contain speech from the deceiver alone, while this corpus consists of the dialogue between the interviewer and deceptive interviewee, allowing us to study entrainment. In addition, each interview consists of half truthful and half deceptive responses, enabling a within-speaker comparison of entrainment in truthful and deceptive speech. The corpus also includes both global and local annotations of deception, as well as interviewer global (i.e. question-level) deception judgments. Thus, we can analyze entrainment with respect to global and local deception labels, and also consider the relationship between interviewer perception of deception and entrainment.

Some of this work was published in Levitan *et al.* [2018b], and was done in collaboration with my co-author Jessica Xiang.

8.1 Method

We examined entrainment in eight acoustic-prosodic features that are commonly studied in speech research: intensity mean, intensity max, pitch mean, pitch max, jitter, shimmer, noise-to-harmonics ratio (NHR), and speaking rate. All acoustic features were extracted

using Praat, and z-score normalized by gender (z = (x-μ)/σ; x = value, μ = gender mean, σ = gender standard deviation). In addition to acoustic-prosodic features, we studied entrainment in four lexical features: 100 most frequent words, 25 most frequent words, *hedge words/phrases*, and *cue phrases*. Entrainment in the use of the most frequent words in a dialogue or corpus has been studied by Nenkova *et al.* [2008] and shown to be predictive of dialogue naturalness and correlated with task success. Hedge words and phrases are used by speakers to express distance or lack of commitment to what they are saying (e.g. "I think," "sort of"), and are a novel domain for entrainment analysis. Cue phrases are linguistic expressions that function as explicit indicators of discourse structure, and have also not been previously studied in the context of entrainment. We used lists of hedge words and affirmative cue lexicons that are found in Appendix C.

There are many ways to quantify entrainment behavior. In this work we followed the methods proposed in Levitan [2014], and differentiated between global and local entrainment. Global entrainment is the phenomenon where a speaker is similar to her partner over the course of a conversation, for a particular feature. This is measured using feature means over the dialogue. Local entrainment refers to a dynamic alignment that occurs within a conversation, regardless of the similarity across the entire conversation. This is measured by looking at similarity at every point in the dialogue. We studied acoustic-prosodic entrainment at both global and local levels, but only examined lexical entrainment at the global level, where there is enough lexical content to compute meaningful lexical entrainment measures.

8.1.1 Local Entrainment Measures

For all local measures of entrainment, features were extracted at the IPU level. We identified the starting IPU of each interviewer and interviewee turn (excluding the first turn of each session) and these formed the set of target IPUs. For each target IPU, IPU_t, we identified the corresponding partner IPU, IPU_p, which was defined as the ending IPU of the speaker's partner's preceding turn (excluding overlapping IPUs).

Local Proximity We calculated *partner difference* and *other difference* for each IPU_t, letting IPU_i be the ending IPU of a random speaker that was not the partner of the speaker

of IPU_t.

$$\text{partner difference} = -|IPU_t - IPU_p| \tag{8.1}$$

$$\text{other difference} = -\frac{\sum_{i=1}^{1000} |IPU_t - IPU_i|}{1000} \tag{8.2}$$

Evidence for local proximity was determined using a paired t-test between partner difference and other (non-partner) difference. If the partner difference was significantly smaller than the difference between the random non-partner, we define that as evidence of local entrainment.

Local Convergence We computed local convergence, the tendency of partners to become more locally similar to each other over time, as the Pearsons correlation coefficient between time and the absolute difference between each target IPU and its corresponding partner IPU.

Local Synchrony We computed local synchrony, the relative alignment of features of conversational partners, as the Pearsons correlation coefficient between each target IPU and its corresponding partner IPU. We repeated each correlation (for local convergence and synchrony) ten times with randomly ordered data to verify that significant results were not just a product of the size of our corpus; we consider a result valid if at least nine of the ten random permutations fail to exhibit significant correlation.

8.1.2 Global Entrainment Measures

For all global measures of entrainment, features were extracted at the IPU level and then averaged over each session. For both speakers in each session, we let S_{avg} equal the mean of all IPU values for the speaker and P_{avg} equal the mean of all IPU values for the speaker's partner. O_{avg} was the average of all IPU values for every speaker in the corpus with the same role (i.e. interviewer or interviewee) as the partner but who was **not** the partner. We calculated **partner difference** as the negated difference between S_{avg} and P_{avg} and **other** difference as the negated difference between S_{avg} and O_{avg}.

Global Proximity Evidence for global proximity was determined using a paired t-test between partner difference and other difference. If the partner difference was significantly

smaller than the difference with other speakers for a particular feature, we considered that to be evidence of global proximity.

Global Convergence Evidence for global convergence was determined using two approaches. The first approach was a paired t-test to compare average partner difference during the first five minutes and last five minutes of each session. The second approach was similar, except that partner differences in the first half of each session was compared with the second half.

All tests for significance correct for family-wise Type I error by controlling the false discovery rate (FDR) at $\alpha = 0.05$. The k^{th} smallest p value is considered significant if it is less than $\frac{k*\alpha}{n}$.

In all the tables in this chapter, we use E to indicate that a feature was entrained on, and D to indicate that a feature was **dis**entrained on (e.g. was significantly more similar to random other speakers that to partner). We consider a result to approach significance if its uncorrected p value is $<= 0.05$ and indicate this with parentheses (e.g. "(E)") in the tables.

8.2 Local Entrainment Results

Feature	*t*	*p*	*Sig.*
Pitch Max	-3.12	0.002	D
Pitch Mean	4.87	1.14E-06	E
Intensity Max	12.82	1.36E-37	E
Intensity Mean	10.67	1.38E-26	E
Speaking Rate	6.04	1.51E-09	E
Jitter	3.95	7.87E-05	E
Shimmer	2.48	0.013	E
NHR	2.75	0.006	E

Table 8.1: T-tests for local proximity: partner vs. non-partner differences.

As shown in Table 8.1, there was evidence of local proximity for all acoustic features except for max pitch. Voice quality features of shimmer and NHR had slightly weaker evidence of entrainment than pitch, intensity, and speaking rate. Adjacent partner turns were not significantly more similar to each other in max pitch than to non-adjacent turns, and in fact were more similar to the max pitch of non-adjacent turns. This is likely because of the interview format of the dialogue, where interviewers asked questions (which were often uttered with a final rising pitch) and interviewees responded with declarative statements (typically using falling pitch).

Feature	*r*	*p*	*Sig.*
Pitch Max	0.003	0.51	
Pitch Mean	-0.006	0.12	
Intensity Max	0.02	6.68E-09	E
Intensity Mean	0.04	3.42E-21	E
Speaking Rate	-0.01	0.004	D
Jitter	-0.01	0.01	D
Shimmer	0.0005	0.91	
NHR	0.01	3.24E-05	E

Table 8.2: Correlation results for local convergence analysis.

As shown in Table 8.2, we observed local convergence for max intensity, mean intensity, and NHR and divergence for speaking rate and jitter. There was no evidence of local convergence for max and mean pitch or shimmer. Again, the lack of entrainment on pitch features is likely due to the question/answer interview format of the dialogue. As with local proximity entrainment, voice quality features were less commonly entrained on.

Feature	*r*	*p*	*Sig.*
Pitch Max	0.02	2.26E-08	E
Pitch Mean	0.03	2.79E-11	E
Intensity Max	0.15	0	E
Intensity Mean	0.16	0	E
Speaking Rate	0.08	3.30E-82	E
Jitter	0.05	8.15E-29	E
Shimmer	0.03	4.64E-11	E
NHR	0.05	5.96E-35	E

Table 8.3: Correlation results for local synchrony analysis.

Table 8.3 shows evidence of local synchrony for all features. Unlike local proximity and local convergence, there was evidence of synchrony for both max and mean pitch. Thus, it seems that in this question-answer dialogue format, speakers did not entrain on pitch by *value*, rather, they entrained *relatively* on pitch, adjusting pitch to a corresponding level within their own range.

All of the correlation coefficients were weak for convergence and synchrony (the highest was .16 for synchrony on mean intensity), indicating a lack of strong trends across all speaker pairs. To better understand the variation across speakers, we analyzed local convergence and behavior for each pair of speakers. For *local convergence*, 51% of pairs converged for at least one feature, and 49% did not converge for any feature. Of the pairs that did converge for at least one feature, 44% only converged positively, 49% only diverged, and 7% converged for some features and diverged for other features. For *synchrony*, 52% of pairs synchronized for at least one feature, while 48% did not exhibit significant synchrony for any feature. Of the pairs that did synchronize for at least one feature, 73% only had positive synchrony, 19% only had negative synchrony, and 8% exhibited positive synchrony for some features and negative synchrony for others.

Although there was evidence of only positive synchrony across all speakers, when we analyzed this by speaker pairs, we observed evidence of both positive and negative synchrony. There was also evidence of both positive and negative convergence for each feature. Neg-

ative convergence, or divergence indicates that speakers adjusted their speech to become *less* similar over time. Negative synchrony indicates *complementary* entrainment, where speakers adjust their speech away from their partners speech at each turn. This may be viewed as "completing" the previous turn.

Feature	*Convergence*		*Synchrony*	
	% Total	%Pos	% Total	% Pos
Max Pitch	14	50	11	56
Mean Pitch	20	33	16	65
Max Intensity	26	47	33	87
Mean Intensity	27	53	32	89
Speaking Rate	13	41	19	91
Jitter	14	57	15	81
Shimmer	10	36	12	66
NHR	12	54	11	68

Table 8.4: Session-level local convergence and synchrony.

Table 8.4 shows the percentage of pairs with significant convergence and synchrony for each feature, considering only pairs that converged or synchronized for at least one feature. It also shows the proportion of positive and negative convergence/synchrony. The feature which partners converged most on was mean intensity, with 27% of pairs exhibiting convergence behavior. The split between positive and negative correlations for mean intensity was roughly balanced, with 53% converging on mean intensity. For some features, it was more common to converge than to diverge (e.g. jitter), while for other features it was more common to diverge (e.g. pitch mean). Max and mean intensity were by far the most commonly synchronized feature, while synchrony for max pitch was the least common. For all features, there was a much greater proportion of positive synchrony than negative synchrony. These findings highlight the lack of strong convergence and synchrony trends across speakers. It seems that speakers were adjusting to their partners' behavior, but in very different ways.

8.2.1 Deception Analysis

Having established the presence and characteristics of local entrainment in dialogue containing deceptive speech, we were interested in exploring the differences in entrainment between deceptive and truthful speech. We computed local proximity entrainment measures for each pair of speaker turns that represented a question and its (immediate) answer from the list of 24 biographical questions asked in the interviews. Question/answer pairs were identified using the question identification approach described in Maredia *et al.* [2017]. Each interviewee answer was labeled as true or false using the biographical questionnaire response sheet prepared by each subject, which was annotated with true and false labels. In addition, each interviewee response was labeled with an interviewer judgment label, indicating whether the interviewer believed that the response was true or false. This resulted in 7260 question/answer pairs. Using this data, we examined the following research questions:

Is there a difference in entrainment behavior between truthful and deceptive speech? Paired t-tests between local proximity measures of truthful and deceptive interviewee responses showed significantly more entrainment on max intensity in deceptive speech than truthful speech ($t(7244) = 3.08; p = 0.002$). In addition, there was significantly more entrainment on jitter in deceptive speech than truthful speech ($t(7226) = 2.66; p = 0.008$). This suggests that acoustic-prosodic entrainment measures, and particularly local proximity of intensity max and jitter, can be useful indicators of deception.

Is there a difference in entrainment behavior between speech that is trusted or not trusted? We repeated the previous analysis, this time comparing entrainment measures between interviewee responses that were *perceived* as truthful and those perceived as deceptive by interviewers, regardless of whether they were in reality truthful or deceptive. Paired t-tests between local proximity measures of trusted and not trusted interviewee responses showed significantly more entrainment on mean intensity in speech judged to be deceptive than in speech judged to be truthful ($t(7222) = 2.45; p = 0.014$). This suggests that entrainment on mean intensity is indicative of an exchange where one speaker does not trust the other, regardless of whether the interlocutor is in fact telling the truth.

Is there a difference in entrainment behavior between successful and unsuccessful lies? In this final analysis, we considered deceptive responses only, and compared en-

trainment measures of lies that were successful (i.e. perceived as truthful by the interviewer) and unsuccessful (i.e. correctly perceived as deceptive by the interviewer). Paired t-tests between successful and unsuccessful deceptive interviewee responses showed no significant differences in entrainment measures for any acoustic-prosodic features. This suggests that interviewees and interviewers were not significantly more coordinated under a successful or unsuccessful deception condition. Despite the fact that there were differences in entrainment behavior between truthful and deceptive speech, it seems that interviewers were not able to perceive these differences and to use them to discriminate between truth and deception. This is consistent with findings that humans in general are very poor at deception detection. In their analysis of over 200 studies of over 24,000 human judges of deception, Bond Jr and DePaulo [2006] reported that detection accuracy is close to 54% on average for judgments of trust and deception. Because of this difficulty in human perception, it is possible that entrainment measures as an indicator of deception will be more useful to a machine learning approach to automatic deception detection than to a human practitioner.

8.3 Global Entrainment Results

Feature	*t*	*p*	*Sig.*
High Frequency 100	0.33	0.74	
High Frequency 25	2.56	0.01	E
Hedge	2.82	0.005	E
Cue	0.18	0.9	
Pitch Max	2.1	0.04	E
Pitch Mean	0.89	0.37	
Intensity Max	3.94	8.53E-05	E
Intensity Mean	4.26	2.17E-05	E
Speaking Rate	3.98	7.30E-05	E
Jitter	3.2	0.001	E
Shimmer	3.44	0.0006	E
NHR	2.31	0.02	E

Table 8.5: T-test results for global proximity: partner vs. non-partner differences.

As shown in Table 8.5, there was evidence of global proximity for all features except the 100 most frequent words, cue words, and mean pitch. There was stronger evidence of entrainment for our novel dimension, hedge words, than for high frequency words, suggesting that this is a useful dimension to use for entrainment analysis. On the other hand, we found no evidence for entrainment for our other novel entrainment dimension, cue words. In addition, high frequency 25 words were entrained on, while high frequency 100 words were not. Perhaps this is because the larger group contained many words pertaining to the interview questions that were used in all dialogues.

Feature	*Beg. vs. End*			*1st vs. 2nd Half*		
	t	*p*	*Sig.*	*t*	*p*	*Sig.*
High Frequency 100	1.99	0.05	(E)	1.72	0.09	
High Frequency 25	2.05	0.04	(E)	1.9	0.06	
Hedge	1.29	0.2		0.53	0.6	
Cue	1.18	0.24		1.32	0.19	
Pitch Max	-0.56	0.58		-0.62	0.54	
Pitch Mean	0.14	0.89		-0.21	0.83	
Intensity Max	0.02	0.99		-0.14	0.89	
Intensity Mean	-0.49	0.63		-0.2	0.84	
Speaking Rate	1.04	0.3		1.26	0.21	
Jitter	0.37	0.71		0.32	0.75	
Shimmer	1.58	0.12		0.87	0.38	
NHR	0.92	0.36		0.42	0.68	

Table 8.6: T-test results for 2 measures of global convergence. "Beg. vs. End" compares first 5 and last 5 min, and "1st vs. 2nd Half" compares features from the first half and second half of each dialogue.

As shown in Table 8.6, we did not find evidence of global convergence using either metric - comparing the first 5 and last 5 minutes ("Beg. vs. End") and comparing the first and second halves of each dialogue ("1st vs. 2nd Half"). We observed a trend approaching significance for "Beg. vs. End": people were less similar in both high frequency entrainment measures in the last 5 min. than the first 5 min. Despite significant evidence of convergence at the local level, we found almost no evidence for global convergence, supporting the view that global and local entrainment are independent phenomena.

8.3.1 Deception Analysis

To further examine the relationship between entrainment and deceptive vs. truthful speech, we computed correlations between partners' global proximity entrainment and the follow-

ing global deception metrics: *Interviewee percent answers believed*: the number of the interviewee's answers that their interviewer thought were true out of a total of 24 answers; *Interviewee percent lies believed*: the number of the interviewees lies that their interviewer thought were true out of the total number of lies the interviewee told; *Interviewer percent guesses correct*: the number of the interviewer's guesses that were correct out of 24 total guesses; and *Interviewer percent lies correctly identified*: the total number of the interviewee's lies that the interviewer guessed correctly out of the total number of lies the interviewee told. The results showed that there was significant correlation between entrainment on high frequency 25 and interviewer percent guesses correct (i.e. interviewer ability to judge deception) ($r = 0.13; p = 0.016$). This indicates that it was easier for interviewers to detect deception in dialogues where the interlocutors entrained lexically. However, there was no relationship between any of the other features and any of these metrics.

8.4 Discussion

In this chapter we presented a study of entrainment in deceptive interview dialogues. This work contributes to our scientific understanding of entrainment as well as deception, two critical components of human communication. Our results show strong evidence of entrainment in deceptive speech, in many acoustic-prosodic and lexical dimensions, at both global and local levels. We identified significant variation in local convergence and synchrony behavior. In our ongoing work, we are exploring the relationship between individual traits, such as gender and native language of both interlocutors, and the nature of convergence and synchrony behavior. It will be interesting to identify clusters of speakers with shared characteristics that exhibit local convergence and synchrony in similar ways. We also identified differences in local entrainment on max intensity and jitter in deceptive and truthful speech, as well differences in local entrainment on mean intensity in trusted and mistrusted speech. These findings have implications for automatic deception detection systems, and for entraining dialogue systems that aim to elicit user trust. Future work can extend these experiments by exploring entrainment as a feature for deception classification. Another area for future work is to examine entrainment in deceptive and truthful dialogue between

human and machine interlocutors. It will be very interesting to explore similarities and differences between entrainment and trust in human-human interaction and human-computer interaction.

Chapter 9

Conclusions and Future Work

Part I of this thesis provides a comprehensive framework for deceptive speech research. Previous research on deception has been limited to small corpora, often with few features, and some studies have used rule-based classification methods. We created a large-scale corpus of deceptive speech, extracted and analyzed a large number of acoustic-prosodic, lexical, and syntactic feature sets, trained statistical machine learning classifiers to automatically identify deceptive speech, and compared human and classifier judgments of deception.

We developed an experimental paradigm for collecting dialogues of cross-cultural deceptive and truthful speech. This paradigm was designed to mitigate some drawbacks of data collected in a laboratory setting: it allows subjects to choose their own lies so they are more genuine, and it provides financial motivation for interviewers and interviewees, tailored to each role. Using this framework, we collected a large-scale corpus of within-subject deceptive and truthful speech, totaling over 122 hours. The previous largest corpus contained about seven hours of subject speech [Enos, 2009]. Our corpus enabled studies of deceptive speech on a scale that was not previously possible. The CXD corpus is a significant contribution of this thesis, and will hopefully be used by others to further the advancement of deceptive speech research.

The systematic analysis of over 150 speech- and text-based features in a large-scale corpus of deceptive speech revealed many significant differences between truthful and deceptive responses. Several of our findings were consistent with previous studies of deceptive language. Some of the features that we examined had not been previously examined in de-

ceptive speech, and were new indicators of deception. And some of our findings contradicted previous observations about deception. The range of results highlights the importance of understanding cues to deception in the context of the data in which they were observed and the underlying goal of the deceivers. We studied cues to deception and truth in two segmentation units: question responses and question chunks. While most cues were consistent across both segmentations, differences between the two suggest that some cues should be treated differently depending on where they appear in a dialogue. This work furthers our scientific understanding of deceptive language and is an important contribution of this thesis.

We focused here on identifying cues to deceptive and truthful speech. However, the poor performance of human judges at deception detection in this corpus and other corpora suggests that perception of deception is distinct from the production of deception. In our ongoing work we are studying cues to perception of deception, or trust, using interviewer judgments of deception as trust labels.

We conducted a series of classification experiments to automatically identify deceptive speech using a variety of acoustic-prosodic and linguistic features. We compared performance across multiple classification algorithms and feature combinations, using four units of analysis for training and evaluation: IPUs, turns, question responses, and question chunks. We reported optimal classifiers and feature sets for each of the different segmentation units, as well as feature ranking results to understand which features contributed most to classification. Our best classifier was a Naive Bayes classifier trained with a combination of lexical and syntactic features extracted from question chunks, and achieved an accuracy of about 70% – well above human performance of 56.75% accuracy. In addition to the contribution of these strong performing deception classifiers, this work contributes to our scientific understanding of deceptive language, and provides useful insights for future experiments with automatic language-based deception detection.

We analyzed the predictions made by the best performing deception classifier, and compared them with the judgments made by human interviewers. The analysis showed that human and classifier judgments were correlated at the segment level but not at the speaker level. We further analyzed interviewee response segments to understand which segments

were easier or more difficult for human judges and for the classifier, and identified segment characteristics that affected judgments. Our findings have implications for practitioners and for training deception classifiers. For example, human judges performed better at detecting deception in response to sensitive questions, suggesting that sensitive questions should be used in interviewing and interrogation. Our analysis showed that classifiers and human judges performed better at detecting deception in response to certain biographical questions, where domain knowledge could be leveraged, suggesting that this is a useful approach for improving domain-specific deception detection. This analysis also highlights the importance of carefully examining the data being classified, which can reveal potential biases. Future work should explore evaluate classifiers trained on the CXD corpus on corpora in other domains, to explore the implications for cross-domain generalization. In future development of deception corpora, these biases should be considered when designing experimental paradigms. It is difficult to draw conclusions from the analysis of human judgments, since each interviewer judged a single interviewee, so there are many confounding factors. Future work can extend this analysis by collecting additional judgments of deception for the corpus from multiple judges.

Our study of entrainment in deceptive speech contributes further insight into the nature of deceptive dialogues. We show that entrainment occurs on global and local levels in deceptive speech, and in acoustic-prosodic and lexical dimensions. We introduced two novel features for entrainment analysis: hedge words and cue phrases. We also highlight differences in entrainment behavior between truthful and deceptive dialogues. Exploring the use of entrainment features, such as proximity measures for acoustic-prosodic and lexical features, is a useful direction for future work. Our analysis of entrainment showed substantial variation in local convergence and synchrony behavior. This work can be extended by studying factors that affect these differences, such as gender, native language, and personality type.

Part I has focused on identifying trends in deceptive speech across all speakers in the corpus, and training classifiers using features that capture those trends to automatically identify deceptive speech. Although there are patterns of deceptive speech that are apparent across all speakers, there are some speakers that do not exhibit those trends. In Part II

of this thesis, we present analyses of individual differences in deceptive speech, considering subgroups of speakers that have the same gender, native language, or personality type. We introduce methods to leverage these differences with the goal of improving automatic deception detection.

Part II

Individual Differences in Deceptive Behavior

Chapter 10

Motivation and Research Goals

In Part I of this thesis we established that there are acoustic-prosodic and linguistic differences between truthful and deceptive speech. We also showed that machine learning classifiers can distinguish between truthful and deceptive speech significantly better than human judges. All feature analysis and classification were performed without considering individual variation in cues to deception. In Part II of this thesis, we present our work on individual differences in spoken deception. The overarching goal of this chapter is twofold: we aim to identify differences in gender, native language, and personality in how people produce and perceive deception, and we aim to leverage these differences to improve deception classification.

Most previous work on deception detection has aimed to identify cues to deception across all speakers. The underlying assumption is that there exist universal indicators of deception. In this work we question that assumption and hypothesize that different groups of speakers produce deception in different ways. People from different backgrounds and cultures, with different genders and personality traits, produce speech in different ways, and we can often identify a speaker's traits using speech processing and machine learning methods. If different speakers produce speech in different ways, it would not be surprising if they produce deception in different ways. And if they do, it is important to identify these differences, and to leverage them in automatic deception classification methods.

Some previous studies of deception have observed individual differences in how people lie. For example, Hirschberg *et al.* [2005] studied deception in American English speech, and

observed differences in the production of deception across speakers. While some subjects raised their pitch when lying, others lowered it significantly; some tended to laugh when deceiving, while others laughed more while telling the truth. However, there have not been significant efforts to empirically study these differences, and understand the factors that affect these differences. An impediment to empirical studies of individual differences in deceptive speech has been the lack of corpora with annotations of individual traits. The Columbia X-Cultural (CXD) corpus was designed and collected with the goal of studying individual differences in deception, and it includes annotations of three categories of speaker traits that might play a role in variation in deception production: (1) gender (2) culture and (3) personality.

There are many differences in speech production between male and female speakers, and there has been extensive research to identify these differences in acoustic-prosodic and linguistic features [Argamon *et al.*, 2003; Shafran *et al.*, 2003]. Gender affects language production significantly, motivating our interest in exploring how gender affects the production of deceptive speech.

Most work on deception has focused on native speakers of Standard American English. The few studies of deception in other languages have largely focused on within-culture deception. We were interested in studying deception both within and across cultures, and identifying differences in cues to deception across culture. Culture is difficult to quantify, and for this study we use native language as a proxy for culture. We studied deception in conversations between native speakers of American English and native speakers of Mandarin Chinese, all speaking in English. We examined similarities and differences in their production of deception, as well as their perception of deception. Although there are people from various cultures included in these groups, this work is a first step in increasing our understanding of the relationship between culture and deceptive behavior. The methods used in this work can be extended and applied to study other cultural groups in the future.

Enos *et al.* [2006] discovered that human judges' accuracy in judging deception could be predicted from their scores on simple personality tests – the NEO-FFI Five Factor Personality Inventory [Costa and McCrae, 1989]. Based on this, it is possible that such personality tests provide useful information in predicting individual differences in deceptive

behavior of speakers rather than judges of deception. We were interested in exploring the role of personality in deception production. We also used the NEO-FFI personality inventory to measure the big five personality traits: Neuroticism, Extroversion, Openness to Experience, Agreeableness, and Conscientiousness.

In this section we aim to answer the following main research questions:

Are there group-specific differences in acoustic-prosodic and linguistic features between truthful and deceptive interviewee responses? We use statistical methods to compare the features of deceptive and truthful speech across gender, native language, and personality. Chapter 12 presents the results of this analysis.

Can we leverage differences across groups to improve deception classification performance? In Chapter 13, we explore methods of incorporating gender, native language, and personality scores in classification models, and compare these results to those presented in Chapter 6. Chapter 14 explores speaker-dependent neural network models for deception classification.

Can we automatically identify gender, native language, and personality of speakers using acoustic-prosodic and linguistic features? In Chapter 15, we report the results of several experiments aimed at identifying speaker traits from short samples of speech, with the goal of using these automatically learned labels to improve deception detection.

Chapter 11

Related Work

This work is motivated by previous studies of deception that observed individual variation in deceptive speech. In their work on automatic deception detection in American English speech, Hirschberg *et al.* [2005] noticed differences in spoken cues to deception across subjects. For example, some subjects raised their pitch when lying, while others lowered it significantly. Some tended to laugh more when lying, others laughed more while telling the truth. We are interested in exploring individual characteristics that might play a role in these differences in deceptive behavior, such as gender, personality, and culture.

11.1 Deception and Gender

Of the possible speaker traits to explore in relation to deception, gender has been the most studied. This is likely due to the ease of obtaining gender labels. For speech corpora, speaker gender is easily identifiable, and gender is standard demographic information that is collected in most studies. Despite several studies of deceptive behavior across gender, the relationship between gender and deceptive behavior is not well understood, with several inconsistent findings in the literature.

Some studies have examined ability to lie and detect lies across gender. DePaulo *et al.* [1985] studied the effects of speaker gender and listener gender in an experiment where subjects described their deceptive and truthful opinions on controversial topics. They found that lies told by female participants were more easily detected than lies told by males. They

also found that same-gender deception was easier than cross-gender deception, i.e. lies were more likely to be detected when the judge was the opposite gender of the deceiver. However, a study of deception using an interactive social media game platform found no significant difference between the success of male and female deceivers [Ho and Hollister, 2013]. Tilley *et al.* [2005] studied gender differences in computer-mediated deception. Subjects participated in a fake job interview session via an online communication platform and provided judgments about whether they thought their partner was honest or dishonest. In contrast to [DePaulo *et al.*, 1985], they did not observe a significant effect of deceiver gender on deception success. However, they did find that female subjects were significantly better at detecting deception than male subjects. They suggest that females are more attentive to details and therefore notice more deception cues than males.

Other studies examined gender differences in motivations for lying and also in the choice of what to lie about. A study by Dreber and Johannesson [2008] examined propensity to deceive using an economics game, and found that men were significantly more likely than women to lie in order to gain a monetary benefit. With the increase in computer-mediated communication, researchers have studied deception in online communications, where it is easy to lie about one's identity. In particular, dating profiles have been a popular area to study gender differences in deception. Hancock *et al.* [2007b] measured the height and weight of subjects and verified their ages by checking their ID (e.g. driver's license), and then compared these verified attributes with those reported on their online dating profiles. They found that 81% of subjects lied about at least one variable and observed these gender differences: Men were more likely to overestimate height, while women underestimated weight. According to the self-presentational model of deception [DePaulo *et al.*, 2003], people lie to portray themselves in a beneficial way to others, and so it is intuitive that there are gender differences in what is considered positive self-presentations. In another study of deception in the context of online dating, Guadagno *et al.* [2012] examined how the expectation of meeting impacted deception. Participants in their study were randomly assigned a dating condition: face-to-face, email, no meeting, or a control group (no relation to dating), and filled out self-reported personality and attractiveness questionnaires. They found that male participants (but not females) exaggerated their positive characteristics

when there was an expectation of dating, and did so most dramatically when the expected modality was email communication.

Recently, people have studied differences in machine learning performance at deception detection across gender. Similar to studying differences in human deception detection ability depending on the gender of the deceiver, here the goal is to see whether there is a difference in automatic deception detection performance for male or female speakers. Abouelenien *et al.* [2017] explored gender-based differences in multimodal deception detection. They reported differences in classification performance between males and females, and observed different patterns in deceptive behavior across gender. A trend in their findings was that deception appeared to be more easily detectable in females. Similarly, Pérez-Rosas and Mihalcea [2014b] trained classifiers to detect deception in short texts, and found that automatic deception detection performance was slightly higher for female deceivers than for male deceivers.

11.2 Deception and Culture

There has been little study of the effects of culture on deceptive behavior. Different cultures often have differing social norms, behaviors, values, and communication patterns, and therefore studying cultural effects on deceptive behavior is an interesting and potentially useful area of research to explore. It is difficult to measure the effects in a reliable way, and studies have used different methods to try to study cultural differences in deceptive behavior.

Some studies used surveys to address whether beliefs about deceptive behavior are universal or culture-specific. Al-Simadi [2000b] asked Jordanians to complete a 20-item questionnaire that assessed their beliefs about behaviors associated with deception, and compared their responses to reported beliefs by Americans. They found many culture-specific beliefs about deception (for example, only Jordanians rated face color as a cue to deception), while only three of the 20 behaviors (e.g. hesitations) were believed to be deceptive in both cultures. A more comprehensive study by Team [2006] recruited participants from 75 countries, speaking 43 languages, to provide beliefs about deceptive behaviors. They

found agreement on some cues to deception across many cultures, and even identified a cue to deception that was shared by all 75 countries – averted eye gaze. They also observed several culture-specific perceptions of deception. These studies contribute toward our understanding of how culture affects the perception of deception; however, it does not address the problem of identifying cultural effects on the production of deception. This is arguably a more useful area of research for deception detection, since perceptions about deception have not been found to correlate with reliable cues to deception Zuckerman *et al.* [1981].

Another method for examining cross-cultural deception cues has been to test whether people detect deception within and across cultures from visual and/or audio information. Bond *et al.* [1990] videotaped Jordanians and Americans telling lies and truths in their native language, and then other Jordanians and Americans were asked to watch the videotapes, without sound, and judge whether the subject was lying or telling the truth. Raters were able to reliably detect lies within their culture but not across cultures, indicating that visual cues may be culture-specific. Follow up studies found that people can detect deception across cultures and languages if visual and audio information are available [Bond Jr and Atoum, 2000]. Al-Simadi [2000a] found that Jordanians and Malaysians were able to detect lies across cultures when they had audio and visual information and were able to judge lies within cultures when they had only audio or audio and visual information. These findings suggest that there may be differences in ability to perceive deception accurately depending upon modality of information and that these abilities may differ when one is judging deception in one's own culture or in another.

It is often difficult to distinguish cultural and language effects. Many studies draw conclusions from comparing people speaking two different languages. A study by Cheng and Broadhurst [2005] found that Cantonese-English bilinguals were more often judged as being deceitful when they spoke in their second language than when they spoke in their first language, regardless of whether they were telling the truth or lying. This indicates that second-language speakers may be perceived differently than native-language speakers. To our knowledge, no study has directly compared the effect of culture on deception behaviors when speakers from different cultures are all speaking one language.

In recent years, studies have been able to research cultural effects on deception using

automated methods. Pérez-Rosas and Mihalcea [2014a] studied the effects of culture on deception by collecting a corpus of deceptive and non-deceptive texts written by people from three countries: United States (American English), India (Indian English) and Mexico (Spanish). They compared the performance of within-culture deception classification with cross-cultural classification (i.e. training on data from one country and testing on another country) and found that within-culture classification was significantly higher performing than cross-cultural classification. They also compared the cues to deception across cultures and observed some common trends across cultures and some culture-specific cues to deception. This study focused on written deception, and the dataset consists of texts without an explicit receiver, so there is no study of the effects of both the deceiver's culture and the target's culture on deceptive behavior.

In summary, there has been little study of cross-cultural deception, compared to the amount of work on deception within cultures. There seem to be some universal perceptions of deception, and many culture-specific beliefs about deception, but these perceptions and beliefs do not always correlate with deceptive behaviors. Studies have indicated that people can successfully detect deception across cultures and languages. However, little work has been done to identify reliable indicators of deception in different cultures. Finally, there has not been work directly comparing the effect of culture on deception behaviors when speakers from different cultures are all speaking one language. Our work aims to fill in these gaps in the literature on culture and deception. We carefully study verbal cues in deception across cultures in an objective manner, examining precisely defined and automatically extractable features, and using statistical and machine learning techniques. Our corpus consists of dialogues between native speakers of English and Chinese, all speaking in English to avoid identifying language-specific rather than culture-specific cues to deception.

11.3 Deception and Personality

Personality is another speaker trait that has been minimally studied in the context of deception. Our personality influences how we communicate, and personality traits can be automatically identified from speech or writing samples [Mairesse *et al.*, 2007; Moham-

madi and Vinciarelli, 2012]. Therefore, it is interesting and potentially useful for deception detection to study the effect of personality on deceptive behavior.

A meta-analysis by Aamodt and Custer [2006] analyzed over 200 studies of deception, examining the relationship between individual differences and accuracy at deception detection. They analyzed the relationship between personality and deception detection, and found that people who had a "self-monitoring" personality were better at deception detection. Self-monitoring measures the degree to which people can regulate their behaviors to accommodate social situations. Due to the lack of studies that examined personality and deception the meta-analysis had no other findings related to personality traits.

Enos *et al.* [2006] studied personality differences in human ability to detection deception. They used the NEO-FFI Costa and McCrae [1989] to measure personality and found that the accuracy of humans at deception detection could be predicted from their NEO-FFI personality scores. They found strong correlations between success in judgments and high scores on Agreeableness and Openness to experience. Judges scoring high on Neuroticism were more reluctant to rate statements as lies. These findings suggest that there are personality differences in ability to judge deception, but it is unclear whether there are personality differences in ability to deceive or in deceptive behavior.

Bradley and Janisse [1981] did study personality differences in ability to deceive. They used a mock-crime paradigm for the deception task, and used the Eysenck Personality Inventory (EPI) Eysenck and Eysenck [1975] to assess extroversion. They found that people with high extroversion scores were more likely to be judged correctly as lying or telling the truth than people who were more introverted. They hypothesized that since introverts have general anxiety in social situations, they would exhibit anxiety both when lying and telling the truth, whereas extroverts would only display anxiety when lying. In contrast, Siegman and Reynolds [1983] observed that extroverted individuals were better at lying than introverts. They hypothesized that extroverts are more socially comfortable and better able to monitor and control their cues to deception than introverts. These conflicting findings demonstrate that the relationship between personality and deception is not well-understood, and further research is needed to draw conclusions.

Vrij and Graham [1997] also examined personality differences in ability to deceive. They

studied the personality traits of public self-consciousness (PSC) and ability to control behavior (ACB). They found that people with high levels of PSC had fewer hand movements when deceiving, while people with low PSC levels had increased hand movements when lying. Further, they found that people who scored high for PSC and ACB had the fewest hand movements when lying, while those who scored low for both traits had the most hand movements when lying. These results were consistent with their hypothesis that people who have high public self-consciousness and are skilled in controlling their behavior would make fewer hand movements when lying than telling the truth. In a follow up study, Vrij and Graham [1997] used this information to train people in deception detection. They told participants that people with these personality traits were found to have decreased hand movements when deceiving, and asked them to assess the personality of potential deceivers as well as the veracity of their statements. The group that was trained with this information performed better at deception detection than the untrained control group. This suggests that knowledge about the effect of personality on deceptive behavior can be useful for deception detection.

11.4 Conclusions

Based on previous work, it seems that gender, culture, and personality can affect the production and perception of deception. However, there has been little work done do identify specific differences across these traits. Do speakers with different traits exhibit different cues to deception? If so, can we identify them, and leverage the differences to improve automatic deception detection? This section addresses these important questions.

Chapter 12

Individual Differences in Cues to Deception

Previous studies of deceptive language have focused on general inferences about deception; this work carefully examines patterns of deception that differ across gender, native language, and personality types. This analysis is critical for understanding how a speaker's individual traits can affect their production of deception. The CXD corpus allowed us to analyze deceptive speech on a scale that had not been previously possible, and in Chapter 5 we reported differences in features of deceptive and truthful speech. Having identified many differences between deceptive and truthful language across all speakers, we were interested in analyzing differences in deceptive language across different groups of speakers. In this chapter we explore differences in cues to deception across groups of speakers. This work aims to answer the following question: *Are there group-specific differences in acoustic-prosodic and linguistic features between truthful and deceptive interviewee responses?*

12.1 Method

We examined groups of speakers defined by gender (male or female), native language (Standard American English or Mandarin Chinese), and personality, defined by the NEO-FFI personality inventory Costa and McCrae [1989]. We computed the participants' NEO-FFI personality scores in five dimensions, Neuroticism (N), Extroversion (E), Openness to Ex-

perience (O), Agreeableness (A), and Conscientiousness (C). The NEO scores are on a continuous scale for each of the five dimensions.

In order to partition speakers into personality groups we binned the numeric personality scores to high, average or low for each dimension, using the thresholds provided in Locke [2015]. These thresholds were determined by psychologists based on population norms from a large sample of administered NEO-FFIs, and are different for males and females. Table 12.1 shows the mapping of numeric NEO scores to the three categorical labels.

Trait	*Gender*	*Low*	*Average*	*High*
N	Male	< 13	$13 =<, <= 21$	> 21
	Female	< 16	$16 =<, <= 25$	> 25
E	Male	< 24	$24 =<, <= 30$	> 30
	Female	< 25	$25 =<, <= 31$	> 31
O	Male	< 23	$23 =<, <= 30$	> 30
	Female	< 23	$23 =<, <= 30$	> 30
A	Male	< 29	$29 =<, <= 35$	> 35
	Female	< 31	$31 =<, <= 36$	> 36
C	Male	< 30	$30 =<, <= 37$	> 37
	Female	< 32	$32 =<, <= 38$	> 38

Table 12.1: Personality mapping from continuous scale to high, average, low.

As expected, the personality bins are highly unbalanced. Table 12.2 shows the distribution of participants in the high, average, and low personality bins for each of the 5 NEO dimensions.

Bin	*N*	*E*	*O*	*A*	*C*
Low	11.93	21.93	6.02	35.30	42.72
Average	40.00	36.77	41.92	45.35	39.93
High	48.07	41.30	52.06	19.35	17.35

Table 12.2: Distribution of participants in high, average, and low personality bins for each of the 5 NEO dimensions.

For each of the three group traits, we conducted two types of analysis. First, we directly compared deception performance measures (ability to deceive as interviewee, and ability to detect deception as interviewer) between speakers with different traits, to assess the effect of individual characteristics on deception abilities. In addition, we compared the features of deceptive and truthful language in subsets of the corpus, considering only people with a particular trait (e.g. all native Chinese speakers) in order to determine group-specific patterns of deceptive language. We examined the following four feature sets for individual differences: (1) Praat (2) LDI (3) LIWC (4) Complexity. These feature sets are described in detail in Chapter 4, Section 4.4. All features were z-normalized by speaker, so that features represent distance from a speaker's mean, measured in standard deviations. For example, we analyzed differences in acoustic-prosodic features between truthful and deceptive responses, considering only male responses and only female responses. We consider a cue to be gender-specific if a feature is significantly different between truthful and deceptive speech for only male speakers or only female speakers, but not both. To avoid noise, we eliminated LIWC features that did not appear in 90% of question response segments. This reduced our analysis to 42 of the 93 LIWC dimensions.

All tests for significance were corrected for family-wise Type I error by controlling the false discovery rate (FDR) at $\alpha = 0.05$. The k^{th} smallest p value is considered significant if it is less than $\frac{k*\alpha}{n}$. All data was balanced by gender and native language for this analysis. However, as shown in Table 12.2, the distribution of speakers across personality bins is unbalanced.

In all the tables in this chapter, we use D to indicate that a feature was significantly increased in deceptive speech, and T to indicate a significant indicator of truth. We consider

a result to approach significance if its uncorrected p value is $<= 0.05$ and indicate this with parentheses (e.g. "(D)") in the tables. We include trends in group-specific cues to deception since segmenting the data by group reduces the size of the data analyzed, and some of these trends might become statistically significant with additional data. Rows shaded in gray indicate that the features in those rows were not significant indicators of deception or truth across all groups of speakers. All analysis was done using the question response segmentation, which is the set of interviewee turn that are direct responses to the 24 biographical questions.

12.2 Gender Analysis

In this section we present the results of our analysis of gender in deceptive speech. We observed no difference across gender in ability to deceive ($t(300) = -0.38, p = 0.70$), nor in ability to detect deception ($t(300) = 0.64, p = 0.52$). There were also no differences in interviewer judgments across interviewee gender; that is, interviewers were not better at detecting deception for male or female interviewees ($t(300 = 0.22, p = 0.83)$). However, we observed many differences in cues to deception between male and female participants. We present an analysis of acoustic-prosodic, LDI, LIWC, and complexity feature sets below.

12.2.1 Acoustic Features

Table 12.3 shows the acoustic-prosodic cues to deception that differ for male and female participants.

Gender	*Feature*	*t*	*df*	*p*	*Sig.*
Male	Pitch Mean	2.22	3555	0.027	(D)
Male	Pitch Median	2.17	3539	0.03	(D)
Male	Pitch SD	3.17	3565	0.0016	D
Male	Intensity Min	-2.29	3572	0.022	(T)
Male	Intensity SD	2.68	3570	0.0075	D
Female	Intensity Mean	2.36	3560	0.018	(D)

Table 12.3: Gender-specific acoustic-prosodic cues to deception. Rows shaded in gray indicate cues that were not present across all speakers.

As shown is this table, there are several gender differences in acoustic-prosodic cues to deception. Across all subjects, standard deviation of pitch and intensity were increased in deceptive speech. However, when we segmented the data by gender and analyzed male and female responses separately, we found that both of these cues were only present in male speech. Intensity min was increased in truthful speech across all speakers, but this trend was only present in male speech. In addition, we found two new cues to deception in male speech that were not found across all speakers - increased pitch mean and median in deceptive speech. We also observed a female-specific cue to deception – intensity mean was increased in deceptive speech across all speakers, but this was true in female-only speech and not male-only speech. Pitch and intensity provided cues to deception for both genders, but in some cases in different ways.

12.2.2 LDI Features

Table 12.4 shows the LDI cues to deception that differed for male and female participants.

Gender	*Feature*	*t*	*df*	*p*	*Sig.*
Male	DAL.imagery	3.74	3561	0.00019	D
Male	hasAbsolutelyReally	2.02	1347	0.044	(D)
Male	hasFalseStart	2.59	2069	0.0097	D
Male	hasHedgePhrase	2.16	3157	0.031	(D)
Male	hasNot	-2.68	2068	0.0074	T
Male	hasWe	2.11	866	0.035	(D)
Male	numHedgePhrases	2.14	3139	0.032	(D)
Female	DAL.wc	-3.4	3538	0.00067	T
Female	hasContraction	-2.84	3294	0.0045	T

Table 12.4: Gender-specific LDI cues to deception. Rows shaded in gray indicate cues that were not present across all speakers.

This table shows several gender differences in LDI cues to deception. As with acoustic-prosodic indicators, there were more male-specific cues than female-specific cues. Of the 20 LDI features that were significantly different between truthful and deceptive responses across all participants, seven were significantly different or trended toward significance in male speakers only. For example, hedge words were increased in deceptive responses overall, but this was due to differences in male speakers' use of hedge words. No difference was observed in the use of hedge words when analyzing female responses alone. We also observed two female-specific cues – contractions and DAL.wc. Interestingly, use of contractions was significantly increased in truthful speech for female speakers, but there was no difference in contraction use between truthful and deceptive responses across all speakers. The Reid and Associates method of interrogation and interviewing Buckley [2000] teaches that contractions are a sign of truthful speech, based on the theory that contractions are indicative of more natural speech (e.g. "I didn't do it" is a more natural way to deny a crime than than "I did not do it"). Here we present a more nuanced finding – female speakers were more likely to use contractions in truthful speech, but male speakers were not.

12.2.3 LIWC Features

Table 12.5 shows the LIWC cues to deception that differed for male and female participants.

Gender	*Feature*	*t*	*df*	*p*	*Sig.*
Male	conj	2.37	3130	0.018	(D)
Male	focuspast	4.47	3177	8.20E-06	D
Male	nonflu	2.11	3531	0.035	(D)
Male	prep	2.84	3412	0.0046	D
Male	pronoun	2.08	3570	0.038	(D)
Male	relativ	2.45	3553	0.014	D
Male	space	3.84	3277	0.00013	D
Female	adj	-2.1	3041	0.036	(T)
Female	allPunc	-2.85	3450	0.0044	T
Female	apostro	-2.83	3286	0.0047	T
Female	netspeak	2.1	2827	0.035	(D)

Table 12.5: Gender-specific LIWC cues to deception. Rows shaded in gray indicate cues that were not present across all speakers.

We observed several gender differences in LIWC cues to deception, and again found more male-specific cues than female-specific cues. Of the 23 LIWC features that were different between truthful and deceptive responses across all participants, six were significantly different or trended toward significance in male speakers only, and two in female speakers only. For example, the *focuspast* dimension, which captures words used in past tense, was more frequent in deceptive responses overall, but this was due to the difference in male speakers' use of past tense when lying and telling the truth. No difference was observed in the use of past tense words when analyzing female responses alone. A female-specific cue that we observed was that female speakers used *apostrophes* more when telling the truth than when lying. Apostrophes only appeared in contractions in the transcriptions, so this supports the finding that contractions were an indicator of truth-telling for female speakers only. We also observed 3 new cues that were not present when analyzing all speakers. The

relativity dimension, which includes words such as "above," "near," and "new," was more frequent in deceptive responses for male speakers only. *Adjectives* and *allPunctuation* were more frequent in truthful speech for female speakers only.

12.2.4 Complexity Features

Complexity features were extracted using a system for automatic syntactic complexity [Lu, 2010]. Chapter 4 describes the complexity features in detail. Table 12.6 shows the complexity cues to deception that differed for male and female participants.

Gender	*Feature*	*t*	*df*	*p*	*Sig.*
Male	W	4.87	3236	1.10E-06	D
Male	DC	3.71	3225	0.00021	D
Male	CT	2.89	3308	0.0039	D
Male	CP	2.99	2768	0.0029	D
Male	CN	3.48	3310	0.00051	D
Male	MLS	4.75	3253	2.10E-06	D
Male	DC.C	2.73	3354	0.0064	D
Male	DC.T	4.28	3177	1.90E-05	D
Male	CT.T	3.07	3303	0.0022	D
Male	CP.T	2.95	2565	0.0032	D
Male	CP.C	2.58	2606	0.0098	D
Male	CN.T	3.83	3288	0.00013	D
Female	CN.C	2.1	3692	0.036	(D)

Table 12.6: Gender-specific complexity cues to deception. Rows shaded in gray indicate cues that were not present across all speakers.

Of the 19 syntactic complexity cues to deception that we observed across all participants, nine were male-specific cues and one was a trend observed in female speakers only. For example, *DC.C* (dependency clauses / number of clauses) and *DC.T* (dependency clauses / number of T-units) were cues to deception across all participants, but this finding was

true only for male speakers. *CN.C* (complex nominals / number of clauses) were more frequent in deceptive speech across all subjects, but this was true only for female speakers. We also observed three new syntactic complexity cues to deception for male participants only: deceptive responses from male speakers were characterized by an increased frequency of *CP* (coordinate phrases), *CP.T* (coordinate phrases / number of T-units) and *CP.C* (coordinate phrases / number of clauses). Coordinate phrases include adjective, adverb, noun, and verb phrases that are joined by a coordinating conjunction.

12.3 Native Language Analysis

Having identified many gender-specific cues to deception, in this section we present the results of our analysis of native language in deceptive speech.

We observed no difference between native speakers of English and Chinese in ability to deceive ($t(300) = -0.99, p = 0.32$). However, we did find a slight difference in ability to detect deception across native language ($t(300) = 1.67, p = 0.09$), although this difference was not statistically significant. Native Chinese speakers were slightly better at detecting deception than native English speakers – the average deception detection performance for native Chinese speakers was 57.8% and 55.58% for native speakers of English. Deception detection performance is defined here as $\frac{\#\text{ correct judgments}}{24} \times 100$.

There were no differences in interviewer judgments across interviewee native language, that is, interviewers were not better at detecting deception for native Chinese or Native English speakers ($t(300) = 0.62, p = 0.53$). However, we did observe a large difference in interviewer judgments across interviewee native language ($t(300) = 3.66, p = 0.0003$): on average, interviewers judged 61.71% of responses of native Chinese speakers as true, while they only judged 57.13% of responses of native English speakers as true. It seems that native Chinese speakers were trusted at a higher rate than native English speakers. To better understand this finding, we ran pairwise comparisons of interviewer judgments between 3 language types of pairs: (1) English-English, (2) English-Chinese, (3) Chinese-Chinese. We found a significant difference between English-English and Chinese-Chinese pairs ($t(90) = -2.29, p = 0.02$) – native Chinese interviewers who were paired with native

Chinese interviewees judged their partners as telling the truth more frequently (61.51%) than native English interviewers paired with native English interviewees (57.59%).

We also observed many differences in cues to deception between native speakers of Chinese and English. We present an analysis of acoustic-prosodic, LDI, LIWC, and complexity feature sets below.

12.3.1 Acoustic Features

Table 12.7 shows the acoustic-prosodic cues to deception that differ for native English and native Chinese speakers.

Native Lang	*Feature*	*t*	*df*	*p*	*Sig.*
English	Pitch Min	-2.15	3547	0.031	(T)
English	Intensity Mean	3.8	3554	0.00015	D
English	Intensity SD	2.07	3554	0.038	(D)
English	Jitter	-2.86	3443	0.0042	T
English	Shimmer	-2.35	3381	0.019	(T)
Chinese	Pitch Mean	2.01	3540	0.044	(D)
Chinese	Pitch SD	3.83	3552	0.00013	D
Chinese	Intensity Min	-2.01	3573	0.044	(T)
Chinese	Speaking Rate	-2.81	3573	0.005	T

Table 12.7: Native language-specific acoustic-prosodic cues to deception. Rows shaded in gray indicate cues that were not present across all speakers.

This table shows several differences in acoustic-prosodic cues to deception across native language. We previously found that for all participants, intensity mean and standard deviation and pitch standard deviation were increased in deceptive speech. In this analysis we find that intensity mean and standard deviation are only cues to deception for native English speakers, and pitch standard deviation are a cue specific to native speakers of Chinese. We also previously found that truthful speech was associated with increased pitch minimum and intensity minimum, but here we find that increased pitch minimum is specific to native

English speakers, and increased intensity minimum is specific to native Chinese speakers.

In addition, we have found four new deception indicators that were not present when studying all speakers. Truthful speech of native English speakers was characterized by an increase in jitter and shimmer. For native Chinese speakers, pitch mean was increased in deceptive speech, and speaking rate was increased in truthful speech. This last finding is intuitive; according to the cognitive theory of deception, we would expect non-native speakers to speak slower when lying and faster when telling the truth, since lying is a more cognitively taxing task. Further, consistent with the theory of Vrij *et al.* [2008], cognitive cues to deception (such as decreased speaking rate) should be more pronounced when deception is combined with a cognitively difficult task – in this case, speaking in one's non-native language.

12.3.2 LDI Features

Table 12.4 shows the LDI cues to deception that differed for native English speakers and native Chinese speakers.

Native Lang.	*Feature*	*t*	*df*	*p*	*Sig.*
English	DAL.wc	-3.59	3538	0.00033	T
English	hasCuePhrase	-3.71	3556	0.00021	T
English	hasHedgePhrase	2.6	2967	0.0093	D
English	hasI	3.41	3321	0.00066	D
English	hasLaugh	2.09	1915	0.037	(D)
English	hasNot	-3.01	1984	0.0027	T
English	numLaugh	2.18	1901	0.029	(D)
English	thirdPersonPronouns	2.95	2413	0.0032	D
Chinese	hasFalseStart	2.75	2221	0.006	D
Chinese	hasYes	6.5	3237	9.40E-11	D

Table 12.8: Native language-specific LDI cues to deception. Rows shaded in gray indicate cues that were not present across all speakers.

This table shows several differences across native language in LDI cues to deception. We found that there were more native English-specific cues than native Chinese-specific cues. Of the 20 LDI features that were significantly different between truthful and deceptive responses across all participants, six were significantly different for native English speakers only. For example, hedge words were increased in deceptive responses overall, but this was due to differences in native English speakers' use of hedge words. No difference was observed in the use of hedge words when analyzing native Chinese responses alone. It is interesting that we previously observed that hedge words were a male-specific cue to deception. Combining the gender and native language analyses, it seems that hedge words were increased in deceptive speech from male native English speakers.

We also observed two native Chinese-specific cues to deception – hasYes and hasFalseStart. The fact that deceptive responses from native Chinese speakers had on average more false starts is again consistent with the cognitive theory of deception and the extension of it by Vrij *et al.* [2008]. False starts are a form of speech disfluency that we would expect to see more of during deception due to the increase in cognitive load associated with lying. And it is also intuitive that this cue should be present in the responses of native Chinese speakers since they are speaking in their non-native language, which adds another level of cognitive load.

In addition, we observed a new cue to deception for native English speakers only – their deceptive responses had on average more instances of laughter. Laughter can be a sign of nervousness, or it can be used in an attempt to sound natural and relaxed. It is interesting that laughter as a cue to deception is specific to native speakers of English, and perhaps there are cultural differences in the use of laughter in dialogue.

12.3.3 LIWC Features

Table 12.9 shows the LIWC cues to deception that differed for native English and native Chinese speakers in our corpus.

Native Lang	*Feature*	*t*	*df*	*p*	*Sig.*
English	adverb	2.13	3133	0.033	(D)
English	conj	2.39	2921	0.017	D
English	focuspresent	-2.33	3538	0.02	T
English	I	1.97	3536	0.049	(D)
English	netspeak	2.28	2583	0.022	D
English	nonflu	2.81	3330	0.005	D
English	posemo	2.62	3005	0.0087	D
English	ppron	2.76	3554	0.0058	D
English	pronoun	2.78	3557	0.0055	D
English	social	2.75	3515	0.0061	D
English	tone	2.93	3228	0.0034	D
Chinese	cogproc	-3.11	3318	0.0019	T
Chinese	space	2.52	3364	0.012	D

Table 12.9: Native language-specific LIWC cues to deception. Rows shaded in gray indicate cues that were not present across all speakers.

We observed several differences across native language in LIWC cues to deception, and again found more cues that were specific to native English speakers than to native Chinese speakers. Of the 23 LIWC features that were different between truthful and deceptive responses across all participants, eight were significantly different in native English speakers only, and two in native Chinese speakers only.

For example, the *tone* and *posemotion* dimensions, which capture words with positive tone and emotion, were more frequent in deceptive responses overall, but this was due to the difference in native English speakers' use of positive words when lying and telling the truth. No difference was observed in the use of positive words when analyzing native Chinese responses alone. On the other hand, a cue that was specific to native Chinese speakers was *cogproc* (cognitive process) words, including "cause," "know," and "aught." These words were used more frequently in the truthful responses of native Chinese speakers. This follows the trend that we previously observed with other features – we found evidence of increased

cognitive load when lying for native Chinese speakers.

We also observed three new deception indicators in native English speakers that were not present when analyzing all speakers, and these are shaded in gray in Table 12.9. For example, the *focuspresent* category, which captures words in present tense, was used more frequently in truthful responses of native English speakers. It is interesting that this difference in usage of tense was only present for native English speakers and suggests that deception indicators that involve nuances in verb tense are specific to native speakers of English, and should not be applied to non-native speakers.

12.3.4 Complexity Features

Table 12.10 shows the complexity cues to deception that differed across native language.

Native Lang.	*Feature*	*t*	*df*	*p*	*Sig.*
English	DC	3.9	3351	9.60E-05	D
English	CT	2.79	3486	0.0054	D
English	CP	2.31	2807	0.021	D
English	CN	3.93	3544	8.60E-05	D
English	DC.C	3.46	3456	0.00054	D
English	DC.T	3.78	3362	0.00016	D
English	CT.T	2.99	3483	0.0028	D
English	CN.T	3.35	3581	0.00083	D
English	CN.C	3.01	3633	0.0026	D

Table 12.10: Native Language-specific complexity cues to deception. Rows shaded in gray indicate cues that were not present across all speakers.

Of the 19 syntactic complexity cues to deception that we observed across all participants, eight were only observed in native English, and there were no syntactic complexity cues that were specific to native Chinese speakers. In addition, we observed a new cue to deception that was only present in the responses of native English speakers, and not in the analysis of all speakers.

Deceptive responses from native English speakers were characterized by an increased frequency of *CP* (coordinate phrases). We previously observed that this was true for male speakers but not female speakers, so it seems that this finding is strongest for male native English speakers.

It is not surprising that we observed many more syntactic complexity cues to deception in native English speakers. Again, this emphasizes the importance of taking into account demographic information when analyzing deceptive speech, and not applying general rules about deception to all populations.

12.4 Personality Analysis

Having identified many gender-specific and native language-specific cues to deception, in this section we present the results of our analysis of personality in deceptive speech.

We ran Pearson's correlations between the five raw NEO scores and ability to deceive and to detect deception, and found no correlation for any trait. We also examined the relationship between personality and how often a person was believed or believed others. We found a slight negative correlation between interviewee Neuroticism and the rate of being believed by the interviewer – as interviewee Neuroticism scores increased, the percent of responses that the interviewer judged as true (i.e. believed) decreased ($r(300) = -0.13, p = 0.02$).

In order to analyze differences in cues to deception across personality types, we used a different method than the one we used for gender and native language analysis. That data was balanced for gender and native language, but the personality bins were highly unbalanced, as shown in Table 12.2. Thus, simply comparing t-test results for speakers that are in a "high" vs "low" personality bin would not be meaningful, since the results could be due to the amount of data in each bin rather than the personality characteristics. Therefore, for each speaker we computed TFdiff_f for each feature f:

$$\text{TFdiff}_f = \frac{\sum_{s_i \in F} f(s_i)}{size_F} - \frac{\sum_{s_i \in T}(s_i)}{size_T} \tag{12.1}$$

where F is the set of a speaker's false responses and T is the set of a speaker's truthful responses. $size_F$ is the number of segments in F and $size_T$ is the number of segments in

T. $f(s_i)$ is the value of feature f in segment s_i. Thus, TFdiff$_f$ represents the difference between the average feature f in a speaker's deceptive responses and truthful responses. A positive value of TFdiff$_f$ indicates that feature f was increased in deceptive speech, while a negative value indicates that f was decreased in deceptive speech.

Computing this measure allows us to identify salient cues to deception across personality bins using the one-way ANOVA. Since all features are speaker normalized, they represent a speaker's distance from their mean, measured in standard deviations. This minimizes the effect of speaker differences.

12.4.1 Acoustic Features

We ran one-way ANOVAs with TFdiff$_f$ as the dependent variable and the personality bin (low, average, high) as the independent variable. We repeated these tests for each feature and each NEO dimension, and corrected for family-wise Type I error by controlling the false discovery rate (FDR) at $\alpha = 0.05$. The k^{th} smallest p value is considered significant if it is less than $\frac{k*\alpha}{n}$.

Table 12.11 shows the ANOVA results for acoustic-prosodic features. "(*)" indicates that the p-value was less than 0.05 before correction for family-wise Type I error.

Trait	*Feature*	*df between*	*df within*	*F*	*p*	*Sig.*
N	Duration	2	297	3.44	0.033	(*)
N	Intensity Max	2	297	3.95	0.02	(*)
E	Shimmer	2	297	3.76	0.024	(*)
O	Duration	2	297	4.16	0.017	(*)
O	Intensity Min	2	297	3.18	0.043	(*)
O	Intensity SD	2	297	3.24	0.041	(*)
C	Intensity Min	2	297	4.15	0.017	(*)
C	Intensity Mean	2	297	3.38	0.035	(*)
C	Intensity SD	2	297	3.17	0.044	(*)
C	NHR	2	297	3.48	0.032	(*)

Table 12.11: ANOVA results comparing differences in acoustic prosodic features in deceptive and truthful responses across personality bins.

We see from this table that there were several differences in acoustic-prosodic indicators of deception across personality bins. However, none of the ANOVAs yielded statistically significant results after correction, so we consider these trends. We note that this analysis is done at the speaker level, and the data is balanced by gender and native language so there are 300 data points in total. Compared to the analysis of gender and native language, which was at the segment level (about 8k segments) we expect to see less statistical power for this analysis.

In order to identify where the differences in deception indicators occurred (i.e. which personality bins were significantly different from each other) we computed Tukey post-hoc tests for all ANOVAs with significant or approaching significant results. The results of the Tukey tests for acoustic-prosodic features are shown in Table 12.12. The columns "Avg-Low," "High-Low," and "High-Avg" represent the pairwise comparisons between those personality bins and the cell values are the p-values of the pairwise comparisons. P-values less than 0.05 are shaded in gray. The column "μ_H" represents the mean feature value for speakers in the high personality bin, "μ_A" for the average personality bin, and "μ_L" for the low personality bin.

Trait	*Feature*	*Avg-Low*	*High-Low*	*High-Avg*	μ_H	μ_A	μ_L
N	Duration	0.067	0.025	0.87	0.16	0.13	-0.06
N	Intensity Max	0.018	0.028	0.92	0.19	0.21	-0.03
E	Shimmer	0.055	0.028	0.97	0	-0.01	-0.16
O	Duration	0.2	0.023	0.2	0.18	0.09	-0.07
O	Intensity Min	0.23	0.047	0.37	-0.09	-0.03	0.13
O	Intensity SD	0.3	0.053	0.28	0.1	0.03	-0.11
C	Intensity Min	0.054	0.041	0.78	0.04	-0.01	-0.13
C	Intensity Mean	0.065	0.1	0.94	0.13	0.1	-0.02
C	Intensity SD	0.26	0.044	0.45	-0.04	0.04	0.12
C	NHR	0.025	0.8	0.38	0	0.08	-0.04

Table 12.12: Tukey post-hoc results for acoustic-prosodic cues to deception. Cells shaded in gray indicate a p-value less than 0.05.

The Tukey post-hoc tests revealed that the majority of the differences came from the High-Low comparison, and none of the differences were from the High-Avg comparison. This analysis revealed interesting differences in cues to deception across all personality types except Agreeableness. For example, we previously observed that duration was a cue to deception across all speakers – deceptive responses were on average longer than truthful responses. Here we see that this behavior was varied across personality. For the trait of Neuroticism, speakers in the high bin had longer deceptive responses, as evidenced by a positive TFdiff value for μ_H. However, speakers in the low bin had *shorter* deceptive responses than truthful responses. The same trend was true for the trait of Openness to Experience - speakers in the high bin for Openness had longer deceptive responses while speakers in the low bin had shorter deceptive responses. It seems that speakers in different personality groups exhibited cues to deception in different ways.

12.4.2 LDI Features

Table 12.13 shows the ANOVA results for LDI features. "(*)" indicates that the p-value was less than 0.05 before correction for family-wise Type I error.

Trait	*Feature*	*df between*	*df within*	*F*	*p*	*Sig.*
N	specScores	2	297	4.3	0.014	(*)
E	hasWe	2	297	5.73	0.0036	(*)
E	thirdPersonPronouns	2	297	3.03	0.05	(*)
O	hasYes	2	297	3.52	0.031	(*)
O	isJustYes	2	297	4.64	0.01	(*)
O	numFilledPauses	2	297	4.8	0.0089	(*)
O	specScores	2	297	3.64	0.027	(*)
A	specificDenial	2	297	3.2	0.042	(*)

Table 12.13: ANOVA results comparing differences in LDI features in deceptive and truthful responses across personality bins.

This table shows several differences in LDI indicators of deception across personality bins. As with the acoustic-prosodic analysis, none of the ANOVAs yielded statistically significant results after correction, so we consider these trends. In order to identify where the differences in deception indicators occurred (i.e. which personality bins were significantly different from each other) we computed Tukey post-hoc tests for all ANOVAs with significant or approaching significant results. The results of the Tukey tests for LDI features are shown in Table 12.14.

Trait	*Feature*	*Avg-Low*	*High-Low*	*High-Avg*	μ_H	μ_A	μ_L
N	specScores	0.011	0.081	0.38	0.1	0.17	-0.08
E	hasWe	0.0072	0.0071	1	0.05	0.05	-0.04
E	thirdPersonPronouns	0.11	0.99	0.077	0.03	0.13	0.02
O	hasYes	0.055	0.023	0.85	0.09	0.12	0.34
O	isJustYes	0.024	0.0071	0.77	-0.01	0.02	0.2
O	numFilledPauses	0.01	0.0074	1	0.17	0.17	-0.11
O	specScores	0.037	0.021	0.95	0.14	0.12	-0.12
A	specificDenial	0.77	0.034	0.12	0.04	-0.01	-0.03

Table 12.14: Tukey post-hoc for LDI cues to deception. Cells shaded in gray indicate a p-value less than 0.05.

As with acoustic-prosodic features, the Tukey post-hoc tests revealed that the majority of the differences came from the High-Low comparison, and none of the differences were from the High-Avg comparison. This analysis revealed interesting differences in cues to deception across all personality types except Conscientiousness. For example, we previously observed that *specScores*, which measure specificity in language, was a cue to deception across all speakers – deceptive responses were on average more specific than truthful responses. Here we see that this behavior was varied across personality. For the trait of Neuroticism, speakers in the high bin used more specific language when lying, as evidenced by a positive TFdiff value for μ_H. However, speakers in the low bin had *less* specific deceptive responses than truthful responses. The same trend was true for the trait of Openness to Experience - speakers in the high bin for Openness had more specific deceptive responses while speakers in the low bin had lower scores for specificity in deceptive responses. These findings support the trend that speakers in different personality groups exhibited cues to deception in different ways.

12.4.3 LIWC Features

Table 12.15 shows the ANOVA results for LIWC features. "(*)" indicates that the p-value was less than 0.05 before correction for family-wise Type I error.

Trait	*Feature*	*df between*	*df within*	*F*	*p*	*Sig.*
N	authentic	2	297	3.48	0.032	(*)
N	relativ	2	297	3.3	0.038	(*)
N	space	2	297	3.68	0.026	(*)
E	focuspast	2	297	3.22	0.042	(*)
O	work	2	297	4.23	0.015	(*)
A	informal	2	297	3.51	0.031	(*)

Table 12.15: ANOVA results comparing differences in LIWC features in deceptive and truthful responses across personality bins.

This table shows several differences in LIWC indicators of deception across personality bins. As with the previous features analyzed, none of the ANOVAs yielded statistically significant results after correction, so we consider these trends. In order to identify where the differences in deception indicators occurred (i.e. which personality bins were significantly different from each other) we computed Tukey post-hoc tests for all ANOVAs with significant or approaching significant results. The results of the Tukey tests for LIWC features are shown in Table 12.16.

Trait	*Feature*	*Avg-Low*	*High-Low*	*High-Avg*	μ_H	μ_A	μ_L
N	authentic	0.23	0.031	0.36	0.09	0.02	-0.12
N	relativ	0.071	0.03	0.89	0.07	0.04	-0.14
N	space	0.074	0.019	0.74	0.1	0.07	-0.11
E	focuspast	0.68	0.044	0.19	0.03	0.12	0.17
O	adverb	0.89	0.21	0.053	0.07	-0.05	-0.09
O	work	0.26	0.99	0.015	-0.03	0.11	-0.04
A	informal	0.024	0.4	0.69	0.16	0.21	0.07

Table 12.16: Tukey post-hoc for LIWC cues to deception. Cells shaded in gray indicate a p-value less than 0.05.

As with acoustic-prosodic and LDI features, the Tukey post-hoc tests revealed that the majority of the differences came from the High-Low comparison. This analysis revealed interesting differences in cues to deception across all personality types except Conscientiousness. For example, *informal* language was a cue to deception across all speakers - deceptive responses had on average less formal language than truthful responses. Here we see that this behavior was varied across personality. For the trait of Agreeableness, speakers in the average bin used more informal language when lying than when telling the truth, as evidenced by a positive TFdiff value for μ_A. However, speakers in the low bin had *less* informal deceptive responses than truthful responses. Again, these findings support the trend that speakers in different personality groups exhibited cues to deception in different ways.

12.4.4 Complexity Features

Table 12.17 shows the ANOVA results for complexity features. "(*)" indicates that the p-value was less than 0.05 before correction for family-wise Type I error.

Trait	*Feature*	*df between*	*df within*	*F*	*p*	*Sig.*
N	VP	2	297	5.67	0.0038	*
N	C	2	297	5.69	0.0038	*
N	DC	2	297	3.29	0.039	(*)
N	MLT	2	297	3.26	0.04	(*)
N	C.S	2	297	5.52	0.0044	*
N	VP.T	2	297	5.98	0.0028	(*)
N	C.T	2	297	5.88	0.0031	*
N	DC.T	2	297	3.3	0.038	(*)

Table 12.17: ANOVA results comparing differences in complexity features in deceptive and truthful responses across personality bins.

We see from this table that there were several differences in complexity indicators of deception across personality bins, but only for the Neuroticism dimension. Unlike previous features analyzed, for complexity we see that some of the ANOVAs yielded statistically significant results after correction. In order to identify where the differences in deception indicators occurred (i.e. which personality bins were significantly different from each other) we computed Tukey post-hoc tests for all ANOVAs with significant or approaching significant results. The results of the Tukey tests for LIWC features are shown in Table 12.18.

Trait	*Feature*	*Avg-Low*	*High-Low*	*High-Avg*	μ_H	μ_A	μ_L
N	VP	0.019	0.0025	0.63	0.18	0.13	-0.11
N	C	0.017	0.0024	0.66	0.19	0.14	-0.09
N	DC	0.12	0.03	0.66	0.1	0.06	-0.1
N	MLT	0.089	0.03	0.81	0.16	0.13	-0.05
N	C.S	0.017	0.0029	0.72	0.18	0.14	-0.09
N	VP.T	0.013	0.0018	0.66	0.18	0.14	-0.1
N	C.T	0.015	0.002	0.65	0.19	0.15	-0.08
N	DC.T	0.2	0.033	0.46	0.12	0.06	-0.08

Table 12.18: Tukey post-hoc for complexity cues to deception. Cells shaded in gray indicate a p-value less than 0.05.

The Tukey post-hoc tests revealed that all of the features were significantly different between bins high and low for Neuroticism, and some between average and low, but there were no significant differences between high and average. Interestingly, Neuroticism was the only personality trait with differences in complexity cues to deception. For all of these measures of syntactic complexity, speakers that were in the high Neuroticism bin produced more syntactically complex deceptive utterances than truthful utterances, as evidenced by a positive value for μ_H, while speakers that were in the low Neuroticism bin produced more syntactically complex *truthful* responses, as evidenced by a negative value for μ_L. This difference is very interesting and highlights the importance of considering individual differences determining the veracity of a speaker's statements.

12.5 Discussion

This chapter aimed to answer the following question: *Are there group-specific differences in acoustic-prosodic and linguistic features between truthful and deceptive interviewee responses?* We carefully analyzed differences in cues to deception across gender, native language, and personality types. We examined a variety of acoustic-prosodic and linguistic features and identified many group-specific cues to deception. In some cases, we found that

previously observed general cues to deception across all speakers were not present when we examined particular groups of speakers. In other cases, we discovered new cues to deception for specific groups of speakers that were not present when we analyzed all speakers.

Gender Pitch and intensity features provided cues to deception in different ways for male and female speakers. Cues to deception in male speakers included increased pitch mean, median, and standard deviation, and increased intensity standard deviation, while female deceptive speech was characterized by increased intensity mean. There were also differences in linguistic cues to deception between male and female subjects. Hedge words and phrases were increased in deceptive speech for male speakers only, as were false starts. Male speakers also used past tense verbs more when lying. Only female speakers had increased frequency of contractions in truthful responses. Adjective usage was increased in truthful responses for female speakers only. We observed differences in syntactic cues to deception across gender. Of the 19 syntactic complexity cues to deception that were observed across all speakers, nine were only present in male speakers, and only one trend was specific to female speakers. In general, there were more male-specific cues than female-specific cues to deception identified with this analysis.

Native Language Several differences in acoustic-prosodic cues to deception between native speakers of SAE and MC were identified. Jitter and shimmer were increased in truthful speech of SAE speakers only, and intensity mean and standard deviation were increased in deceptive speech of SAE speakers only. Native MC speakers had increased pitch mean and standard deviation in deceptive speech, and increased speaking rate in truthful speech. It is intuitive that only MC speakers spoke faster when telling the truth, since we expect cognitive cues to deception to be more pronounced when deception is combined with a cognitively difficult task – in this case, speaking in one's non-native language. There were also several differences in linguistic cues to deception between native SAE and MC speakers. For example, positive emotion words and laughter were increased in deceptive speech of SAE speakers only. Present tense verbs were also increased in SAE deceptive speech only. False starts and cognitive process words were increased in deceptive responses on MC speakers only, reinforcing the trend of cognitive cues to deception that were only present in MC speakers. Of the 19 syntactic complexity cues to deception across all subjects, eight were

only present in native SAE speakers, and none were specific to MC speakers. It seems that syntactic complexity features are more useful indicators of deception in native SAE speakers. In general, there were more SAE-specific cues than MC-specific cues to deception identified with this analysis.

Personality Differences in acoustic-prosodic and linguistic cues to deception were observed between subjects with different personality types, however these differences were not statistically significant after correction for multiple comparisons. The majority of the differences in cues to deception were between speakers who scored high vs. low for the five personality dimensions. For example, subjects who scored high in Neuroticism had increased intensity max when lying, while subjects who scored low in Neuroticism had decreased intensity max when lying. Subjects who scored high for Extroversion used "we" more when lying, while those who scored low used "we" more when telling the truth. Subjects who scored high in Openness used filled pauses more when lying, and those who scored low in Openness used filled pauses more when telling the truth. Differences in syntactic cues to deception between personality types were statistically significant, and they were all for speakers in high vs. low or average vs. low on the Neuroticism scale. All measures of syntactic complexity were increased in deceptive responses of speakers who scored high in Neuroticism. Overall, the greatest number of trait-specific cues were observed for Neuroticism (14), and the fewest for Agreeableness (2).

The findings presented in this chapter suggest that gender, native language, and personality all play a role in how people produce deceptive speech. Because of this, practitioners should be cautious about applying blanket rules about deceptive language to all populations. For example, some cues to deception that involve nuances in language such as verb tense changes were only present in native English speakers, and should not be applied to non-native speakers. Ideally, gender, native language, and personality should be taken into account when detecting deception. This is a difficult task for human practitioners. In the next chapter, we explore incorporating these individual differences into our machine learning models, with the goal of improving automatic deception detection.

While previous studies of deception have observed some variation in cues to deception across speakers, this work is the first comprehensive analysis of gender, native language,

and personality differences in acoustic-prosodic and linguistic cues to deception.

Chapter 13

Classification: Exploring Speaker Differences

This chapter presents the results of deception classification experiments that explore speaker differences in deceptive behavior. This work is motivated by our findings in Chapter 12, which showed a wide range of deceptive behavior across speakers of different genders, cultures, and personality types. Having identified these differences, we aimed to determine whether these speaker differences can be incorporated in our classification methods to improve the performance of automatic deception detection. In Chapter 6 we trained deception classifiers using acoustic-prosodic, lexical, and syntactic feature sets. In this chapter we aimed to answer the following question: **Can we use information about speaker characteristics to improve automatic deception detection?**

We explored three approaches to incorporate differences in deceptive behavior across speakers:

1. Classification with individual traits as features
2. Classification with homogenized data
3. Classification with speaker dependent features

In the first approach, we included features that indicate the gender, native language, and personality scores of the speaker. In the second approach, we trained gender-specific and

native language-specific classifiers using homogeneous training data from subjects sharing the same trait (e.g. female speakers). In the third approach, we included speaker-dependent features extracted from a baseline sample of speech for each speaker, to capture a speaker's deviation from their natural speaking style during deception.

For each approach, we trained classifiers using the combined feature sets detailed previously in Section 6.2. The feature sets are:

- Acoustic
- Lexical
- Syntactic
- Acoustic+Lexical
- Acoustic+Syntactic
- Lexical+Syntactic
- Acoustic+Lexical+Syntactic

For each of these feature sets, we compared the results of the models that incorporate speaker differences with the results of generic models presented in Chapter 6. We did this for each of the four segmentation units: IPU, turn, question response, and question chunk.

We used the same training and evaluation framework for all experiments in this chapter as in Chapter 6 (unless otherwise noted). We used the same folds for our 10-fold cross-validation setup, as well as the same classifiers and parameters, to ensure that these experimental results are directly comparable with our previous results.

13.1 Classification with Individual Traits as Features

In this section we present the results of classification experiments that use speaker differences by including individual traits as features. This first approach is straightforward – we append a 7-element vector to the existing feature vector that represents the speaker's gender (male or female), native language (English or Chinese), and five raw personality scores

from the NEO-FFI (Neuroticism, Extroversion, Openness to Experience, Agreeableness, and Conscientiousness).

The motivation for this approach is that we observed differences in cues to deception between male and female speakers, native English and native Chinese speakers, and between speakers with different personality types. By including these speaker traits as features, perhaps the classifiers can learn how the acoustic-prosodic and linguistic characteristics of a speaker's responses interact with their unique combination of speaker traits to signal deception or truth.

The tables below show the classification results. For each feature set, we show the performance of the generic classifier without speaker traits (*Generic*) along with the results of the classifier trained with the additional speaker trait features (*SpeakerTrait*). The column labeled $|Generic - ST|$ shows the absolute value of the difference between the performance of the generic classifier and the performance of the classifier trained with the additional speaker trait features. Shaded gray cells indicate which model performed better. The performance metric shown in the tables is accuracy. We compared accuracy because this enabled us to test whether the differences in classifier performance were statistically significant, using a two-tailed .95 confidence interval. (This is not possible with F1, which does not have a probabilistic interpretation.)

Table 13.1 shows the results of combined feature sets with speaker traits for IPU classification.

Feature	*Generic*	*SpeakerTrait*	$\lvert Generic - ST \rvert$
Acoustic	52.90	52.87	0.03
Lexical	56.01	56.07	0.06
Syntactic	51.12	51.09	0.03
Acoustic+Lexical	56.25	56.25	0.00
Acoustic+Syntactic	52.72	52.72	0.00
Lexical+Syntactic	56.00	56.01	0.01
All	56.29	56.32	0.03

Table 13.1: IPU classification accuracy with combined feature sets + speaker traits (ST). Shaded cells indicate which model performed better.

Overall, we did not find that adding speaker traits was useful at the IPU-level. The differences between the generic and speaker trait classifiers were minuscule for all feature sets, with an average difference in performance of .02% This suggests that combining speaker traits with acoustic-prosodic and linguistic features is not useful for detecting deception in IPU segments.

Table 13.2 shows the results of combined feature sets with speaker traits for turn classification.

Feature	*Generic*	*SpeakerTrait*	$\lvert Generic - ST \rvert$
Acoustic	52.98	53.00	0.02
Lexical	58.03	57.93	0.10
Syntactic	52.15	52.14	0.01
Acoustic+Lexical	59.77	59.46	0.31
Acoustic+Syntactic	53.03	53.02	0.01
Lexical+Syntactic	57.86	57.91	0.05
All	57.86	57.84	0.02

Table 13.2: Turn classification accuracy with combined feature sets + speaker traits (ST). Shaded cells indicate which model performed better.

This table shows a similar trend for turn classification that we saw for IPU classification. There were very slight differences between the generic and speaker trait classifiers for turn classification, none of which were significant at a .95 confidence interval. The mean difference in classifier performance was only 0.07%, and in some cases the classifier performance was lower after adding speaker traits. Including speaker traits for turn classification did not significantly improve performance.

Table 13.3 shows the results of combined feature sets with speaker traits for classification of question responses.

Feature	*Generic*	*SpeakerTrait*	$\lvert Generic - ST \rvert$
Acoustic	56.40	56.36	0.04
Lexical	64.43	64.63	0.20
Syntactic	66.05	66.11	0.06
Acoustic+Lexical	63.47	63.47	0.00
Acoustic+Syntactic	64.31	64.36	0.05
Lexical+Syntactic	65.77	65.64	0.13
All	63.69	63.77	0.08

Table 13.3: Question response classification accuracy with combined feature sets + individual traits. Shaded cells indicate which model performed better.

As with IPU and turn classification, we observed no significant differences between the generic and speaker trait classifiers for question response classification. The mean difference in classifier performance was .08% and in some cases performance was lower after including speaker traits. It seems that including speaker traits for question response classification was not helpful in improving deception detection performance.

Table 13.4 shows the results of combined feature sets with speaker traits for classification of question chunks.

Feature	*Generic*	*SpeakerTrait*	$\lvert Generic - ST \rvert$
Acoustic	58.10	58.09	0.01
Lexical	64.96	64.99	0.03
Syntactic	69.34	69.29	0.05
Acoustic+Lexical	66.31	66.33	0.02
Acoustic+Syntactic	69.24	69.29	0.05
Lexical+Syntactic	69.81	69.73	0.08
All	69.43	69.42	0.01

Table 13.4: Question chunk classification accuracy with combined feature sets + individual traits. Shaded cells indicate which model performed better.

Question chunk classification followed the same trend as the other segmentations, where differences between the generic and speaker trait models were marginal. The mean difference in classifier performance was 0.04%. It seems that including speaker traits for question chunk classification did not improve classification performance.

Overall, our experimental results suggest that adding speaker traits as features was not useful for IPUs, turns, question responses, or question chunk segmentation. None of the differences between the results of the generic classifier and the speaker trait classifier were statistically significant. It is possible that this approach was too simplistic. The generic classifiers were trained using hundreds of segment-level features, some of which were significantly different between truthful and deceptive speech. It seems that adding a handful of speaker trait features, which do not differentiate between truthful and deceptive segments on their own, was not helpful for the classification.

13.2 Classification with Homogenized Data

We previously observed that including traits as features did not significantly improve deception classification performance. In this section we explore a second method to leverage speaker variability in deception classification, namely, data homogenization. Motivated by our findings that male and female speakers, as well as native Chinese and native English

speakers, exhibit cues to deception differently, we hypothesized that training gender-specific and language-specific deception classifiers could improve performance over generic classifiers. Further, data homogenization was successfully used by An and Levitan [2018] for personality identification using the CXD corpus, motivating our experiments in deception classification. To test this hypothesis, we trained three versions of each deception classifier:

1. Generic
2. Gender-specific
3. Language-specific

For the gender-specific models, we trained a male classifier using only male speakers and a female classifier using only female speakers. At inference time, we used the male classifier to classify deception for male test speakers, and the female classifier to classify deception for female test speakers. We used the same approach for the language-specific classifier, training an English classifier and a Chinese classifier using only speakers with that native language, and applying the appropriate classifier at inference time. Because the data is balanced by gender and native language, the gender-specific and language-specific models were trained using half of the training data available. To ensure a fair comparison between generic and homogenized classifiers, we trained the generic classifier on a randomly selected half of the training data, so all classifiers were trained using the same amount of data.

We compared generic, gender-specific, and language-specific classifiers for all four segmentation units (IPU, turn, question response, question chunk) as well as for the seven feature sets used in the above experiments (acoustic, lexical, syntactic, and combinations).

Table 13.5 shows the results comparing generic and homogenized models.

Feature	*Generic*	*Gender-specific*	*Lang-specific*	$\lvert Generic - best \rvert$
Acoustic	52.24	52.45	52.49	0.25
Lexical	54.86	54.39	54.85	0.00
Syntactic	51.08	50.97	50.91	0.00
Acoustic+Lexical	55.13	54.79	55.03	0.00
Acoustic+Syntactic	52.22	52.22	52.46	0.24
Lexical+Syntactic	54.80	54.51	54.81	0.01
All	55.00	54.84	55.05	0.05

Table 13.5: IPU generic vs. homogenized classification accuracy accuracy with combined feature sets. Shaded cells indicate which model performed best.

We see from this table that the language-specific classifier outperformed the generic classifier for several feature combinations, including acoustic, acoustic+syntactic, lexical+syntactic, and all features combined. However, the margins of improvement were very small (the largest was .25%) and none were statistically significant at a .95 confidence interval.

Table 13.6 shows the comparison between generic and homogenized models for classification of turns.

Feature	*Generic*	*Gender-specific*	*Lang-specific*	$\lvert Generic - best \rvert$
Acoustic	52.42	51.92	52.77	0.35
Lexical	56.35	56.50	57.02	0.67
Syntactic	52.10	51.86	51.87	0.00
Acoustic+Lexical	57.09	57.06	56.90	0.00
Acoustic+Syntactic	52.35	52.19	52.23	0.00
Lexical+Syntactic	56.77	56.42	57.08	0.31
All	56.19	56.00	56.13	0.00

Table 13.6: Turn generic vs. homogenized classification accuracy with combined feature sets. Shaded cells indicate which model performed best.

As shown in this table, the language-specific classifiers outperformed the generic classifier

for the acoustic, lexical and lexical+syntactic feature sets. However, as with IPUs, the margins of improvement were small (less than 1% improvement for all feature combinations) and none were statistically significant at a .95 confidence interval.

Table 13.7 shows the comparison between generic and homogenized models for classification of question responses.

Feature	*Generic*	*Gender-specific*	*Lang-specific*	$\lvert Generic - best \rvert$
Acoustic	55.21	54.91	55.22	0.01
Lexical	60.47	60.84	61.76	1.29
Syntactic	63.21	63.48	63.09	0.27
Acoustic+Lexical	58.81	58.94	60.17	1.36
Acoustic+Syntactic	60.03	61.51	61.54	1.51
Lexical+Syntactic	62.34	62.43	62.81	0.47
All	59.64	60.65	60.87	1.23

Table 13.7: Question response generic vs. homogenized classification accuracy with combined feature sets. Shaded cells indicate which model performed best.

As shown in this table, the gender- and language-specific classifiers outperformed the generic classifier for all feature combinations. The gender-specific model was preferred for one feature set (syntactic), and the language-specific model was preferred for all other feature combinations (acoustic, lexical, acoustic+lexical, acoustic+syntactic, lexical+syntactic, and all features). The margins of improvement were larger for question responses than for IPUs and turns for some feature sets (as high as 1.5% for acoustic+syntactic). However, none were statistically significant at a .95 confidence interval.

Table 13.8 shows the comparison between generic and homogenized models for classification of question chunks.

Feature	*Generic*	*Gender-specific*	*Lang-specific*	$\lvert Generic - best \rvert$
Acoustic	57.24	56.55	56.83	0.00
Lexical	61.90	61.64	61.87	0.00
Syntactic	68.73	67.77	68.10	0.00
Acoustic+Lexical	62.85	62.94	62.78	0.09
Acoustic+Syntactic	68.84	67.65	68.17	0.00
Lexical+Syntactic	69.69	68.41	68.11	0.00
All	68.99	68.30	68.19	0.00

Table 13.8: Question chunk generic vs. homogenized classification accuracy with combined feature sets. Shaded cells indicate which model performed best.

We see from this table that unlike the results for IPUs, turns, and question responses, the generic classifier was preferred for all feature sets except one. It seems that the classifiers trained and evaluated with the question chunk segmentation did not benefit from using homogenized data. Although none of the performance differences between the generic and homogenized models were statistically significant, we observed a trend that the homogenized models were the most useful for question responses, and the least useful for question chunks. It is possible that question chunks benefit the least because they have the advantage of the most contextual information, which potentially outweighs the benefits of leveraging speaker trait information. Another trend that we observed is that the language-specific models tended to perform better than the gender-specific models, suggesting that there were more benefits from incorporating language-specific deceptive behaviors than gender-specific deceptive behaviors.

Overall, we conclude that classification with homogenized models did not significantly improve deception detection performance. This is contrast to the personality detection work of An and Levitan [2018], which found that homogenized models significantly outperformed the baseline generic personality classifiers. It is possible that our generic deception classifiers were more optimized than the generic personality classifiers, and therefore had less room for improvement. Another possibility is that there are greater gender and cultural differences in personality expression than in deception behavior.

13.3 Classification with Speaker-Dependent Features

In this section we explore a third method to leverage speaker variability in deception classification: classification with speaker-dependent features. Practitioners are often trained to establish a baseline behavior for a subject, and then look for deviation from the baseline to assess the veracity of a subject's statements. For example, Reid and Associates [Buckley, 2000] train interviewers to begin interviews by asking neutral questions that are easily verifiable (such as the subject's name, age, occupation) and observe how the subject behaves when responding truthfully to establish baseline behavior. These training instructions motivated Enos [2009] to develop subject-dependent features. These features captured each speaker's tendency toward certain behaviors when lying and telling the truth. For example, they developed features that captured speaker ratios of laughter and filled pauses in deceptive and truthful speech. They reported that these features improved classification performance. A drawback of this approach is that it requires data annotated with truth and deception labels for each speaker in order to train the classifier. To use a speaker-dependent classifier to detect deception in a new, unseen speaker, one would first have to obtain labeled truth and deception data for that new speaker and compute these ratios. This is an unrealistic expectation in a real-world situation.

In our work we aimed to avoid this constraint. In our experimental paradigm for collecting the CXD corpus, we included an initial "baseline" interview between an experimenter and the subject. During this interview, the experimenter asked the subject open-ended questions that were designed to elicit spontaneous speech (e.g. "What do you like best/worst about living in NYC?"). Subjects were instructed to respond truthfully during this baseline session. We collected 3-4 minutes of subject speech for each participant, and this enables us to establish a baseline behavior for each subject and look for deviations from this baseline to help with classification decisions. To do this, we extracted features from the baseline session and combined those features with the features extracted from the lying game. In a real-world application, it is conceivable that one can obtain a sample of a speaker speaking truthfully by asking them to answer a few neutral, verifiable questions.

We extracted the following feature sets from the baseline session:

1. Acoustic: Praat, openSMILE (IS09)
2. Lexical: N-grams, LIWC, LDI
3. Syntactic: Complexity, POS, word+POS

These features were extracted from IPU segments. (We did not define turns in the baseline data since it was not a dialogue between the participant and the experimenter, but rather the participant responding to a series of prompts.) The features are described in detail in Chapter 4, Section 4.4. We computed mean feature vectors, representing the mean value of each feature in the baseline data, for a particular speaker. For example, the mean acoustic feature vector of a speaker consisted of the mean value of each Praat and openSMILE feature across all subject IPUs from the baseline data. We then subtracted the baseline feature vector from each feature vector extracted from the interview session, to capture a speaker's deviation from their baseline behavior. To evaluate the performance of these speaker-dependent features, we compared three approaches:

1. Generic: trained classifier with only session features, and no baseline features
2. Speaker-dependent: trained classifier with only speaker-dependent features (i.e. session features minus baseline features)
3. Combined: trained classifier with session features concatenated with speaker-dependent features

All classifier parameters were consistent across the three conditions, except for the number of features used for classification, which was increased for the combined features. The classifiers and parameters used here were the same as those used in our original deception classification experiments, described in Chapter 6. All models were evaluated with 10-fold cross-validation.

Table 13.9 shows the comparison between generic, speaker-dependent, and combined models for classification of IPUs.

Feature	*Generic*	*Speaker-dependent*	*Combined*	$\lvert Generic - Best \rvert$
Acoustic	52.90	52.90	52.88	0.00
Lexical	56.01	55.91	56.43	0.42
Syntactic	51.12	51.14	51.13	0.02
Acoustic+Lexical	56.25	56.16	56.73	0.48
Acoustic+Syntactic	52.72	52.73	52.94	0.22
Lexical+Syntactic	56.00	56.01	56.56	0.56
All	56.29	56.28	56.76	0.47

Table 13.9: IPU speaker-dependent classification accuracy with combined feature sets. Shaded cells indicate which model performed best.

As shown in this table, the classifiers trained with speaker-dependent features or combined features outperformed the generic classifiers for all feature sets except for acoustic. The model trained with only speaker-dependent features did best for the syntactic feature set, while the models trained with speaker-dependent and generic features combined performed best for all other feature combinations. However, the differences in classifier performance were marginal (the mean improvement was 0.32%), suggesting that adding speaker-dependent features was not very helpful for improving deception classification for IPU segments.

Next, we repeated these experiments using the turn segmentation. Table 13.10 shows the comparison between generic and speaker-dependent models for classification of turns.

Feature	*Generic*	*Speaker-dependent*	*Combined*	$\lvert Generic - Best \rvert$
Acoustic	52.98	52.97	52.91	0.00
Lexical	58.03	57.98	59.03	1.00
Syntactic	52.15	52.07	52.17	0.02
Acoustic+Lexical	59.77	59.43	59.92	0.15
Acoustic+Syntactic	53.03	53.03	53.03	0.00
Lexical+Syntactic	57.86	58.19	59.01	1.15
All	57.86	57.83	58.38	0.52

Table 13.10: Turn speaker-dependent classification accuracy with combined feature sets. Shaded cells indicate which model performed best.

As shown in this table, the classifiers trained with both session features and speaker-dependent features (*Combined*) performed best for all feature sets except for the acoustic feature set (for which the generic model performed marginally better). The largest margin of improvement was 1.15% for lexical+syntactic features. It seems that speaker-dependent features were more helpful for turn classification than for IPU classification.

Next, we examined the impact of training with speaker-dependent features on question response classification. Table 13.11 shows the comparison between generic and speaker-dependent models for classification of question responses.

Feature	*Generic*	*Speaker-dependent*	*Combined*	$\lvert Generic - Best \rvert$
Acoustic	56.40	56.41	56.52	0.12
Lexical	64.43	64.06	65.95	1.52
Syntactic	66.05	65.82	66.02	0.00
Acoustic+Lexical	63.47	63.49	64.08	0.61
Acoustic+Syntactic	64.31	64.39	64.49	0.18
Lexical+Syntactic	65.77	65.65	65.65	0.00
All	63.69	63.61	63.95	0.26

Table 13.11: Question response speaker-dependent classification accuracy with combined feature sets. Shaded cells indicate which model performed best.

As shown in this table, speaker-dependent features improved question response classification accuracy for all feature sets except for syntactic and lexical+syntactic. Using only speaker-dependent features achieved the best performance for acoustic features, and a combination of speaker-dependent and generic features yielded the best performance for all other feature sets. The margin of improvement was largest for lexical features (1.5%). As with turn classification, we found that adding speaker-dependent features improved deception classification performance for question response segmentation.

Finally, we examined the impact of training with speaker-dependent features on question chunk classification. Table 13.12 shows the comparison between generic, speaker-dependent, and combined generic+speaker-dependent models for classification of question chunks.

Feature	*Generic*	*Speaker-dependent*	*Combined*	$\lvert Generic - Best \rvert$
Acoustic	58.10	58.10	58.69	0.59
Lexical	64.96	65.06	67.96	3.00
Syntactic	69.34	69.19	69.69	0.35
Acoustic+Lexical	66.31	66.36	67.80	1.49
Acoustic+Syntactic	69.24	69.06	69.62	0.38
Lexical+Syntactic	69.81	69.59	70.22	0.41
All	69.43	69.49	69.90	0.47

Table 13.12: Question chunk speaker-dependent classification accuracy with combined feature sets. Shaded cells indicate which model performed better.

As shown in this table, combining speaker-dependent features with generic features improved question chunk classification performance for all feature sets. The margin of improvement was greatest for lexical features (3%).

13.4 Discussion

This chapter aimed to answer the question: *Can we use information about speaker characteristics to improve automatic deception detection?* We tested three approaches to incorporate speaker-dependent information in deception classification: adding speaker traits as features, training homogenized models, and adding speaker-dependent features. We found that adding speaker traits did not improve classification performance. The classifiers were trained with many acoustic-prosodic, lexical, and syntactic features, and simply adding speaker traits as features, which were the same for all truthful and deceptive speaker utterances, was not useful for improving deception classification. Homogenized models improved performance under some conditions, particularly for the question response segmentation. In almost all cases where the homogenized model improved over the generic model, we found that it was the language-specific model that did best, not the gender-specific model. It seems that there were larger gains from training classifiers with data from speakers with the same native language, than from training classifiers with data from speakers with the

same gender.

The largest improvements were obtained from the third approach of adding speaker-dependent features, and particularly for the question chunk segmentation. The speaker-dependent features were computed by extracting features from the 3-4 minute initial interview with each subject, in which subjects were instructed to answer truthfully to each question, in order to establish baseline speaking behavior when telling the truth. The baseline features were then subtracted from the session features, to capture distance from the baseline. We found that combining speaker-dependent features with session features was better than using only speaker-depending features. The most useful speaker-dependent features were the lexical and syntactic features. There were smaller improvements across all segmentations from adding speaker-dependent acoustic features to interview session acoustic features.

Overall, we conclude that adding speaker-dependent features that captured speakers' deviation from their baseline speaking behavior improved deception classification performance. For IPUs and question chunks, the improvements were marginal, while for turn, question response, and question chunk segmentations the improvements were larger (1-3% for some feature combinations). For some segmentations, we achieved a new best performance using a combination of generic and subject-dependent features, supporting the hypothesis that capturing deviations from baseline behavior is helpful for deception detection. The improvements from adding speaker-dependent features were not statistically significant at a .95 confidence interval, so we consider them trends that should be further studied. Practitioners have advocated for interviewing practices that establish baseline behavior of subjects while telling the truth, and then looking for differences from the baseline to detect deception. Baseline behavior is often elicited by first asking neutral questions that the subject is expected to answer truthfully. In this work we operationalized a method to automatically capture deviations from the baseline, instead of relying on human judgment to determine deviation from the baseline.

In conclusion, not only are there differences in production of deception across speakers, but our experimental results suggest that those speaker differences can be leveraged to improve automatic deception classification. Future work can explore modeling speaker

traits in other ways. For example, recent work by An *et al.* [2018] used a multi-task learning framework to jointly predict speaker personality and utterance deception. They found that this approach performed better for deception classification than including personality scores as features. This is very promising work, and can be extended to include gender and native language. Modeling speaker differences in creative ways can help further push the state-of-the-art in automatic deception classification.

Chapter 14

Speaker-Dependent Deception Classification Using Neural Network Models

In Chapter 13, we explored three approaches to incorporate speaker differences in deception classification. The three methods had varying degrees of success, with the best approach using speaker-dependent features extracted from the baseline sample of speech. In this chapter we explore another approach for speaker-dependent deception classification: training classifiers using training instances from the same speakers that we evaluated the models on.

In all prior experiments described in this thesis, we trained models using features extracted from a set of speakers, and evaluated them on a distinct set of speakers. This was done to ensure that the models do not overfit to a specific set of speakers, but rather they learn generalizable patterns of deceptive speech that extend to unseen test speakers. In this section we explored a method of speaker-dependent classification, where instead of splitting train and test sets by speaker, we split the data randomly by instances, so that there were segments in train and test from the same speakers. Although this paradigm is difficult to replicate in a real-world scenario, these experiments were conducted to see whether deception classification could greatly benefit from having some labeled training data available for

a particular speaker. We refer to the two approaches as "speaker split" and "random split."

We first compared speaker and random split experiments for the feature sets and classifiers described in Chapter 6. Those experiments used standard statistical machine learning algorithms (e.g. Random Forest, Logistic Regression, Support Vector Machines, and Naive Bayes) and acoustic, lexical, and syntactic feature sets. Our results showed no significant differences between the models trained on speaker split and random split data. This suggests that the classifiers trained and evaluated using the same speakers (but not the same instances) did not learn speaker-specific patterns of deception.

In our next set of experiments, we compared classifiers trained and evaluated on speaker split data vs. random split data using a new set of classification models: neural network models. Neural network models are currently the state-of-the-art in many computer vision, speech recognition, and NLP tasks such as POS tagging. They have not been previously explored in the context of deception detection, probably because they typically require large training sets, which are not available for deception. Given the relatively large size of the CXD corpus, this was not a constraint for our work. An advantage of deep neural networks is that multiple feature streams can be combined in a single architecture. This is especially useful for handling both lexical content from the speech transcription jointly with acoustic-prosodic features extracted from the speech signal. In the remainder of this chapter, we first describe three neural network architectures that we developed for deception classification. We then present classification results for both speaker-split evaluation and random-split evaluation.

Some of this work was published in Mendels *et al.* [2017], and was done in collaboration with my co-authors Gideon Mendels and Kai-Zhan Lee.

14.1 Neural Network Architectures

In this section we describe the three neural network models that we developed for deception classification.

1. LSTM-lexical, trained on word embeddings

2. DNN-acoustic, trained on openSMILE features (IS09)

3. Hybrid, a combination of the LSTM-lexical and DNN-acoustic models

We used Keras [Chollet and others, 2015] with a TensorFlow backend for all model implementations. We used Bayesian hyper-parameter optimization [Snoek *et al.*, 2012], as implemented by the `spearmint` library [Group, 2017] to select the optimal hyper-parameters for our models.

LSTM-lexical

In our previous statistical machine learning experiments, we observed that lexical features, and particularly n-grams, were useful for deception classification. However, lexical features have the disadvantage of capturing domain-specific trends. Another drawback of n-grams is that they do not capture context or semantic relationships between words. Therefore, we designed this lexical neural model trained on word embeddings, a distributed representation of words that capture context and semantic similarity between words. The model is based on the bidirectional long short-term memory (BLSTM) architecture. Recurrent models have been successful in related tasks of sentiment classification [Tang *et al.*, 2015], speech recognition [Graves *et al.*, 2013] and emotion detection [Trigeorgis *et al.*, 2016]. The BLSTM model [Schuster and Paliwal, 1997] is a modification of the original long short-term memory (LSTM) model [Hochreiter and Schmidhuber, 1997] in that it analyzes input simultaneously in the forward and reverse time directions. The effectiveness of both models comes from the capacity of an LSTM node to retain memory of its prior values with an internal state, bridging long temporal gaps. For every node at a given time-step t, with output gate y_{out}, input gate y_{in}, forget gate net, and differentiable activation functions g, h, output is defined as $y(t) = y_{out}(t)h(s(t))$ with internal state $s(t) = s(t-1) + y_{in}(t)g(net(t))$ [Hochreiter and Schmidhuber, 1997]. We used pre-trained word embeddings described in Chapter 4, since our corpus is relatively small for training word embedding models. These GloVe embeddings were used to initialize the weights, and we allowed back-propagation to update embedding values during training. We used a single softmax layer that operated on the final output and state of the LSTM for prediction. Our final model used a cell size of 256.

DNN: openSMILE

In our statistical machine learning experiments, we observed that the openSMILE feature set was somewhat discriminative between truthful and deceptive speech. We designed a

deep neural network model (MLP) using the the same feature set. Prior to training, we normalized our features by removing the mean and scaling to unit variance. Centering and scaling were done independently on each feature. Our model consisted of six fully connected layers, each with 1095 hidden units followed by a Relu activation. For prediction we used a softmax layer with two outputs that corresponds to the two classes in our task. We used categorical cross-entropy as our loss function. During training the output of each layer was normalized using Batch Normalization [Ioffe and Szegedy, 2015] and passed through a Dropout layer [Srivastava *et al.*, 2014] with a 0.497 probability. Our model has many parameters and a high dropout rate reduces the risk of over-fitting. Additionally, we added L2 regularization on the weights with a value of 0.2. We trained our model using stochastic gradient descent with a learning rate of 0.00134 that reduced by 50% for every 10 epochs with no improvement on training loss. The above hyper parameters were obtained using the Bayesian Optimization method implemented by the Spearmint library [Group, 2017].

Hybrid: LSTM + DNN

One of the advantages of neural networks is the ability to tailor the architecture to the task and combine sequential and discrete features in a single model. In our final model, we combined our LSTM-lexical and DNN-openSMILE models. Unlike most ensemble methods, our hybrid model was trained jointly without explicit voting between the acoustic and lexical based areas. We first experimented with merging the two models by taking the output of the last hidden layer in our DNN model and concatenating it with the output of LSTM, using the softmax function to normalize the last layer's output and generate class probabilities.

However, this architecture failed to improve on the original DNN model, which led us to the hypothesis that during back-propagation, the acoustic-based area of the network was being penalized more than the lexical area. To test our hypothesis, we attached an auxiliary softmax prediction layer to the LSTM output and used it to predict the test set. We observed that this area of the network achieved lower performance than the original LSTM-lexical model. This result confirmed our hypothesis that although the overall loss seemed to converge, the lexical area of the network was not optimized. Although it is possible to freeze the weights of the acoustic area and continue training the lexical area, that approach is not preferred due the manual intervention required. Instead, we computed

the loss of the network twice: once for the main softmax and once for the auxiliary softmax. Using a parameter λ we computed a weighted sum of the two error matrices. This approach allowed us to train the network without manual intervention. We treated λ as a hyper-parameter and using Bayesian optimization found an optimal value of 0.67 which doubles the significance of the loss computed from the auxiliary softmax compared to the main softmax. The architecture of this hybrid model is illustrated in Figure 14.1.

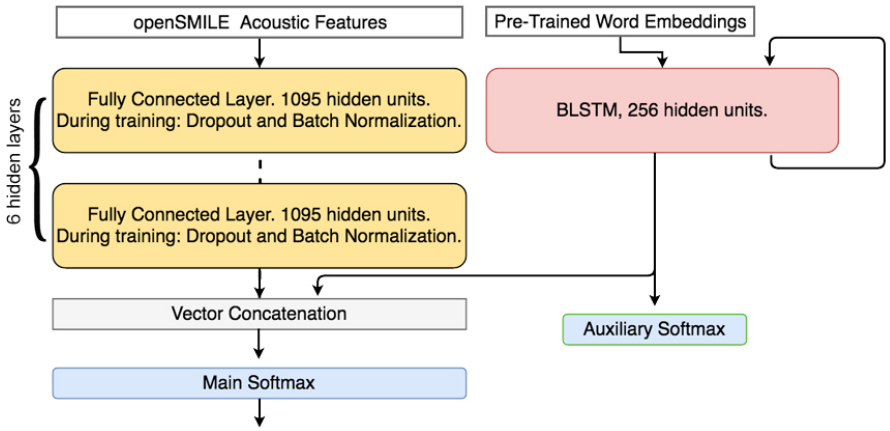

Figure 14.1: Hybrid acoustic lexical model architecture.

14.2 Neural Network Deception Classification

In this section we present the classification results using the three neural network models. We first present the speaker-independent evaluation, which uses speaker-split data, and then we present the speaker dependent evaluation, which uses the random-split data. For both speaker-split and random-split conditions, the data was partitioned into 80% training data and 20% test data. 5% of the training data was used as a validation set for selecting model parameters.

14.2.1 Speaker-Independent Evaluation

This section presents the results of the speaker-independent classification experiments. For each of the three neural network architectures, DNN, LSTM, and hybrid, we trained and evaluated the classifier using the speaker split data, with train and test data from non-

overlapping sets of speakers. The classifiers were trained and evaluated for each of the four segmentation units: IPU, turn, question response, and question chunk. The results are shown in Table 14.1.

Model	*Segmentation*	*P*	*R*	*F*
DNN	IPU	52.55	52.54	52.51
	Turn	54.04	54.00	53.91
	Question Response	58.23	58.23	58.23
	Question Chunk	56.93	56.93	56.93
LSTM	IPU	53.44	53.43	53.40
	Turn	55.54	55.44	55.26
	Question Response	58.81	58.79	58.77
	Question Chunk	59.76	59.48	59.19
Hybrid	IPU	55.43	55.03	54.33
	Turn	54.92	54.50	53.75
	Question Response	59.60	59.43	59.18
	Question Chunk	59.21	59.04	58.83

Table 14.1: Speaker-independent classification results for DNN, LSTM and hybrid neural network classifiers.

The results for the DNN model were almost the same as the results for the statistical machine learning models trained with openSMILE features, reported in Chapter 6. The results here ranged from 52.51 F1 for IPUs to 58.23 for question responses. Similarly, the LSTM results using embeddings were very similar to our previous results from models trained with n-gram features. The LSTM results ranged from 53.4 F1 for IPUs to 59.19 F1 for question chunks. As with other classification experiments, we found that classification performance improved as the size of the segmentation units increased, with question responses and question chunks performing better than IPUs and turns.

The hybrid model achieved the best performance for IPUs (54.33 F1) and question responses (59.18 F1), but not for turns or question chunks. It seems that training the

hybrid model from acoustic-prosodic and lexical feature streams jointly was not a very useful approach for speaker-independent deception classification. Unlike many other speech classification tasks, where large improvements are achieved by using a neural network model instead of a statistical machine learning model, here we did not see large improvements from using neural network models. However, these classifiers were trained using a subset of the features that we explored for deception detection, and it is possible that neural networks trained using additional feature sets (such as syntactic features) would achieve better performance.

14.2.2 Speaker-Dependent Evaluation

This section presents the results of the speaker-dependent classification experiments. For each of the three neural network architectures, DNN, LSTM, and hybrid, we trained and evaluated the classifier using the random split data, with train and test data from the same speakers. The classifiers were trained and evaluated for each of the four segmentation units: IPU, turn, question response, and question chunk. The results are shown in Table 14.2.

Model	*Segmentation*	*P*	*R*	*F*
DNN	IPU	60.59	60.54	60.49
	Turn	62.94	62.65	62.37
	Question Response	63.50	63.50	63.50
	Question Chunk	70.92	70.93	70.93
LSTM	IPU	60.98	60.94	60.89
	Turn	61.26	60.87	60.60
	Question Response	67.44	67.44	67.44
	Question Chunk	68.22	68.23	68.21
Hybrid	IPU	62.93	62.94	62.94
	Turn	62.41	62.41	62.41
	Question Response	71.14	71.14	71.14
	Question Chunk	68.89	68.89	68.89

Table 14.2: Speaker-dependent classification results for DNN, LSTM and hybrid neural network classifiers.

The speaker-dependent classification results were substantially better than the speaker-independent results, for all three neural network models and across the four segmentation units. The DNN trained on openSMILE features produced strong results, ranging from 60.49 F1 for IPUs to 70.93 F1 for question chunks. These results are the best performance obtained using only acoustic-prosodic features. Using statistical machine learning models trained with openSMILE features, we previously obtained F1 scores ranging from 52.03 for IPUs to 56.06 for question chunks (as reported in Chapter 6. It seems that the DNN was able to accurately model speaker-specific patterns of deceptive speech using only acoustic-prosodic features.

The LSTM trained on word embeddings also produced strong results, ranging from 60.89 F1 for IPUs to 68.21 F1 for question chunks. Using statistical machine learning models trained with n-gram features, we previously obtained F1 scores ranging from 53.28 for IPUs to 60.92 for question chunks (as reported in Chapter 6). The LSTM model trained on word embeddings was able to accurately model speaker-specific patterns of deceptive

word usage.

The hybrid model, which combined the DNN and LSTM models trained with openSMILE features and embeddings, resulted in the best performance for all segmentations except for question chunks (which were best classified by the DNN model). It achieved 62.94 F1 for IPUs, 62.41 F1 for turns, 71.14 F1 for question responses, and 68.89 F1 for question chunks. It seems that training the hybrid model from acoustic-prosodic and lexical feature streams jointly was a useful approach for speaker-dependent deception classification.

As with other classification experiments, we found that classification performance improved as the size of the segmentation units increased, with question responses and question chunks performing better than IPUs and turns.

14.3 Discussion

We developed three neural network models for deception classification: a DNN trained on openSMILE features, an LSTM trained on word embeddings, and a hybrid model that combined the DNN and LSTM. These models were motivated by our experimental results with statistical machine learning classifiers reported in Chapter 6, that showed that acoustic-prosodic and lexical features were discriminative between truthful and deceptive speech.

A possible reason for the lack of strong performance of the speaker-independent models is that neural network models require a lot of training data. Although the CXD corpus is relatively large for deception research, the number of training samples varies with the segmentation units, and there only 8,092 question response and question chunk segments. Training with 80% of the data resulted in only 6,473 training instances, which is small for training a neural network model.

IPUs and turns have more segments: 111,428 IPUs (89,142 for training) and 43,673 turns (34,938 for training). However, our previous experiments showed that despite the increased number of training samples, IPUs and turns were more difficult to classify – possibly because they include segments with ambiguous veracity labels.

We found that speaker-dependent models performed strongly for all segmentations. The best results for IPUs (62.9 F1), turns (62.4 F1), and question responses (71.1 F1), were ob-

tained with the hybrid model, and the best result for question chunks (70.9) was obtained with the DNN model. On average, the speaker-dependent models performed 8.8% better than the speaker-independent models. Speaker-independent models performed similarly to statistical machine learning models. The results suggest that the speaker-independent models were not optimized for deception classification, perhaps because of the lack of quantity of the training data for question responses and chunks, or the lack of quality of the training data for IPUs and turns. On the other hand, the speaker-dependent models performed very well, with large improvements in the DNN-openSMILE model over the statistical results using openSMILE features. This suggests that the DNN was able to accurately learn speaker-specific patterns of deceptive speech.

Although this speaker-dependent training paradigm is difficult to replicate in a real-world scenario, and training and evaluating classifiers with data from the same speakers is generally a poor practice, these experiments suggest that this might be a fruitful area of research to pursue. If there are large performance gains from leveraging a small amount of training data from a target speaker, perhaps we should invest in training classifiers that can be easily optimized for a target speaker. This can be useful in a scenario where verifiable language samples of a potential deceiver, such as a politician or other high-profile individual, can be obtained. The hybrid model requires speech features along with embeddings extracted from the transcription of the speech, while DNN model only requires speech samples, without any transcription or annotation. Further research can explore how much training data per speaker is needed to obtain good performance. In addition, experiments can be conducted using "found" data, such as recordings of political speeches, to study the utility of these models on real-world data in the wild.

Chapter 15

Identification of Speaker Traits

In this chapter we present approaches to identifying speaker traits including gender, native language, and personality from short samples of speech. Identifying speaker attributes is useful for many computational applications, including speaker identification and personalization of human-machine interactions. In particular, we are interested in leveraging individual information about a speaker in order to improve deception detection approaches. In Chapter 13 we showed that the speaker traits of gender, native language, and personality can be leveraged, along with acoustic-prosodic and linguistic features, to improve automatic deception detection. Such work is promising, but requires ground-truth knowledge of these speaker traits. For example, it requires NEO-FFI personality scores, which may be impractical to collect in a real-world deception situation.

We address this problem in this chapter. Specifically, we aimed to answer the following question: How much information can be automatically learned from a short dialogue with a subject? We use a portion of the CXD corpus for this study. This part is an initial dialogue between an experimenter and each subject, a 3-4 minute truthful conversation in which the subject answered simple, open-ended questions. There are an average of about 550 words per baseline sample of speech. Using this subset, we extracted acoustic-prosodic and lexical features, and trained classifiers to identify gender, native language (American English or Chinese), and personality. All of this information can be useful for downstream deception detection.

We used three feature sets for the machine learning experiments:

1. Acoustic
2. Lexical
3. Syntactic

Acoustic features include Praat and openSMILE (IS09) feature sets, lexical features are a combination of LIWC and LDI features, and syntactic features include measures of syntactic complexity and part-of-speech (POS) tag ngrams. These feature sets are described in detail in Chapter 4, Section 4.4. A list of the tags and their descriptions is found in Appendix B. These features were extracted from IPUs and features were aggregated per speaker by computing the averaging of each feature across all speaker IPUs in the baseline session. We used n-gram features and word+POS features in our initial set of experiments. However, we found that the results were inflated because of domain-specific n-grams. For example, the token "Barnard" was a very strong indicator that the speaker was female. Therefore, we decided to exclude n-gram and word+POS features for our trait identification experiments.

For the machine learning experiments, we used 10-fold cross validation to train and evaluate the models. Each training example consisted of a feature vector for a single speaker, and each fold contained features from unique speakers. We compared the performance of three classification models: Random Forest (RF), Support Vector Machines (SVM), and Naive Bayes (NB). We used the scikit-learn implementation for these models, and the default parameters.

Some of this work was published in Levitan *et al.* [2016]; An *et al.* [2016], and was done in collaboration with my co-authors.

15.1 Gender Identification

The problem of gender identification was framed as a binary classification problem: given a feature vector extracted from a speaker's baseline speech sample, can we determine whether the speaker is male or female? We used the self-identified gender labels provided by each participant in the demographic survey at the start of the experiment. Table 15.1 shows the gender classification performance, measured by accuracy, precision, recall, and F1-score.

The baseline performance, obtained by always predicting the majority class (Female) is 54.41% accuracy.

Feature	*CLF*	*A*	*P*	*R*	*F1*
Acoustic	SVM	95.88	95.91	95.85	95.85
Lexical	NB	66.47	68.32	67.52	66.28
Syntactic	NB	69.71	72.07	70.75	69.26
Lexical+Syntactic	NB	71.47	72.8	72.27	71.37
All	NB	95.29	95.3	95.31	95.26
Majority Baseline	-	54.41	27.21	50	35.24

Table 15.1: Gender classification with combined feature sets. (SVM=Support Vector Machine, NB=Naive Bayes)

Intuitively, the acoustic-prosodic features were highly predictive of gender, with an SVM classifier achieving 95.88% accuracy. It is interesting that the text-based lexical and syntactic feature sets were also somewhat predictive of gender. A Naive Bayes classifier trained with a combination of lexical and syntactic features achieved 71.47% accuracy, about 17% better than the baseline performance. This was despite the fact that all subjects answered almost the same questions in the baseline session.

Having demonstrated that acoustic-prosodic, lexical, and syntactic features are highly effective at gender classification, we were interested in analyzing which features were most useful at discriminating between male and female speakers. For each of three main feature groups – acoustic-prosodic, lexical, and syntactic, we ranked the features using the SelectKBest function in scikit-learn. We used a score function which scores features using the ANOVA F-value between the class label and each feature. Below we show the top 20 features and their F-values for each group of features. The top 20 ranked acoustic-prosodic features are shown in Figure 15.1.

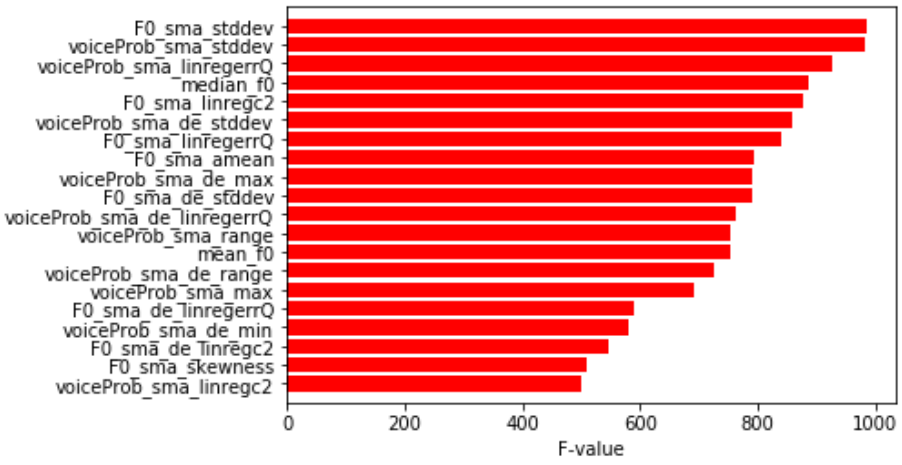

Figure 15.1: Top 20 acoustic features for gender classification, ranked by ANOVA F-values.

There were two Praat features (median-f0 and mean-f0) in the top 20 acoustic-prosodic features, and the rest were from the openSMILE feature set. Interestingly, 10 of the top features were functionals computed over the probability of voicing, which indicates how close the signal is to an ideal harmonic signal (high probability) or to a noise-like signal (low probability).

All top 20 acoustic features were significantly different with $p < 0.05$ (after FDR correction for multiple comparisons). However, an SVM trained with only the top single feature – F0-sma-stddev – yielded an accuracy of 93.16%. This feature alone was highly discriminative between male and female speakers: the mean value was 29.35 for male speakers, and 80.70 for female speakers. Voice probability features were also significantly higher on average for female speakers than for male speakers.

A more challenging problem than gender identification from acoustic-prosodic features is gender identification from text-based features. The top 20 ranked lexical features are shown in Figure 15.2.

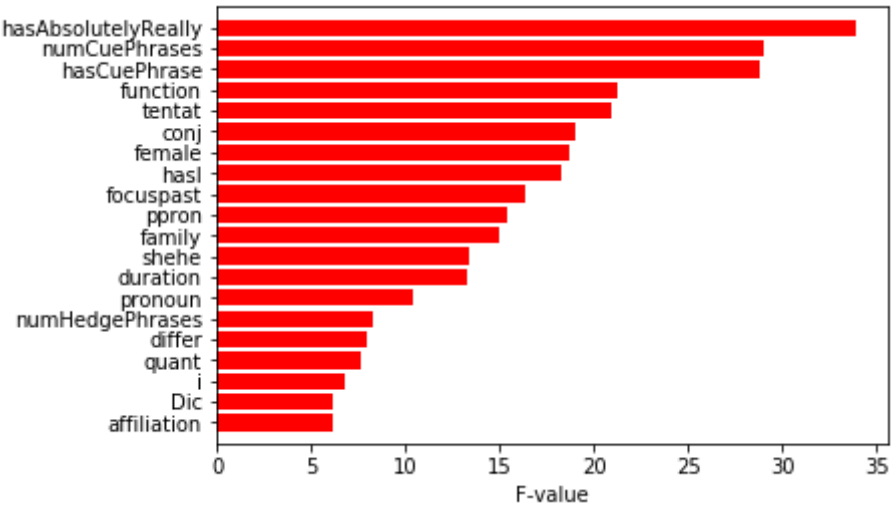

Figure 15.2: Top 20 lexical features for gender classification, ranked by ANOVA F-values.

Five of the top 20 lexical features were from the LDI feature set (hasAbsolutelyReally, numCuePhrases, hasCuePhrase, hasI and numHedgePhrases), and the rest were from the LIWC feature set. Female speakers tended to use "absolutely" and "really" more frequently than male speakers, used more first person singular pronouns (e.g. I, me, my), and also used more cue phrases and hedge phrases. Intuitively, the LIWC dimension of "female," which captures references to females (e.g. girl, her, mom) was more frequent in female language. All top 20 lexical features were significantly different with $p < 0.05$ (after FDR correction for multiple comparisons). A Naive Bayes classifier trained on only lexical features achieved an accuracy of 66.47%.

We also examined the top 20 syntactic features, shown in Figure 15.3.

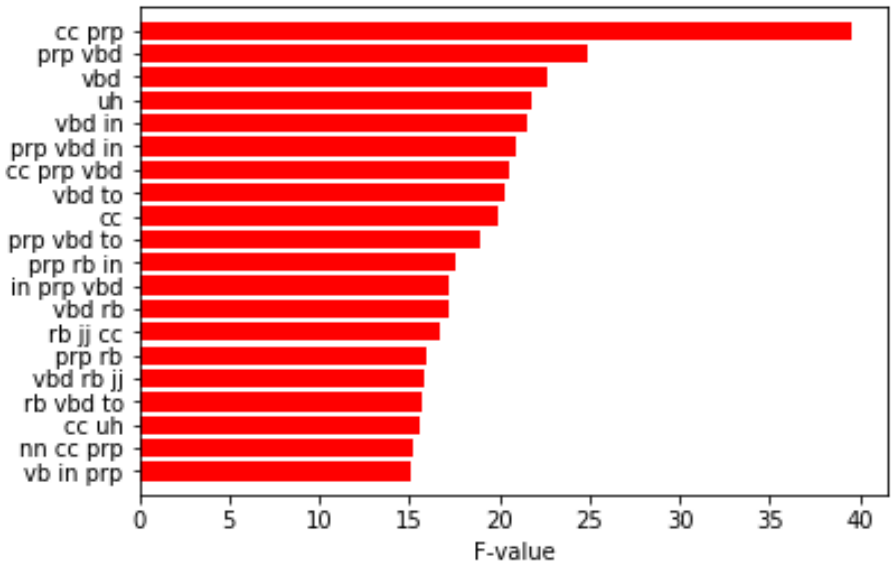

Figure 15.3: Top 20 syntactic features for gender classification, ranked by ANOVA F-values.

Interestingly, the top ranked syntactic feature was "cc prp" – a coordinating conjunction (e.g. and, but, not) followed by a personal pronoun (e.g. I, we, they). This syntactic pattern was used significantly more frequently by female speakers than male speakers. Past tense verbs ("vbd") were also used more frequently by female speakers, while interjections ("uh") were used more frequently by male speakers. 17 of the top 20 complexity features were significantly different with $p < 0.05$ (after FDR correction for multiple comparisons). A Naive Bayes classifier trained on only syntactic features achieved an accuracy of 69.71% The best performance of 71.47% accuracy was achieved using a combination of syntactic and lexical features.

As expected, acoustic-prosodic features were very predictive of speaker gender. More surprisingly, we were able to automatically identify speaker gender using a combination of syntactic and lexical features extracted from short samples of transcribed speech. This suggests that not only are there acoustic-prosodic markers of gender, but there are significant differences in syntactic and lexical patterns across gender, which we can leverage to classify gender from transcribed speech.

15.2 Native Language Identification

Having successfully classified gender from short samples of speech, we used the same feature sets to classify native language. The problem of native language identification was framed as a binary classification problem: given a feature vector extracted from a speaker's baseline speech sample, can we determine whether the speaker is a native English speaker or a native speaker of Mandarin Chinese? Nativeness was determined from the language background survey that each participant filled out at the start of the experiment.

Table 15.1 shows the gender classification performance, measured by accuracy, precision, recall, and F1-score. The baseline performance, obtained by always predicting the majority class (Female) is 54.12% accuracy.

Feature	*CLF*	*A*	*P*	*R*	*F1*
Acoustic	RF	74.99	75.21	74.39	74.49
Lexical	RF	85.29	86.15	85.03	85.02
Syntactic	RF	86.16	86.78	85.7	85.93
Lexical+Syntactic	RF	87.05	87.47	86.73	86.86
All	RF	87.04	87.65	86.71	86.86
Majority Baseline	-	54.12	27.06	50	35.11

Table 15.2: Native language classification with combined feature sets. (RF=Random Forest)

Random Forest was the best classification algorithm for native language identification. An RF classifier trained on acoustic-prosodic features achieved an accuracy of 74.99%, over 20% better than the majority class baseline. Unlike the results for gender classification, here we found that text-based features performed better than acoustic-prosodic features. A classifier trained on lexical features achieved an accuracy of 85.29%, and a classifier trained on syntactic features resulted in 86.16% accuracy. The best performance of 87.05% accuracy was obtained using a combination of lexical and syntactic feature sets.

Next, we examined which features were most useful at discriminating between native English and native Chinese speakers. For each of three main feature groups – acoustic-prosodic, lexical, and syntactic, we ranked the features using the SelectKBest function in

scikit-learn. We used a score function which scores features using the ANOVA F-value between the class label and each feature. Below we show the top 20 features and their F-values for each group of features.

The top 20 ranked acoustic-prosodic features are shown in Figure 15.4.

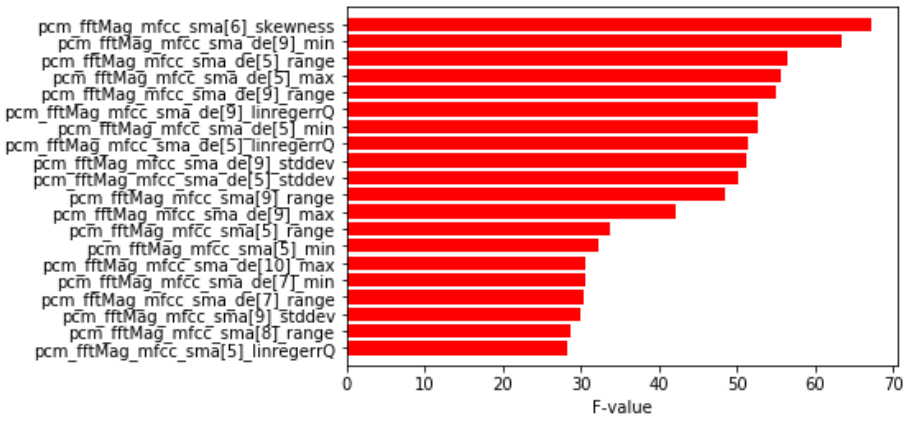

Figure 15.4: Top 20 acoustic features for lang classification, ranked by ANOVA F-values.

All top 20 acoustic-prosodic features were MFCC features from the openSMILE feature set. These top 20 acoustic features were significantly different with $p < 0.05$ (after FDR correction for multiple comparisons). It seems that MFCC features, which are commonly used for speech recognition and speaker recognition, are useful for distinguishing between native speakers of English and native speakers of Chinese.

Next, we examined text-based features that were predictive of native language. The top 20 ranked lexical features are shown in Figure 15.5.

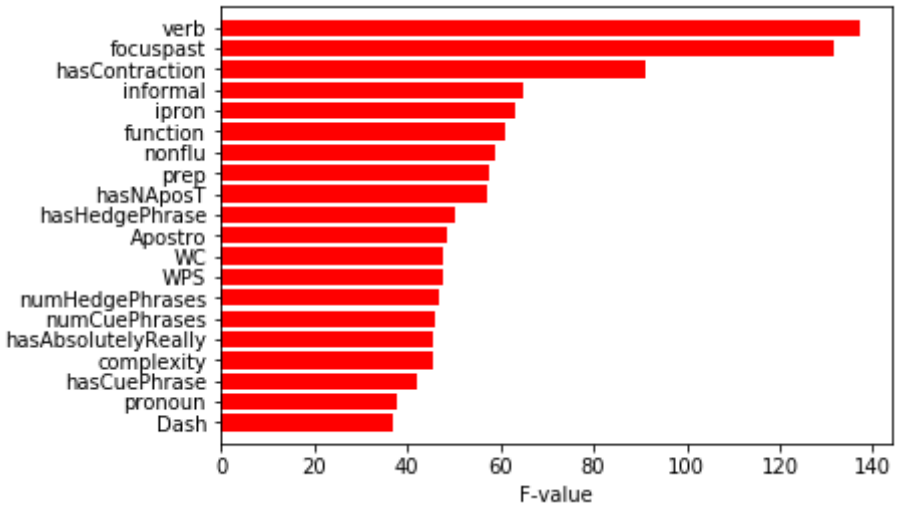

Figure 15.5: Top 20 lexical features for native language classification, ranked by ANOVA F-values.

Eight of the top 20 lexical features were from the LDI feature set (e.g. hasContraction, hasHedgePhrase), and the rest were from the LIWC feature set. The top ranked feature was verb usage: native Chinese speakers use fewer verbs on average than native English speakers. In particular, past tense verbs, captured by the LIWC dimension "focuspast," were used significantly more frequently by native English speakers. Another useful feature for native language identification was "hasContraction" – native English speakers were much more likely to use contractions in their baseline speech than native Chinese speakers. All top 20 lexical features were significantly different with $p < 0.05$ (after FDR correction for multiple comparisons). An RF classifier trained on only lexical features achieved an accuracy of 85.29%

We also examined the top 20 syntactic features, shown in Figure 15.6.

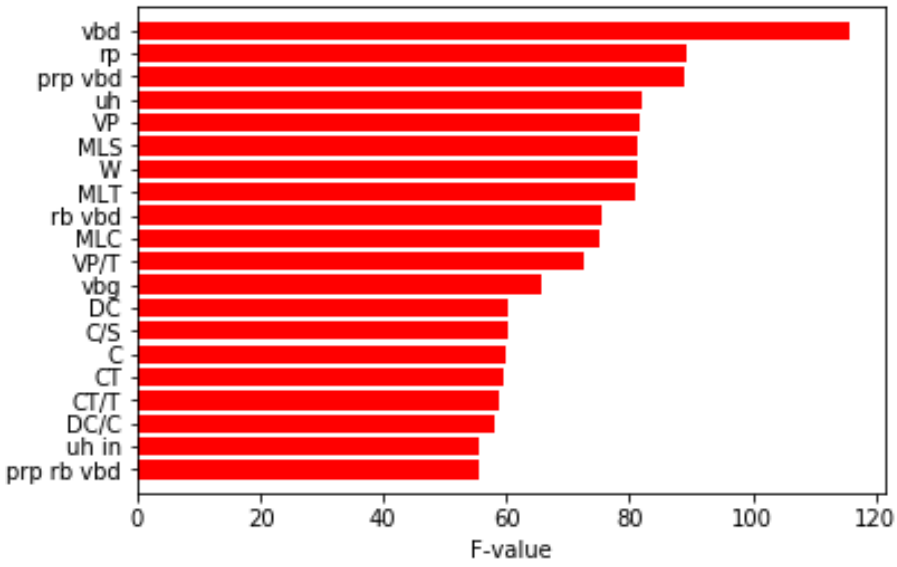

Figure 15.6: Top 20 syntactic features for native language classification, ranked by ANOVA F-values.

11 of the 20 features came from the syntactic complexity feature set (e.g. MLS–mean length of sentence, C/S–clauses per sentence), and the remaining 9 features were part of speech ngram features. The complexity features are described in detail in Chapter 4, Section 4.4. Consistent with our analysis of lexical features, the top ranked syntactic feature was "vbd" – the part of speech representing past tense verbs. Interestingly, particle usage ("rp") was higher for native speakers of English. All 11 measures of syntactic complexity that were ranked in the top 20 features were increased for native speakers of English. All of the top 20 complexity features were significantly different with $p < 0.05$ (after FDR correction for multiple comparisons). An RF classifier trained on only syntactic features achieved 86.16% accuracy. It seems that these syntactic features were highly predictive of native language. The best native language classification results – 87.05% accuracy – were achieved using a combination of lexical and syntactic features.

Our analysis of useful features for native language identification highlighted interesting trends. We found that MFCC features were the most useful acoustic-prosodic feature set. Verb usage, contractions, particles, and several measures of syntactic complexity were the most useful text-based features. We were able to train Random Forest classifiers to leverage

these differences and distinguish between native speakers of English and native speakers of Chinese with high accuracy. In future work, it will be interesting to test whether these differences that we observed between native speakers of English and native speakers of Chinese hold true for other L2 speakers of English, who are not native Chinese speakers.

15.3 Personality Identification

In this final set of trait identification experiments, we aimed to automatically identify speaker personality traits from short samples of speech. Unlike gender and native language classification, which we modeled as binary classification tasks, personality cannot be easily modeled as a binary classification problem. Personality labels in the CXD corpus were defined using the NEO-FFI personality inventory, which was administered to each participant at the beginning of each experimental session. A psychologist scored the personality tests, giving each participant five numeric scores, one for each of the Big Five personality dimensions: Neuroticism (N), Extroversion (E), Openness to Experience (O), Agreeableness (A), and Conscientiousness (C). The NEO scores are on a continuous scale for each of the five dimensions.

As described previously in Chapter 12, we partitioned speakers into personality groups by binning the numeric personality scores to "high," "average," or "low" for each dimension. The thresholds for each bin were obtained from a prior study of population norms from a large sample of administered NEO-FFI, and are different for males and females Locke [2015]. Table 12.1 shows the mapping of numeric NEO scores to the three categorical labels. As expected, the personality bins are highly unbalanced. Table 12.2 shows the distribution of participants in the high, average, and low personality bins for each of the 5 NEO dimensions. We framed the personality identification task as a 3-way classification problem for each personality dimension. That is, we aimed to identify whether a speaker scored high, average, or low for each personality trait. Because of the unbalanced distribution of personality bins, we evaluated the performance of our classifiers using average F1 across the three classes.

Table 15.3 shows the classifier performance for personality classification, measured by average F1-score. The baseline performance was obtained by always predicting the majority

class. The majority class was "High" the dimensions of N, E, and O; "Average" for the A dimension; and "Low" for the C dimension.

Feature	*CLF*	*N*	*E*	*O*	*A*	*C*
Acoustic	SVM	34.43	39.01	35.21	37.06	34.42
Lexical	SVM	35.06	34.25	43.64	38.74	34.36
Syntactic	NB	50.62	78.32	**52.14**	70.80	64.96
Lexical+Syntactic	NB	**56.84**	78.51	40.86	**73.38**	**69.45**
All	NB	32.61	**78.69**	43.60	63.95	63.95
Majority Baseline	-	22.66	18.64	23.24	19.93	20.11
Improvement	-	34.18	60.05	28.90	53.45	49.34

Table 15.3: Personality bin classification with combined feature sets. (SVM=Support Vector Machine, NB=Naive Bayes)

As shown in Table 15.3, SVM models performed best using acoustic and lexical feature sets, and the NB models performed best using the syntactic, lexical+syntactic, and all features combined. For each of the five factors, our classifiers obtained performance well above the majority baseline. The best performance was achieved for Extroversion – a NB model trained with acoustic, lexical, and syntactic features combined achieved an F1-score of 78.69, an improvement of 60.05 over the majority baseline. Agreeableness classification also performed very strongly – a NB classifier trained on a combination of lexical and syntactic features achieved an F1-score of 73.38, an improvement of 53.45 over the baseline. Classification of Conscientiousness also achieved strong performance. A NB classifier trained on lexical and syntactic features achieved an F1-score of 69.45, an improvement of 49.34 over the majority baseline. We achieved more moderate improvements for classification of Neuroticism and Openness to Experience. The best classifier for Neuroticism identification was a NB classifier trained with lexical and syntactic features, which achieved an F1-score of 56.84, 34.18 points above the baseline. Openness to Experience was the most difficult to classify. The best performance of 52.14 F1 was obtained with a NB classifier trained on syntactic features. This result was 28.9 points above the baseline.

Overall, we observed that text-based features were much more effective for personality identification than acoustic-prosodic features. In particular, the best individual feature set was syntactic. However, the SVM classifiers trained using acoustic features all achieved performance above the baseline.

Next, we examined which features were most useful at discriminating between speakers who scored high, average, or low for each personality dimension. For each personality trait, we ranked the features of the best performing classifier using the SelectKBest function in scikit-learn. The score function scores features using the ANOVA F-value between the class label and each feature. In the figures below, we show the top 20 features and their F-values for the acoustic+lexical+syntactic feature set for each of the five trait classification tasks.

Figure 15.7 shows the top 20 acoustic+lexical+syntactic features for classification of Neuroticism.

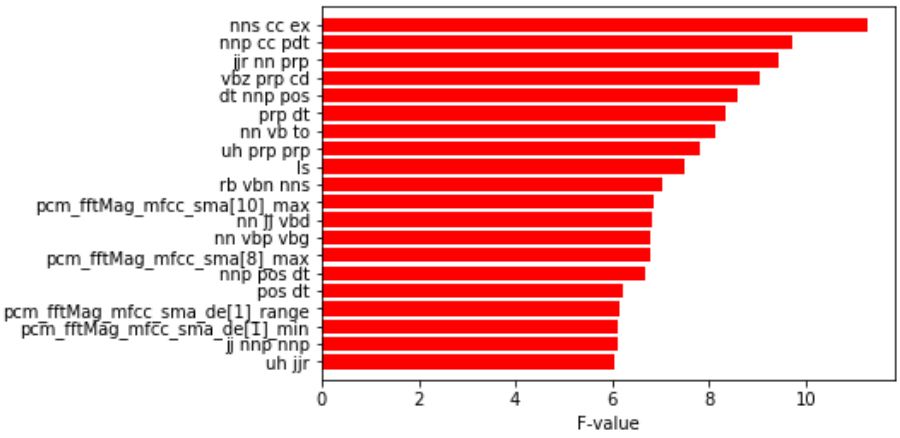

Figure 15.7: Top 20 acoustic+lexical+syntactic features for Neuroticism classification, ranked by ANOVA F-values.

16 of the top 20 features came from the syntactic feature set, and specifically the POS tag n-grams. The remaining four features were from the acoustic-prosodic feature set. The top ranked feature was "nns cc ex" – the POS tag trigram of a plural noun followed by a coordinating conjunction and then an existential there. "nnp cc pdt" was also highly ranked, and it represents a proper noun followed by a coordinating conjunction and then a predeterminer. The third ranked feature, "jjr nn prp," indicates an adjective followed

by a singular noun and then a preposition. All three sequences of POS tags were most frequently used in individuals who scored low on Neuroticism. A trend that we observed is that plural nouns ("nns") and proper singular nouns ("nnp") appear in six of the top features, all of which appeared more frequently in the speech of individuals who scored low on Neuroticism. These findings support previous work by Gill [2003], which found that individuals who scored low on Neuroticism tended to use more plural nouns and proper singular nouns than individuals who scored high on Neuroticism.

All of the top 20 complexity features were significantly different with $p < 0.05$ (after FDR correction for multiple comparisons). An SVM classifier trained on lexical+syntactic features achieved an F1-score of 56.84. Syntactic features were the most useful for Neuroticism classification.

Figure 15.8 shows the top 20 acoustic+lexical+syntactic features for classification of Extroversion.

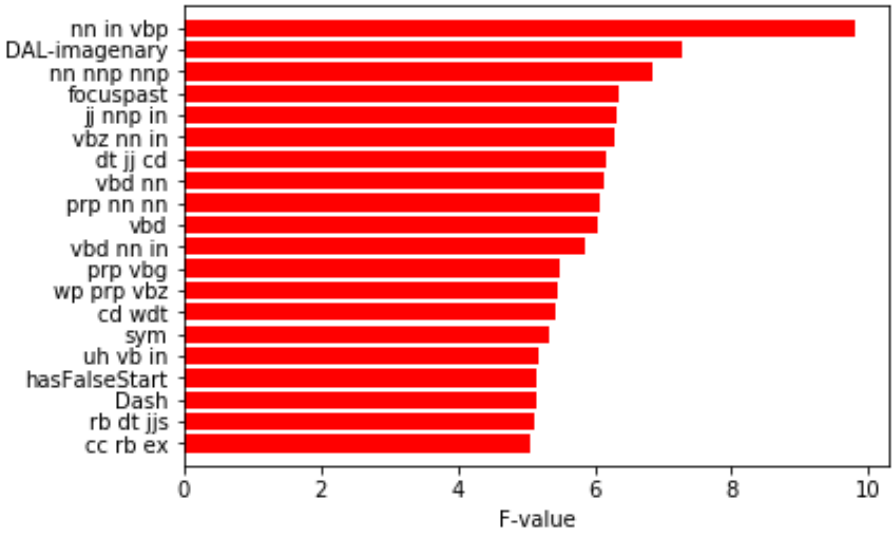

Figure 15.8: Top 20 acoustic+lexical+syntactic features for Extroversion classification, ranked by ANOVA F-values.

16 of the top 20 features came from the syntactic feature set, and the remaining four features were from the lexical feature set. The top ranked feature was "nns in vbg" – the POS tag trigram of a singular noun followed by a preposition and then a present tense verb.

This formulation was used more frequently by individuals who scored low for Extroversion. "DAL-imagery" was another highly ranked feature. It captures words that are used to create vivid descriptions. Individuals who were highly extroverted used these words more frequently than those who were introverted. "focuspast" is a LIWC category that captures past tense verbs. This feature was most frequent in individuals who were in the average Extroversion bin, followed by those who were in the high Extroversion bin, and it was the least frequently used by individuals who scored low on Extroversion. Interestingly, "hasFalseStart" and "Dash" (which was used by transcribers to indicate false starts) were most frequent in speech of highly Extroverted individuals. False starts are a type of speech disfluency where the speaker begins an utterance and then stops it prematurely. This sometimes occurs when the speaker changes their mind about what they are saying. Another trend in the feature analysis is that verb usage seems to be important for Extroversion identification; various verb forms appear in 9 of the top features.

Extroversion classification using all of the features achieved an F1-score of 78.69, which was an improvement of 60.05 over the majority class baseline. This was the "easiest" trait to predict in our classification experiments, and suggests that there are salient lexical and syntactic markers of Extroversion that are present in spontaneous speech.

Figure 15.9 shows the top 20 acoustic+lexical+syntactic features for classification of Openness.

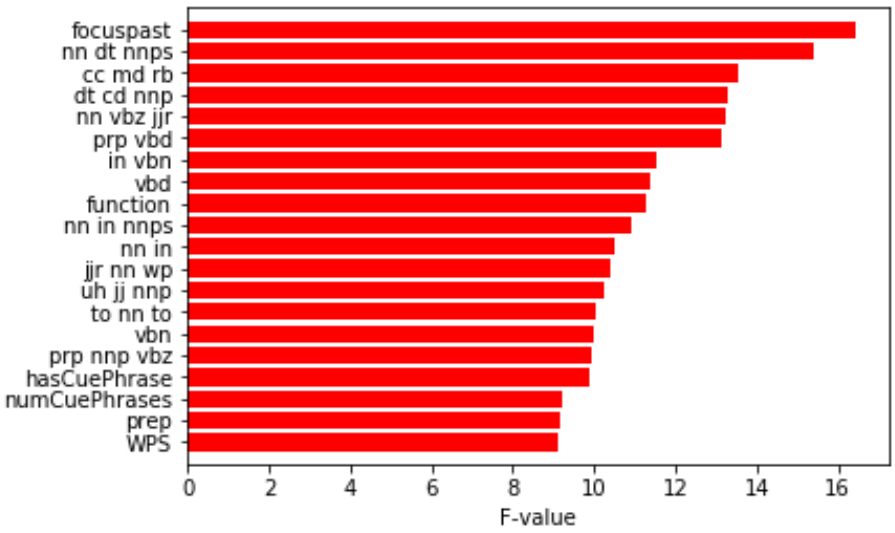

Figure 15.9: Top 20 acoustic+lexical+syntactic features for Openness classification, ranked by ANOVA F-values.

15 of the top 20 features came from the syntactic feature set, and the remaining five features were from the lexical feature set. The LIWC dimension "focuspast" was the top ranked feature for Openness, and words in this dimension were most frequently used by speakers who scored high for Openness. Function words and cue phrases were used most frequently by speakers who scored high for Openness. "WPS" (words per sentence) was significantly higher for speakers in the high bin for Openness. Prepositions ("prep") were also used most frequently be speakers in the high Openness bin. For syntactic features, past tense verbs ("vbd") and past participle verbs ("vbn") were most frequently used by speakers who were high on the Openness scale. The n-grams containing these verbs (e.g. "prp vbd") were also most frequently used by individuals who scored high for Openness. N-grams containing proper nouns ("nnp" and "nnps") were most frequently used by speakers who were low in Openness.

We found that Openness to Experience was the most difficult trait to classify; the best result of 52.14 F1-score was obtained using a Naive Bayes classifier trained on syntactic features. Of the five personality dimensions, the distribution of subjects was the most skewed for Openness. Only 6% of subjects scored low for Openness, 42% were average, and

52% were high. Perhaps there were fewer differences between the speaking styles of subjects in average vs. high bins, making it more difficult to classify subjects in this dimension.

Figure 15.10 shows the top 20 acoustic+lexical+syntactic features for classification of Agreeableness.

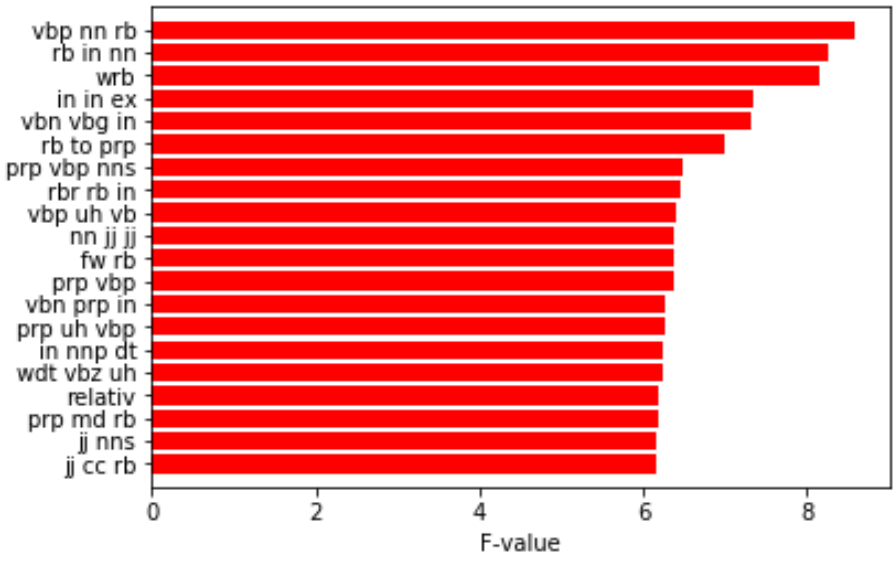

Figure 15.10: Top 20 acoustic+lexical+syntactic features for Agreeableness classification, ranked by ANOVA F-values.

19 of the top 20 ranked features were POS n-grams. The n-grams that contained interjections ("uh") were used more frequently by individuals who scored high for Agreeableness. N-grams that contained prepositions ("in") were also used more frequently by highly agreeable speakers. N-grams that contained personal pronouns ("prp") and adverbs ("rb"), such as "rb to prp" and "rb to prp" were most frequent in speakers with low Agreeableness. Adjectives ("jj") appeared in n-grams that were most frequently used by speakers with high Agreeableness (e.g. "jj nns"). The best performance for Agreeableness classification was 73.38 F1-score; it was achieved using a Naive Bayes classifier trained with a combination of lexical and syntactic features. It seems that there are strong linguistic markers of Agreeableness.

Figure 15.11 shows the top 20 acoustic+lexical+syntactic features for classification of Conscientiousness.

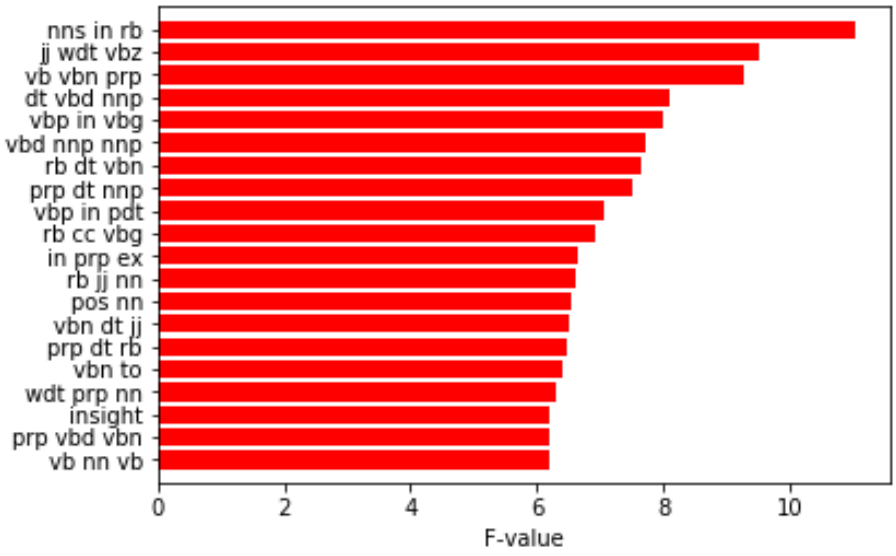

Figure 15.11: Top 20 acoustic+lexical+syntactic features for Conscientiousness classification, ranked by ANOVA F-values.

19 of the top 20 ranked features were from the POS n-gram feature set. Only one POS n-gram ("vbn to") was most frequent for speakers in the low bin for Conscientiousness, and two POS n-grams were most frequently used by speakers who were in the average bin. All of the other POS n-grams were used most frequently by speakers who scored high for Conscientiousness. The best performance for Conscientiousness classification was obtained using a Naive Bayes classifier trained on a combination of lexical and syntactic features. This model achieved an F1-score of 69.45, which was an improvement of 49.34 above the majority baseline.

It is difficult to draw strong conclusions from the syntactic feature analysis for personality trait identification. First, the dependency parses from which the syntactic features were derived were noisy, due to the nature of the corpus (transcribed speech, including non-native speakers, no capitalization or punctuation). In addition, we excluded word n-grams because we found that there were many corpus-specific tokens that would not generalize to other domains. However, it is also possible that some of the POS n-grams that were useful for personality trait classification captured specific patterns for this corpus. Many of the POS n-grams were very sparse. It remains to be seen whether these same syntactic patterns

are predictive of personality traits in other corpora.

There are also some limitations of this paradigm for personality trait classification. In our experiments we treated each personality dimension as independent, and attempted to classify each speaker as high, average, or low for each trait. This independence assumption is questionable – it is intuitive that a speakers personality traits are related to each other. In our ongoing work we have explored identifying clusters of personality traits and then classifying speakers into personality clusters. The high, average, and low bins that were used for this analysis were defined using thresholds from a large and diverse population. It is possible that these thresholds were not a good fit for the population studied in the CXD corpus, which was mostly college students.

An open question in personality identification is how to define ground truth personality labels. This work used self-identified personality labels derived from the NEO-FFI personality test taken by each subject. Others have used observer-identified personality labels, by having people annotate speech or language samples for perceived personality traits. Mairesse *et al.* [2007] compared personality classification results for self-reported vs. observer-reported personality traits, and found that they were able to accurately identify observer labeled personality scores, but the results were much lower for self-labeled scores. Although there are inherent biases when a person assesses their own personality traits, self-reported personality labels are likely more representative of an individual's personality than observer-reported labels, since it is questionable whether personality can be accurately labeled by others, especially strangers. Modeling personality is a difficult problem, and how it is modeled has important ramifications for automatic personality identification.

15.3.1 Discussion

This chapter aimed to answer the question: *How much information can be automatically learned from a short dialogue with a subject?* In this chapter we presented the results of several speaker trait classification experiments. These experiments aimed to automatically identify the gender, native language, and personality of a speaker, using a short sample of speech. The data used was the initial baseline interview that was conducted with each subject, where subjects were instructed to answer truthfully to the questions. There was

3-4 minutes of subject speech collected per speaker in the baseline session.

We obtained strong gender classification performance. As expected, we achieved as high as 95.88% accuracy using acoustic-prosodic features. We also obtained strong performance using only text-based features derived from the transcribed speech. A Naive Bayes classifier trained using a combination of lexical and syntactic features achieved 71.47% accuracy – well above a majority class baseline of 54.41% accuracy. In addition to the gender classification experiments, we identified the best acoustic, lexical, and syntactic features for distinguishing between male and female speakers.

We presented classification results for native language identification – specifically, distinguishing between native speakers of Standard American English (SAE) and native speakers of Mandarin Chinese (MC). The best performance of 87.05% was achieved using a Random Forest classifier trained with a combination of lexical and syntactic features. We also trained a speech-based classifier that achieved an accuracy of 74.99% using only acoustic-prosodic features. In addition to the classification experiments, we identified the best acoustic, lexical, and syntactic features for distinguishing between native speakers of SAE and MC. For example, use of contractions was an indicator of SAE speakers. Further experiments are needed to determine whether these differences are specific to non-native speakers of SAE who are native speakers of MC, or whether they generalize to all non-native speakers of SAE.

Finally, we presented classification results for personality trait identification. We modeled this task as five independent 3three-way classification tasks, where we classified each speaker as falling into the high, average, or low bin for each of the Big Five personality traits. We obtained results well above a majority class baseline for all five personality traits: N-score 56.84 (+34.18 from baseline), E-score 78.69 (+60.05), O-score 52.14 (+28.9), A-score 73.38 (+53.45), and C-score 69.45 (+49.34). We also analyzed the features that discriminated between the high, average, and low bins for each trait. Finally, we discussed limitations of this approach of modeling personality and ways to overcome these limitations.

Although these experiments were conducted for the purpose of providing speaker trait information for deception detection, this work has implications beyond deception detection. For example, speaker trait identification can be very useful for speech analytics and person-

alization of human-machine interactions. Our results show that gender, native language, and to some degree, personality, can be inferred from a short sample of speech. Our feature analysis provides insight into the acoustic-prosodic, lexical, and syntactic characteristics that help distinguish between groups of speakers, and can help further research in speaker trait identification.

Chapter 16

Conclusions and Future Work

Part II of this thesis provides a comprehensive framework for identifying individual differences in deceptive speech and leveraging those differences for classification of deceptive speech. Most previous research on deceptive communication has identified cues to deception across all speakers. Some previous studies have observed individual differences in how people lie, but there have not been significant efforts to empirically identify these differences, understand the factors that affect these differences, and leverage these differences for automatic deception detection.

Using the CXD corpus, which is annotated with speaker traits, we carefully analyzed differences in cues to deception across gender, native language, and personality type. We compared several approaches to leverage speaker differences in deception classification, including speaker-dependent neural network models. We also trained models to automatically identify speaker gender, native language, and personality from short samples of speech, with the goal of using this information for downstream deception detection.

We systematically analyzed over 150 acoustic-prosodic, lexical, and syntactic cues to deception and truth, and identified many differences between male and female speakers, between native speakers of Standard American English (SAE) and Mandarin Chinese (MC), and between speakers who scored high, average, or low for each of the Big Five personality traits. In some cases, we found that previously identified cues to deception across all speakers were not present when we examined particular groups of speakers. In other cases, we discovered new cues to deception for groups of speakers with shared traits, which were

not present when we analyzed all speakers. These findings suggest that gender, native language, and personality all play a role in how people produce deceptive speech. This work is the first comprehensive analysis of gender, native language, and personality differences in acoustic-prosodic and linguistic cues to deception, and is an important contribution of this thesis.

We compared three approaches to leverage speaker-dependent information in deception classification: adding traits as features, training models using homogenous data, and using speaker-dependent features. The largest improvements were obtained from adding speaker-dependent features. These features were computed by subtracting baseline features, where subjects spoke truthfully, from interview session features, in order to capture deviations from their baseline speaking behavior. Practitioners have advocated for interviewing practices that establish baseline behavior of subjects while telling the truth, and then looking for differences from the baseline to detect deception. Baseline behavior is often elicited by first asking neutral questions that the subject is expected to answer truthfully. In this work we operationalized a method to automatically capture deviations from the baseline, instead of relying on human judgment to determine deviation from the baseline. Future work can explore modeling speaker traits in additional ways. For example, there has been promising work modeling personality with deception in a multi-task learning framework [An *et al.*, 2018]; this idea can be extended to learn gender and native language as well.

We developed three neural network models for deception classification: a DNN trained on openSMILE features, an LSTM trained on word embeddings, and a hybrid model that combined the DNN and LSTM. We found that these models performed similarly to the statistical models when trained and evaluated on distinct speaker sets, but were able to accurately model speaker-dependent patterns of deceptive behavior. These results suggest that under conditions where training data can be obtained for a target speaker, neural network models can be used to achieve strong speaker-dependent deception detection performance. Further research can explore how much training data per speaker is needed to obtain good performance. In addition, experiments can be conducted using "found" data, such as recordings of political speeches, to study the utility of these models on real-world data in the wild.

We conducted a series of speaker trait classification experiments, aimed at automatically identifying gender, native language, and personality traits from a short sample of speech. Our results show that gender, native language, and to some degree, personality, can be inferred from a short sample of speech. We also conducted feature ranking analyses, providing insight into the acoustic-prosodic, lexical, and syntactic characteristics that help distinguish between groups of speakers. These can help further research in speaker trait identification. An area for further research is modeling the five personality traits jointly instead of treating each trait independently. Although these experiments were conducted for the purpose of providing speaker trait information for deception detection, this work has implications beyond deception detection. For example, speaker trait identification can be very useful for speech analytics and personalization of human-machine interactions.

Part II of this thesis provides a framework for identifying speaker differences in cues to deception, and explores ways to leverage speaker differences in deception classification. Hopefully this work will lay the groundwork for continued research on individual differences in deceptive speech, which will lead to further improvements of automatic deception detection.

Part III

Conclusions

Chapter 17

Conclusions

Despite much research, deception remains a problem that is not well understood. Human performance at deception detection is about chance level, and current deception detection technologies are not much better. A challenging problem in deception research is that different people exhibit different cues when lying. In order to develop technologies that can accurately identify deception, we need a better understanding of deceptive communication. Furthermore, it is important to study the individual and cultural factors that affect deception production and perception.

In this thesis, we presented a comprehensive framework for studying deceptive communication and developing automated technologies for deception detection. In addition, we presented a study of individual differences in cues to deception, with methods to leverage individual differences for automatic deception detection.

This thesis contains the following six major contributions:

- **A large-scale corpus of deceptive speech.** We created the Columbia X-Cultural Deception (CXD) Corpus, with over 122 hours of subject speech. This corpus enabled studies of deceptive speech on a scale that was not previously possible. The cross-cultural nature of the corpus and the personality trait information that was collected enabled a study of individual differences in deceptive speech.

- **Acoustic-prosodic and linguistic cues to deception.** Our systematic analysis of over 150 speech- and text-based features in a large-scale corpus of deceptive speech

identified many significant differences between truthful and deceptive responses. This furthers our scientific understanding of deceptive language.

- **Automatic deception classification.** We trained classifiers to automatically identify deceptive speech using a variety of acoustic-prosodic and linguistic features, for four segmentation units. Our best classifier was a Naive Bayes classifier trained with a combination of lexical and syntactic features extracted from question chunks, and achieved an accuracy of almost 70% – well above human performance of 56.75%. In addition to the contribution of strong performing deception classifiers, our work provides useful insights for future experiments with automatic language-based deception detection.

- **A study of entrainment in deceptive dialogue.**Our study of acoustic-prosodic and lexical entrainment in the CXD corpus is, to our knowledge, the first to investigate entrainment in those dimensions in deceptive dialogues. We found evidence of global and local entrainment in deceptive speech, and some differences in entrainment between truthful and deceptive speech. This motivates modeling entrainment behavior in future work on automatic deception detection.

- **Individual differences in cues to deception.** We present the first comprehensive analysis of gender, native language, and personality differences in acoustic-prosodic and linguistic cues to deception. This work identified many differences in cues to deception between male and female speakers, between native speakers of Standard American English (SAE) and Mandarin Chinese (MC), and between speakers who scored high, average, or low for each of the Big Five personality traits. These findings suggest that gender, native language, and personality all play a role in how people produce deceptive speech.

- **Deception classification leveraging speaker differences.** We introduced speaker-dependent features that capture a speaker's deviation from their natural speaking style, in order to improve deception classification. We also developed neural network models that accurately modeled speaker-specific patterns of deceptive speech.

These features and models are novel approaches for modeling individual differences in deceptive speech.

17.1 Future Work

Throughout the thesis we discussed suggestions for future work. Here we describe future research directions that arise from this thesis.

- **Real-world data.** All of the experiments in this thesis were conducted using the CXD corpus. An important next step is to evaluate the classifiers on real-world deception, which can be substantially different from deception produced in a laboratory environment. Aside from the problem of data quality (e.g. poor audio recording conditions), real-world deception is often high-stakes, and therefore the cues to deception might differ from low stakes deception in a lab environment.

- **Dialogue features.** This work, along with almost all other studies of deception, focused on the speech produced by the deceiver. However, as our study of entrainment in deceptive speech suggests, it might be useful to also consider the speech produced by the interlocutor. The CXD corpus is unique in that it includes both the interviewer and interviewee channels. Future work should explore deception classification using features from both dialogue partners, such as acoustic-prosodic entrainment measures, or measures of linguistic similarity between interlocutors.

- **Trustworthy speech.** This thesis focused on identifying verbal indicators of deceptive speech. A less-studied, complementary phenomenon, is the task of identifying verbal indicators of trust. Trust is a fundamental component of human communication, and understanding the characteristics of trustworthy speech is useful for improving human-computer interactions. The framework that was introduced in this thesis for studying deceptive speech and individual differences can be applied to the study of trustworthy speech. The CXD corpus is well-suited for the study of trustworthy speech, as it includes interviewer judgments of deception which can be used as trust annotations.

17.2 Epilogue

In Part I of this thesis, we introduced the CXD corpus, identified verbal indicators of deception across all speakers in the corpus, and developed machine learning classifiers to automatically identify deceptive speech. In Part II of this thesis, we analyzed gender, native language, and personality differences in deceptive speech, and introduced methods to leverage these differences to improve automatic deception detection. The contributions of this work add substantially to our scientific understanding of deceptive speech, and have practical implications for human practitioners and automatic deception detection.

Part IV

Bibliography

Bibliography

Michael G Aamodt and Heather Custer. Who can best catch a liar? *Forensic Examiner*, 15(1):6, 2006.

Mohamed Abouelenien, Verónica Pérez-Rosas, Bohan Zhao, Rada Mihalcea, and Mihai Burzo. Gender-based multimodal deception detection. 2017.

Susan H Adams. Statement analysis: What do suspects' words really reveal. *FBI L. Enforcement Bull.*, 65:12, 1996.

Fayez A Al-Simadi. Detection of deceptive behavior: A cross-cultural test. *Social Behavior and Personality: an international journal*, 28(5):455–461, 2000.

Fayez A Al-Simadi. Jordanian students beliefs about nonverbal behaviors associated with deception in jordan. *Social Behavior and Personality: an international journal*, 28(5):437–441, 2000.

Shahin Amiriparian, Jouni Pohjalainen, Erik Marchi, Sergey Pugachevskiy, and Björn Schuller. Is deception emotional? an emotion-driven predictive approach. *Interspeech 2016*, pages 2011–2015, 2016.

Guozhen An and Rivka Levitan. Comparing approaches for mitigating intergroup variability in personality recognition. *arXiv preprint arXiv:1802.01405*, 2018.

Guozhen An, Sarah Ita Levitan, Rivka Levitan, Andrew Rosenberg, Michelle Levine, and Julia Hirschberg. Automatically classifying self-rated personality scores from speech. *Interspeech 2016*, pages 1412–1416, 2016.

Guozhen An, Sarah Ita Levitan, Julia Hirschberg, and Rivka Levitan. Deep personality recognition for deception detection. *Proc. Interspeech 2018*, pages 421–425, 2018.

Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. Gender, genre, and writing style in formal written texts. *TEXT-THE HAGUE THEN AMSTERDAM THEN BERLIN-*, 23(3):321–346, 2003.

Joan Bachenko, Eileen Fitzpatrick, and Michael Schonwetter. Verification and implementation of language-based deception indicators in civil and criminal narratives. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 41–48. Association for Computational Linguistics, 2008.

Stefan Benus, Frank Enos, Julia Hirschberg, and Elizabeth Shriberg. Pauses in deceptive speech. In *Speech Prosody*, volume 18, pages 2–5, 2006.

Paulus Petrus Gerardus Boersma et al. Praat, a system for doing phonetics by computer. *Glot international*, 5, 2002.

Sissela Bok. *Lying: Moral choice in public and private life*. Vintage, 1999.

Charles F Bond, Adnan Omar, Adnan Mahmoud, and Richard Neal Bonser. Lie detection across cultures. *Journal of nonverbal behavior*, 14(3):189–204, 1990.

Charles F Bond Jr and Adnan Omar Atoum. International deception. *Personality and Social Psychology Bulletin*, 26(3):385–395, 2000.

Charles F Bond Jr and Bella M DePaulo. Accuracy of deception judgments. *Personality and social psychology Review*, 10(3):214–234, 2006.

MT Bradley and Michel Pierre Janisse. Extraversion and the detection of deception. *Personality and Individual Differences*, 2(2):99–103, 1981.

Joseph P Buckley. The reid technique of interviewing and interrogation, 2000.

David B Buller and Judee K Burgoon. Interpersonal deception theory. *Communication theory*, 6(3):203–242, 1996.

Judee K Burgoon, JP Blair, Tiantian Qin, and Jay F Nunamaker. Detecting deception through linguistic analysis. In *International Conference on Intelligence and Security Informatics*, pages 91–101. Springer, 2003.

Tanya L Chartrand and John A Bargh. The chameleon effect: the perception–behavior link and social interaction. *Journal of personality and social psychology*, 76(6):893, 1999.

Danqi Chen and Christopher Manning. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 740–750, 2014.

Keens Hiu Wan Cheng and Roderic Broadhurst. The detection of deception: The effects of first and second language on lie detection ability. *Psychiatry, Psychology and Law*, 12(1):107–118, 2005.

Gokul Chittaranjan and Hayley Hung. Are you awerewolf? detecting deceptive roles and outcomes in a conversational role-playing game. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5334–5337. IEEE, 2010.

François Chollet et al. Keras: Deep learning library for theano and tensorflow. *URL: https://keras. io/k*, 7(8), 2015.

PT Costa and RR McCrae. Neo five-factor inventory (neo-ffi). *Odessa, FL: Psychological Assessment Resources*, 1989.

Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. Mark my words!: linguistic style accommodation in social media. In *Proceedings of the 20th international conference on World wide web*, pages 745–754. ACM, 2011.

Bella M DePaulo, Julie I Stone, and G Daniel Lassiter. Telling ingratiating lies: Effects of target sex and target attractiveness on verbal and nonverbal deceptive success. *Journal of Personality and Social Psychology*, 48(5):1191, 1985.

Bella M DePaulo, Deborah A Kashy, Susan E Kirkendol, Melissa M Wyer, and Jennifer A Epstein. Lying in everyday life. *Journal of personality and social psychology*, 70(5):979, 1996.

Bella M DePaulo, James J Lindsay, Brian E Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. Cues to deception. *Psychological bulletin*, 129(1):74, 2003.

Anna Dreber and Magnus Johannesson. Gender differences in deception. *Economics Letters*, 99(1):197–199, 2008.

Paul Ekman and Wallace V Friesen. Nonverbal leakage and clues to deception. *Psychiatry*, 32(1):88–106, 1969.

Paul Ekman, Wallach V Friesen, and Klaus R Scherer. Body movement and voice pitch in deceptive interaction. *Semiotica*, 16(1):23–28, 1976.

Paul Ekman, Maureen O'Sullivan, Wallace V Friesen, and Klaus R Scherer. Invited article: Face, voice, and body in detecting deceit. *Journal of nonverbal behavior*, 15(2):125–135, 1991.

Frank Enos, Stefan Benus, Robin L Cautin, Martin Graciarena, Julia Hirschberg, and Elizabeth Shriberg. Personality factors in human deception detection: comparing human to machine performance. In *INTERSPEECH*, 2006.

Frank Enos, Elizabeth Shriberg, Martin Graciarena, Julia Hirschberg, and Andreas Stolcke. Detecting deception using critical segments. In *Eighth Annual Conference of the International Speech Communication Association*, 2007.

Frank Enos. *Detecting deception in speech.* PhD thesis, Columbia University, 2009.

Anders Eriksson and Francisco Lacerda. Charlatanry in forensic speech science: A problem to be taken seriously. *International Journal of Speech, Language and the Law*, 14(2):169–193, 2007.

Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462. ACM, 2010.

Hans Jurgen Eysenck and Sybil Bianca Giuletta Eysenck. *Manual of the Eysenck Personality Questionnaire (junior and adult).* Hodder and Stoughton, 1975.

Song Feng, Ritwik Banerjee, and Yejin Choi. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 171–175. Association for Computational Linguistics, 2012.

Jonathan G Fiscus. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover). In *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*, pages 347–354. IEEE, 1997.

Charles V Ford, Bryan H King, and Marc H Hollender. Lies and liars: Psychiatric aspects of prevarication. *The American journal of psychiatry*, 145(5):554, 1988.

Elizabeth B Ford. Lie detection: Historical, neuropsychiatric and legal dimensions. *International Journal of Law and Psychiatry*, 29(3):159–177, 2006.

Tommaso Fornaciari and Massimo Poesio. Automatic deception detection in italian court cases. *Artificial intelligence and law*, 21(3):303–340, 2013.

Emma Franklin. Some theoretical considerations in off-the-shelf text analysis software. In *Proceedings of the Student Research Workshop associated with RANLP*, pages 8–15, 2015.

Alastair James Gill. *Personality and language: The projection and perception of personality in computer-mediated communication.* PhD thesis, Citeseer, 2003.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE, 2013.

Harvard Intelligent Probabilistic Systems Group. Spearmint. `https://github.com/HIPS/Spearmint`, 2017.

Rosanna E Guadagno, Bradley M Okdie, and Sara A Kruse. Dating deception: Gender, online dating, and exaggerated self-presentation. *Computers in Human Behavior*, 28(2):642–647, 2012.

Jeffrey T Hancock, Lauren E Curry, Saurabh Goorha, and Michael T Woodworth. Lies in conversation: An examination of deception using automated linguistic analysis. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 26, 2004.

Jeffrey T Hancock, Lauren E Curry, Saurabh Goorha, and Michael Woodworth. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, 45(1):1–23, 2007.

Jeffrey T Hancock, Catalina Toma, and Nicole Ellison. The truth about lying in online dating profiles. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 449–452. ACM, 2007.

Jeffrey T Hancock. Digital deception. *Oxford handbook of internet psychology*, pages 289–301, 2007.

Julia Bell Hirschberg, Stefan Benus, Jason M Brenier, Frank Enos, Sarah Friedman, Sarah Gilman, Cynthia Girand, Martin Graciarena, Andreas Kathol, Laura Michaelis, et al. Distinguishing deceptive from non-deceptive speech. 2005.

Shuyuan Mary Ho and Jonathan M Hollister. Guess who? an empirical study of gender deception and detection in computer-mediated communication. *Proceedings of the American Society for Information Science and Technology*, 50(1):1–4, 2013.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Frank Horvath. Detecting deception: the promise and the reality of voice stress analysis. *Journal of Forensic Science*, 27(2):340–351, 1982.

Fred Inbau, John Reid, Joseph Buckley, and Brian Jayne. *Criminal interrogation and confessions*. Jones & Bartlett Publishers, 2011.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015.

Chi-Chun Lee, Matthew Black, Athanasios Katsamanis, Adam C Lammert, Brian R Baucom, Andrew Christensen, Panayiotis G Georgiou, and Shrikanth S Narayanan. Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

Rivka Levitan, Agustín Gravano, Laura Willson, Stefan Benus, Julia Hirschberg, and Ani Nenkova. Acoustic-prosodic entrainment and social behavior. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, pages 11–19. Association for Computational Linguistics, 2012.

Sarah I Levitan, Guozhen An, Mandi Wang, Gideon Mendels, Julia Hirschberg, Michelle Levine, and Andrew Rosenberg. Cross-cultural production and detection of deception from speech. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, pages 1–8. ACM, 2015.

Sarah Ita Levitan, Michelle Levine, Julia Hirschberg, Nishmar Cestero, Guozhen An, and Andrew Rosenberg. Individual differences in deception and deception detection. *Proceedings of Cognitive*, 2015.

Sarah Ita Levitan, Yocheved Levitan, Guozhen An, Michelle Levine, Rivka Levitan, Andrew Rosenberg, and Julia Hirschberg. Identifying individual differences in gender, ethnicity, and personality from dialogue for deception detection. In *Proceedings of NAACL-HLT*, pages 40–44, 2016.

Sarah Ita Levitan, Angel Maredia, and Julia Hirschberg. Linguistic cues to deception and perceived deception in interview dialogues. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 1941–1950, 2018.

Sarah Ita Levitan, Jessica Xiang, and Julia Hirschberg. Acoustic-prosodic and lexical entrainment in deceptive dialogue. In *Proc. 9th International Conference on Speech Prosody 2018*, pages 532–536, 2018.

Rivka Levitan. *Acoustic-prosodic entrainment in human-human and human-computer dialogue*. Columbia University, 2014.

Junyi Jessy Li and Ani Nenkova. Fast and accurate prediction of sentence specificity. In *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence (AAAI)*, pages 2281–2287, January 2015.

Sichu Li. Recent developments in human odor detection technologies. *Journal of Forensic Science & Criminology*, 1(1):1–12, 2014.

Kenneth Locke. Neo scoring, 2015.

Shan Lu, Gabriel Tsechpenakis, Dimitris N Metaxas, Matthew L Jensen, and John Kruse. Blob analysis of the head and hands: A method for deception detection. In *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*, pages 20c–20c. IEEE, 2005.

Xiaofei Lu. Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, 15(4):474–496, 2010.

François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, pages 457–500, 2007.

BE Malone, RB Adams, DE Anderson, ME Ansfield, and BM DePaulo. Strategies of deception and their correlates over the course of friendship. In *Poster presented at the annual meeting of the American Psychological Society, Washington, DC*, 1997.

Angel Maredia, Kara Schechtman, Sarah Ita Levitan, and Julia Hirschberg. Comparing approaches for automatic question identification. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, pages 110–114, 2017.

Ewout H Meijer and Bruno Verschuere. Deception detection based on neuroimaging: Better than the polygraph? *Journal of Forensic Radiology and Imaging*, 8:17–21, 2017.

Gideon Mendels, Sarah Ita Levitan, Kai-Zhan Lee, and Julia Hirschberg. Hybrid acoustic-lexical deep learning approach for deception detection. *Proc. Interspeech 2017*, pages 1472–1476, 2017.

Thomas O Meservy, Matthew L Jensen, John Kruse, Judee K Burgoon, Jay F Nunamaker, Douglas P Twitchell, Gabriel Tsechpenakis, and Dimitris N Metaxas. Deception detection through automatic, unobtrusive analysis of nonverbal behavior. *IEEE Intelligent Systems*, 20(5):36–43, 2005.

Gelareh Mohammadi and Alessandro Vinciarelli. Automatic personality perception: Prediction of trait attribution based on prosodic features. *IEEE Transactions on Affective Computing*, 3(3):273–284, 2012.

Ani Nenkova, Agustin Gravano, and Julia Hirschberg. High frequency word entrainment in spoken dialogue. In *Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies: Short papers*, pages 169–172. Association for Computational Linguistics, 2008.

Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards. Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin*, 29(5):665–675, 2003.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 309–319. Association for Computational Linguistics, 2011.

James W Pennebaker and Laura A King. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296, 1999.

James W Pennebaker, Tracy J Mayne, and Martha E Francis. Linguistic predictors of adaptive bereavement. *Journal of personality and social psychology*, 72(4):863, 1997.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of liwc2015. Technical report, 2015.

JW Pennebaker, CK Chung, M Ireland, A Gonzales, and RJ Booth. Liwc. *Austin, Texas; 2007. LIWC2007: Linguistic inquiry and word count [software program for text analysis] URL: http://liwc. wpengine. com/[accessed 2017-02-27]*, 2015.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

Verónica Pérez-Rosas and Rada Mihalcea. Cross-cultural deception detection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 440–445, 2014.

Verónica Pérez-Rosas and Rada Mihalcea. Gender differences in deceivers writing style. In *Mexican International Conference on Artificial Intelligence*, pages 163–174. Springer, 2014.

Verónica Pérez-Rosas and Rada Mihalcea. Experiments in open domain deception detection. In *Proceedings of EMNLP 2015*, pages 1120–1125. ACL, 2015.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

Bashar A Rajoub and Reyer Zwiggelaar. Thermal facial analysis for deception detection. *IEEE transactions on information forensics and security*, 9(6):1015–1023, 2014.

David Reitter and Johanna D Moore. Priming of syntactic rules in task-oriented dialogue and spontaneous conversation. In *Proceedings of the Cognitive Science Society*, volume 28, 2006.

Victoria Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 7–17, 2016.

Björn Schuller, Stefan Steidl, and Anton Batliner. The interspeech 2009 emotion challenge. In *Tenth Annual Conference of the International Speech Communication Association*, 2009.

Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

Kim B Serota, Timothy R Levine, and Franklin J Boster. The prevalence of lying in america: Three studies of self-reported lies. *Human Communication Research*, 36(1):2–25, 2010.

Izhak Shafran, Michael Riley, and Mehryar Mohri. Voice signatures. In *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*, pages 31–36. IEEE, 2003.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.

Aron W Siegman and Mark A Reynolds. Self-monitoring and speech in feigned and unfeigned lying. *Journal of Personality and Social Psychology*, 45(6):1325, 1983.

Nicky Smith. *Reading between the lines: An evaluation of the scientific content analysis technique (SCAN)*. Home Office, Policing and Reducing Crime Unit, Research, Development and Statistics Directorate, 2001.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

Lynn A Streeter, Robert M Krauss, Valerie Geller, Christopher Olson, and William Apple. Pitch changes during attempted deception. *Journal of personality and social psychology*, 35(5):345, 1977.

Duyu Tang, Bing Qin, and Ting Liu. Document modeling with gated recurrent neural network for sentiment classification. In *EMNLP*, pages 1422–1432, 2015.

Global Deception Research Team. A world of lies. *Journal of cross-cultural psychology*, 37(1):60–74, 2006.

Patti Tilley, Joey F George, and Kent Marett. Gender differences in deception and its detection under varying electronic media conditions. In *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*, pages 24b–24b. IEEE, 2005.

Roger Tourangeau and Ting Yan. Sensitive questions in surveys. *Psychological bulletin*, 133(5):859, 2007.

George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 5200–5204. IEEE, 2016.

Gabriel Tsechpenakis, Dimitris Metaxas, Mark Adkins, John Kruse, Judee K Burgoon, Matthew L Jensen, Thomas Meservy, Douglas P Twitchell, Amit Deokar, and Jay F Nunamaker. Hmm-based deception recognition from visual cues. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 824–827. IEEE, 2005.

Morgan Ulinski, Seth Benjamin, and Julia Hirschberg. Using hedge detection to improve committed belief tagging. In *Proceedings of the Workshop on Computational Semantics beyond Events and Roles*, pages 1–5, 2018.

Aldert Vrij and Sally Graham. Individual differences between liars and the ability to detect lies. *Expert Evidence*, 5(4):144–148, 1997.

Aldert Vrij, GUN R SEMIN, and Ray Bull. Insight into behavior displayed during deception. *Human Communication Research*, 22(4):544–562, 1996.

Aldert Vrij, Ronald Fisher, Samantha Mann, and Sharon Leal. A cognitive load approach to lie detection. *Journal of Investigative Psychology and Offender Profiling*, 5(1-2):39–43, 2008.

Cynthia Whissell, Michael Fournier, René Pelland, Deborah Weir, and Katherine Makarec. A dictionary of affect in language: Iv. reliability, validity, and applications. *Perceptual and Motor Skills*, 62(3):875–888, 1986.

Maria Yancheva and Frank Rudzicz. Automatic detection of deception in child-produced speech using syntactic complexity features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 944–953, 2013.

Xiang Yu, Shaoting Zhang, Zhennan Yan, Fei Yang, Junzhou Huang, Norah E Dunbar, Matthew L Jensen, Judee K Burgoon, and Dimitris N Metaxas. Is interactional dissynchrony a clue to deception? insights from automated analysis of nonverbal visual cues. *IEEE transactions on cybernetics*, 45(3):492–506, 2015.

Lina Zhou, Judee K Burgoon, Jay F Nunamaker, and Doug Twitchell. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group decision and negotiation*, 13(1):81–106, 2004.

Miron Zuckerman, Bella M DePaulo, and Robert Rosenthal. Verbal and nonverbal communication of deception1. In *Advances in experimental social psychology*, volume 14, pages 1–59. Elsevier, 1981.

Part V

Appendices

Appendix A

CXD Corpus Forms

A.1 Questions for Baseline Data Collection

Tell me how you decided to come to Columbia.

What do you like most about living in New York City?

What do you like least about living in New York City?

Describe a typical weekend for you, from Friday night through Sunday night.

What was the best food you ever ate. Where did you have it? What made it so good?

Where was the last place you traveled? What are some things you did while you were there?

What was the last movie you saw and what was the plot?

Besides work or school, what do you do with your time?

What did you do this past summer?

A.2 Participant Information

SUBJ #: _____ EXPER: _____ DATE: _____

PARTICIPANT INFORMATION

[Your responses on this sheet are intended only to provide background information about our participants, and do not in any way affect your status as a participant in this study. All information will be kept strictly confidential, and will not have your name attached to it.]

1. Male Female

2. Approximate age (circle one): 20 25 30 35 40 45 50

3. Which hand do you use for writing? Right Left

Language Background:

4. Were you born and raised in the United States? Yes No (if no, list country)

5. What is the first language/dialect you learned to speak fluently?

6. What language(s) did your *mother* speak at home to you while growing up?

7. What language(s) did your *father* speak at home to you while growing up?

8. What language(s) did your *mother and father* speak to each other at home while you were growing up?

9. Do you speak more than one language fluently? Yes No

If yes, please list all languages/dialects you speak (including English), noting in each case whether you first acquired that language through instruction in school (SCH), or due to hearing and using the language while immersed in an everyday social environment (ENV) where it was spoken. Please note also the approximate age from which you acquired the language.

a. _____ SCH ENV (from age: _____)
b. _____ SCH ENV (from age: _____)
c. _____ SCH ENV (from age: _____)

A.3 Gender and Minority Information

SUBJ #: _____ EXPER: _____ DATE: _____

GENDER AND MINORITY INFORMATION

[Our funding agencies (National Institutes of Health, National Science Foundation, etc.) require that all studies maintain records of the gender, race, and ethnicity of all participants. If you decline to provide this information, it will in no way affect your status as a participant in this study. Your cooperation is appreciated. All information will be kept strictly confidential, and will not have your name attached to it.]

Sex/Gender: **Please select one of the following:**

____ Female ____ Male ____ No Report

Ethnicity:

Do you consider yourself to be Hispanic or Latino? (see definition below) Please select one.

Hispanic or Latino: A person of Mexican, Puerto Rican, Cuban, South or Central American, or other Spanish culture or origin, regardless of race.

____ Hispanic or Latino ____ Not Hispanic or Latino ____ Unknown/No Report

Race:

What race do you consider yourself to be? Please select all that apply.

____ **American Indian or Alaska Native.** A person having origins in any of the original peoples of North, Central, or South America, and who maintains tribal affiliation or community attachment.

____ **Asian.** A person having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent.

____ **Native Hawaiian or Other Pacific Islander.** A person having origins in any of the original peoples of Hawaii, Guam, Samoa, or other Pacific Islands.

____ **Black or African American.** A person having origins in any of the black racial groups of Africa.

____ **White.** A person having origins in any of the original peoples of Europe, the Middle East, or North Africa.

____ **Other.**

____ **Unknown/No Report.**

A.4 Sample Biographical Questionnaire

No.	Questions	True Answer	False Answer
1	Where were you born?		[redacted]
2	How many years did you live in your first home?		
3	What is your mother's job?		
4	What is your father's job?		
5	Have your parents divorced?		
6	Have you ever broken a bone?		
7	Do you have allergies to any foods?		
8	Have you ever stayed overnight in a hospital as a patient?		
9	Have you ever tweeted? (posted a message on twitter)		[redacted]
10	Have you ever bought anything on eBay?		
11	Do you own an e-reader of any kind?		[redacted]
12	Who was the last person you were in a physical fight with?		
13	Have you ever gotten into trouble with the police?		
14	Who ended your last romantic relationship?		
15	Whom do you love more, your mother or father?		
16	What is the most you have ever spent on a pair of shoes?		
17	What is the last movie you saw that you really hated?		[redacted]
18	Have you ever gone ice-skating?		
19	Do you currently own a tennis racket?		
20	How many roommates do you have?		
21	If you attended college, what was your major?		
22	Did you ever have a cat?		[redacted]
23	Have you ever watched a person or pet die?		
24	Did you ever cheat on a test in high school?		[redacted]

A.5 Biographical Questionnaire Guidelines

Biographical Questionnaire Guidelines

Instructions
Please use these guidelines to come up with a false answer that is sufficiently different from your true answer. You only need to make up false answers for the questions indicated on the questionnaire.

No.	Questions	Guidelines for False Answers
1	Where were you born?	Not a place you have ever been
2	How many years did you live in your first home?	Add or subtract at least 5 years
3	What is your mother's job?	Pick a field you are not familiar with
4	What is your father's job?	Pick a field you are not familiar with
5	Have your parents divorced?	If Yes, say No. If No, say Yes
6	Have you ever broken a bone?	If Yes, say No. If No, say Yes
7	Do you have allergies to any foods?	If Yes, say No. If No, say Yes
8	Have you ever stayed overnight in a hospital as a patient?	If Yes, say No. If No, say Yes
9	Have you ever tweeted? (posted a message on twitter)	If Yes, say No. If No, say Yes
10	Have you ever bought anything on eBay?	If Yes, say No. If No, say Yes
11	Do you own an e-reader of any kind?	Choose the opposite answer
12	Who was the last person you were in a physical fight with?	Pick someone you haven't fought with
13	Have you ever gotten into trouble with the police?	If Yes, say No. If No, say Yes
14	Who ended your last romantic relationship?	Choose the opposite answer
15	Whom do you love more, your mother or father?	Choose the opposite answer
16	What is the most you have ever spent on a pair of shoes?	Add or subtract at least $200
17	What is the last movie you saw that you really hated?	Pick a movie you have recommended
18	Have you ever gone ice-skating?	If Yes, say No. If No, say Yes
19	Do you currently own a tennis racket?	If Yes, say No. If No, say Yes
20	How many roommates do you have?	Add or subtract at least 2 roommates
21	If you attended college, what was your major?	Pick a subject you have not studied
22	Did you ever have a cat?	If Yes, say No. If No, say Yes
23	Have you ever watched a person or pet die?	If Yes, say No. If No, say Yes
24	Did you ever cheat on a test in high school?	If Yes, say No. If No, say Yes

A.6 Participant Instructions

Participant Instructions

Aim
In this experiment, our goal is to
(a.1)Evaluate how well different people can deceive others
(a.2)Evaluate how well different people can detect when others are being deceptive.

Instructions
Step 1: Please fill out the following 'Biographical Questionnaire'. Answer each question truthfully for each question in the 'True Answer' column. In the 'False Answer' column some rows will be blacked out and others will be blank. For the questions that are blank, and these questions only, you should make up a lie. Please check the 'Biographical Questionnaire Guidelines' when coming up with the lies for these questions.

Step 2: Take a few minutes when you are done to remind yourself of the answers that you just wrote. You want to be able to convince your partner that your answers are true, so greater familiarity is helpful. When you feel comfortable with your modified biography, let the experimenter know. You will be able to look at your answers during your interview.

Step 3: You and your partner will play a game where you take turns playing the role of the interviewer and interviewee.

As Interviewer:
Your aim is to find out when the other person is telling the truth and when they are lying.
Each time you guess correctly, you will earn $1. For every time that you guess incorrectly, you will lose $1. You may ask as many follow up or probing questions as you need to, to help you make each decision.

As Interviewee:
Your aim is to convince the Interviewer that everything in your (modified) biography is true. When you are being interviewed, there will be a keyboard in front of you, which your interviewer cannot see. During each sentence, you must press the 'T' key if what you are saying is true, and the 'F' key if what you are saying is false. While answering a question with a false answer, some of the things you say to justify your answer may still be true. You should press the 'T' key during these sentences. While answering a question truthfully, you should only press the 'T' key and tell no lies.
For every question the interviewer's guesses to be true, you earn $1. For every question that the interviewer guesses to be a lie, you lose $1.

A.7 Interviewer Report

Interviewer Report

Participant No. _____ Date _____

Instructions
Please ask your partner the following questions and listen to his or her answer to each question carefully. You must decide whether you think your partner is lying or not. In order to do this you may ask as many questions as you want about their answers, as well as ask them to provide details.
Mark each row of the "True or False" column with a "T" or "F" indicating whether you think your partner's answer to the question is true or a lie. Indicate your confidence in the correctness of your decision in the "Confidence" column with a number 1-5, with 1 being extremely uncertain and 5 being extremely certain.

No.	Questions	True or False	Confidence
1	Where were you born?		
2	How many years did you live in your first home?		
3	What is your mother's job?		
4	What is your father's job?		
5	Have your parents divorced?		
6	Have you ever broken a bone?		
7	Do you have allergies to any foods?		
8	Have you ever stayed overnight in a hospital as a patient?		
9	Have you ever tweeted? (posted a message on twitter)		
10	Have you ever bought anything on eBay?		
11	Do you own an e-reader of any kind?		
12	Who was the last person you were in a physical fight with?		
13	Have you ever gotten into trouble with the police?		
14	Who ended your last romantic relationship?		
15	Whom do you love more, your mother or father?		
16	What is the most you have ever spent on a pair of shoes?		
17	What is the last movie you saw that you really hated?		
18	Have you ever gone ice-skating?		
19	Do you currently own a tennis racket?		
20	How many roommates do you have?		
21	If you attended college, what was your major?		
22	Did you ever have a cat?		
23	Have you ever watched a person or pet die?		
24	Did you ever cheat on a test in high school?		

A.8 Post Experiment Survey

Post Experiment Survey
Participant ID: _____

1. In your opinion, how many of the judgments that you made today are correct? (Choose the answer that best describes your opinion.)

1	2	3	4	5
almost none	a few	about half	most	almost all

2. In your opinion, how many of the lies that you told today do you think your interviewer believed?

1	2	3	4	5
almost none	a few	about half	most	almost all

3. What strategy did you use in making judgments?

Appendix B

Penn Treebank POS Tag Set

POS Tag	Description	Example
CC	coordinating conjunction	and
CD	cardinal number	1, third
DT	determiner	the
EX	existential there	there is
FW	foreign word	d'hoevre
IN	preposition/subordinating conjunction	in, of, like
JJ	adjective	big
JJR	adjective, comparative	bigger
JJS	adjective, superlative	biggest
LS	list marker	1)
MD	modal	could, will
NN	noun, singular or mass	door
NNS	noun plural	doors
NNP	proper noun, singular	John
NNPS	proper noun, plural	Vikings
PDT	predeterminer	both the boys
POS	possessive ending	friend's
PRP	personal pronoun	I, he, it

PRP$	possessive pronoun	my, his
RB	adverb	however, usually, naturally, here, good
RBR	adverb, comparative	better
RBS	adverb, superlative	best
RP	particle	give up
TO	to	to go, to him
UH	interjection	uhhuhhuhh
VB	verb, base form	take
VBD	verb, past tense	took
VBG	verb, gerund/present participle	taking
VBN	verb, past participle	taken
VBP	verb, sing. present, non-3d	take
VBZ	verb, 3rd person sing. present	takes
WDT	wh-determiner	which
WP	wh-pronoun	who, what
WP$	possessive wh-pronoun	whose
WRB	wh-abverb	where, when

Table B.1: Note: This table is from `https://www.winwaed.com/blog/2011/11/08/part-of-speech-tags/`.

Appendix C

Linguistic Deception Indicator Feature Lexicons

C.1 Hedge Words

completely	hear	likes	estimates	seem
expect	hears	liked	estimated	seemingly
expected	heard	might	fairly	seldom
expects	somebody	general	frequently	several
recall	could	likely	generally	somewhat
recalls	somewhere	sure	guess	speculate
recalled	know	think	guesses	suggest
somehow	knows	thought	guessed	suggests
totally	knew	thinks	largely	suggested
remember	much	may	maybe	suppose
remembers	most	almost	mostly	supposed
remembered	some	apparently	nearly	supposes
should	someone	appear	necessarily	technically
understand	really	appears	occasionally	unlikely
understands	find	appeared	often	unsure

understood	finds	approximately	partial	usually
about	found	arguably	perhaps	virtually
read	imagine	assume	possibly	
reads	imagines	assumes	practically	
sometimes	imagined	assumed	probable	
fair	basic	basically	probably	
possible	believe	consider	propose	
feel	believes	considers	rarely	
feels	believed	considered	rough	
felt	like	estimate	roughly	

C.2 Hedge Phrases

my thinking	sound like
they say	sounds like
they said	sounded like
kind of	the like
sort of	their impression
look like	and the rest
looks like	i would say
looked like	a whole bunch
a little	and all that
a couple	and so forth
a bunch	and so on
a bit	and such like
a few	in my mind
among other	in my opinion
it's say	in my understanding
my understanding	in my view
pretty much	more or less
so far	something or other
somebody says	to be honest
somebody said	

C.3 Cue Phrases

actually	next
also	no
although	now
and	ok
basically	or
because	otherwise
but	right
essentially	say
except	second
finally	see
first	similarly
further	since
generally	so
however	then
indeed	therefore
like	well
look	yes