Detecting Deception in Interrogation Settings

by

Carolyn Elizabeth Lamb

A thesis submitted to the
School of Computing
in conformity with the requirements for
the degree of Master of Science

Queen's University
Kingston, Ontario, Canada
December 2012

Abstract

Bag-of-words deception detection systems outperform humans, but are still not always accurate enough to be useful. In interrogation settings, present models do not take into account potential influence of the words in a question on the words in the answer. According to the theory of verbal mimicry, this ought to exist. We show with our research that it does exist: certain words in a question can "prompt" other words in the answer. However, the effect is receiver-state-dependent. Deceptive and truthful subjects in archival data respond to prompting in different ways. We can improve the accuracy of a bag-of-words deception model by training a machine learning algorithm on both question words and answer words, allowing it to pick up on differences in the relationships between these words. This approach should generalize to other bag-of-words models of psychological states in dialogues.

Contents

List of Tables

List of Figures

List of Acronyms and Abbreviations

CBCA Criteria-Based Content Analysis

EM Expectation Maximization

excl Exclusive words other than "but" and "or"

FPP First Person Plural

FPS First Person Singular

LSA Latent Semantic Analysis

LSM Linguistic Style Matching

MATLAB MATrix LABoratory, a computer program

neg Negative emotion words

RM Reality Monitoring

SCAN Scientific Content ANalysis

SVD Singular Value Decomposition

TPP Third Person Plural

TPS Third Person Singular

wh "Wh" words ("who", "what", "when", "where", "why", "how")

Chapter 1

Introduction

We all would like to be able to spot liars. From court cases to airport security, many high-stakes real-world situations rely on the human ability to figure out who is deceptive and who is not. Sadly, human ability in this area is not very good. There is variation between individuals, but overall, humans in controlled trials detect deception at chance levels [33].

Automating the detection of deception would bring many benefits. First, if this automated method performed better than humans—and better than, say, a polygraph machine—it could be used to augment human judgement in high-stakes situations. Second, automated deception detection could be used in lower-stakes situations where paying for expert human judgement is infeasible. For instance, a filter alerting users to potential deception could be used with chat rooms or online job postings.

Sadly, none of our deception-detection methods have reached this level of performance. One problem is that deception seems to work differently in different situations. The cues which predict deception online, for instance, are different from the ones that emerge in face-to-face deception [45]. Similarly, the cues shown by people who are

highly motivated to be deceptive are different from the cues shown by people who have little to gain from it [30].

The variable we focus on here is a slightly different one. In many studies of deception, participants are asked to extemporize a deceptive monologue. For example, they might be asked to tell a story of a personal experience that did not actually happen to them. But many interesting real-world cases—from police interrogations to online chat—do not involve monologues. Instead, the deceptive person is in a dialogue, answering questions from another person. As in any other dialogue, each participant responds and adjusts to the linguistic style of the other. This makes deception detection in a dialogue more complicated. If one participant shows signs of deception, are they being deceptive, or are they responding to some aspect of the other participant's prompting?

We have undertaken an exploratory study of these issues by analyzing archival transcripts of real-world interrogations and applying a deception model developed by Newman *et al.* [70]. In this model, deception is marked by changes in the frequency of certain words. However, Newman *et al.* developed this model by looking only at monologue data. Its applicability to dialogues is an open question. We have looked into this question by applying the model to archival question-and-answer data from recent political debates, the Nuremberg trials, and a civil suit.

Our first goal was to detect the presence of a prompting effect in words relevant to Newman *et al.*'s model. For any of these words in the answer, could it be shown that their rates of use covaried with the rates of use of a word in the question? We discovered that for some words, particularly first-person pronouns, this prompting does occur. The more a questioner uses certain words, for example, second-person

pronouns, the more the respondent tends to use first-person pronouns.

This implies that a model like Newman *et al.*'s should not be applied to dialogue data in its present form, because the words in the answer will deviate from monologue-based expectations if significant prompting occurs. Models of processes other than deception which use some of the same kinds of words are presumably sensitive to the same effects.

Our second goal was to fix this problem by removing the effect of the question from each answer, thereby expanding the applicability of the deception detection model. If the respondent has a mental state which would produce certain words in a monologue, and the questioner has an influence which distorts the use of these words, we wanted to undo this distortion. We developed a method of removing it through a series of affine transformations on the data. However, using this correction method on some of our real-world question-and-answer data revealed that our assumptions had been overly simplistic. The correction method actually reduced differences between deceptive and non-deceptive answers, when we had hoped it would increase them.

Delving further, we discovered the reason for this: namely, deceptive and non-deceptive people in the data responded to prompting in different ways. Removing the prompting effect also removed some of the signs of difference between these subgroups. In deception, it is not the case that the respondent has an underlying mental state whose effects are distorted in a uniform way by prompting. Instead, it appears to be the case that prompting occurs, but the respondent's mental state influences the effect of that prompting.

Since respondents with different mental states respond differently to questioning, we ought to be able to improve deception detection by taking into account both the

words in the answer and the words in the question. In this way, differences in how respondents respond to similar questions can become part of the model. We tested this hypothesis by running two of our data sets through a machine-learning-based classifier. In both cases, the classifier performed markedly better when taking both question and answer words into account, supporting our predictions.

We have made a contribution to deception detection in showing that existing word-based models are incomplete, and that we can make needed improvements to their accuracy by taking question words into account. We have also contributed by discovering the reason this is so: the relationships between question and answer words, and not merely the answer words themselves, are indicative of deception. In a broader sense, we have discovered a way to track the influence of words in one utterance on the words in a responding utterance. We expect this to generalize to other word-based psychological models which data miners or psychologists might wish to apply to dialogues.

In Chapter 2 of this thesis, we review previous work in deception detection and in other relevant aspects of word-based psychological models, including verbal mimicry. We also introduce the mathematical and computational ideas we will be using in later chapters: singular value decomposition, bag-of-words models, bivariate Gaussian and gamma distributions, and random forests. Chapter 3 describes our experimental setup and the three data sets we used. Chapter 4 describes our results and the discovery that the prompting effect is receiver-state-dependent. Lastly, in Chapter 5 we discuss implications, applications, and limitations of these results.

Chapter 2

Background

2.1 Singular Value Decomposition

We will begin by explaining some mathematics that are needed to understand our analysis. A singular value decomposition is a factorization of a real matrix A into three other matrices:

$$A = USV^T$$

such that U and V^T are unitary matrices and S is a diagonal matrix [88].

The matrices U and V contain singular vectors in each column: these represent a series of axes and coordinates corresponding to linearly independent components of the data. (In other words, they are a representation of the data as vectors in a many-dimensional space.) What is more, these axes appear in the matrices in decreasing order of size: the first rows of V^T represent the axes of maximal variation in the data. The nonzero values of the matrix S, meanwhile, correspond to the amount of variation along each axis. One can therefore truncate a set of SVD matrices at an arbitrary number of dimensions and be confident that the truncated version is the

most faithful possible representation of A in the chosen number of dimensions. The S matrix values provide a suggestion of how much data is being lost [86].

SVD was discovered independently five times between 1873 and 1912 [88], but it is used now for data mining in ways its discoverers could not have foreseen. In particular, when dealing with data of very high dimensionality, it is useful to use SVD to reduce the number of dimensions. This not only compresses the data but elucidates correlations between different parts of it, because documents or variables that vary together will wind up being represented as closely aligned vectors in the SVD space [28].

In the next section, we discuss how singular value decomposition has been applied in text mining.

2.2 A Sample Bag of Words Model

A frequent technique in text mining is to analyze a document based on the words it contains. The order of the words does not matter: only the words themselves and the frequencies with which they appear are considered. These techniques are therefore called "bag-of-words" models.

An example of this technique is Latent Semantic Analysis (LSA) [28]. In LSA, one first takes a large corpus of language data and converts it into a term-document matrix. Each row i represents a document, such as a short essay or chapter in a book; each column j represents a word; and each element ij in the matrix represents the number of times that word appears in the document. Conceptually, this produces a "semantic space" with as many dimensions as words, with each document represented as a vector in the space. One can then compute similarities between documents by

taking the vectors for two documents and calculating their dot product in the semantic space.

However, the semantic space is not very useful in this form. A space representing the entire English language will have hundreds of thousands of dimensions, and calculations will be slow; more importantly, the space cannot represent synonymy. If one document contains multiple copies of the word "doctor", for instance, and another contains multiple copies of the word "physician", these documents are probably dealing with roughly similar topics, but a semantic space will not "know" that the words are similar and that documents containing either one should be grouped together.

This is where SVD comes in. Recall that SVD allows one to truncate a matrix into an arbitrarily small number of maximally representative dimensions, and that documents or variables which covary will end up close together in these new dimensions. By running a document-word matrix through SVD and compressing the space to a few hundred dimensions, LSA techniques produce a new semantic space in which words that are used together, and in similar ways, appear close together. Since "doctor" and "physician" are used in the same way and with reference to the same topics—and, frequently, together in the same document—the angle between them will be small. LSA's semantic space thus provides an approximate representation for the underlying meaning of words.

LSA proves to be a surprisingly powerful technique for processing these meanings. Perhaps most strikingly, training an LSA model on representative examples of good and bad essays produces an automated essay scoring system that agrees with skilled human scorers [36]. The underlying model can be used as part of an electronic tutoring system, performing other tasks like suggesting appropriate texts for the student to

read next and locating possible flaws in their understanding [42].

Note that when constructing an LSA corpus—or a data set to analyze using LSA techniques—one must decide how large each document should be. The official LSA corpora truncate most documents at several hundred words, but it is entirely possible to use much larger documents. Any "bag-of-words" model that faces this kind of decision: to make a bag-of-words, one must first choose a size for the bag. In theory a bag can be any size. One can put all the words in a book, for instance, into a single bag and analyze them as a unit. At the opposite end of the size scale, one can analyze single words or "N-grams" consisting of two or three words together at a time. This latter technique is often used with techniques that do not involve dimensionality reduction, such as naive Bayes classifiers (e.g. Zhou *et al.*'s deception model discussed in Section 2.4 [104]).

With N-grams and other small bags, one often uses a "sliding window" algorithm. That is to say, at $N = 3$, the first N-gram consists of the first three words in the document. But the second N-gram does not start where the first left off: it takes the second, third, and fourth words in the document, thereby overlapping with the first N-gram. The sliding window algorithm increases the granularity of the analysis without losing the ability to look at groups of words together or to take word order into account. It can be used at various window sizes as well as various overlap sizes [1].

A final note about LSA is that not all words are included in the matrix on which SVD is performed. It is common practice to remove a few "stop words": words like "the", "I", and "if", usually including all pronouns, which are thought to be too common to be worth analyzing [28]. It is also common practice to perform "stemming"—removing tense, number, and other information from a word, leaving

only the root. For example, the word "stemming" itself would be shortened to "stem".

2.3 The Secret Life of Stop Words

LSA is not the only technique that involves removing "stop words". All other gold standard "bag-of-words" models, such as Latent Dirichlet Allocation [5], include this step. It is thought that these stop words have only a grammatical function, and thus are unnecessary in a model that does not take word order into account. But stop words contain interesting meanings of their own.

Consider the following sentence:

I took him to the theatre after that.

The words "I", "him", and "that" are stop words, but they are not simply indicators of grammatical structure. "I" and "him" refer to people, and "that" to an event: they have as much referential meaning as the word "theatre", which refers to a place. However, the word "theatre" has a meaning every English speaker knows. In contrast, the meanings of "I", "him", and "that" are almost entirely opaque without context. "I" presumably refers to the person speaking, and "him" to a male person who is not the one speaking or being addressed, but when the sentence is taken out of context, we can't tell who these people are, and we certainly do not know anything about the events to which "that" refers. The article "the" also indicates that a particular theatre is being discussed, but without the social context, we do not know which one. These words are thus inherently social, having a layer of meaning that is specific to the people in the conversation.

"I", "him", and "that" perform referential functions in a sentence; "to" and "after" perform other functions, mostly to do with the grammatical relationship between

other words. Social psychologists call all of these words "function words". Function words—as opposed to "content words" like "theatre" —are words that lack a direct lexical meaning, instead serving grammatical functions. Function words are also called "style words", and it is this second name that gives us a hint as to their uses in text mining.

Consider these three sentences:

1. I think there would probably be a lot of people living in Tokyo, wouldn't there?

2. One can certainly see a great number of people living in Tokyo.

3. So, like, oh my God, I can't believe how many, like, people live in Tokyo.

Each of these sentences conveys the same information (that a lot of people live in Tokyo). But each sentence is written in a different style. From the stylistic markers in each sentence, we can infer things about the three speakers. The speaker in the third sentence is probably young and female and speaking in a casual context, while the second speaker appears older and more formal. The first speaker, meanwhile, seems to feel unsure of him or herself. Some of these differences in style are expressed in content words, but the majority come from function words: the avoidance of first-person pronouns in the second sentence, for instance, is one signifier of formality, and the use of words like "would" and "probably" in the first sentence convey uncertainty. Since the meanings of function words are inherently social, it is perhaps not surprising that this is where much of our social intuition about the speakers comes from. If we ran these sentences through LSA and removed stop words, we would lose a great deal of the social and psychological distinction between them.

Function words are particularly useful for the discovery of underlying mental states. They are convenient for this analysis for two reasons. First of all, they are more plentiful than content words. Although a normal English speaker has only 500 function words in their vocabulary (out of 100,000 total English words), 55% of all the words they speak or write on a given day will be function words [92]. These words are produced in a different part of the brain than content words [65] and can "leak" information about a person's mental state even without their awareness [25].

Social psychologists in the past twenty years have been analyzing people's speaking and writing styles by counting both function words and common content words. The foremost tool for this analysis is Pennebaker's [74] Linguistic Inquiry and Word Count program (LIWC) which processes text and outputs word counts for more than 70 linguistic categories, including both function and content words. Users can also augment the dictionary with their own special-purpose lists of words. While the task of counting words is not in itself very complicated, creating a dictionary of words from different categories is difficult and time-consuming. LIWC saves researchers from this labor by providing standard dictionaries which have been vetted by psychological judges. The program thus makes it simple to conduct research correlating the use of different kinds of words with all sorts of other interesting variables.

The use of easily-overlooked function words to discern mental state is consistent with Ekman's [32] work on facial microexpressions and "leakage". Although people are mostly in control of their speech and facial expression, unconscious processes modify these signals in subtle ways. A person attempting to convey one mental state experiences "leakage" when their true mental state briefly shows through. Ekman studies this in the context of changes in facial expression, but the idea of verbal leakage

is also consistent with his theories, and in fact goes all the way back to Freud [41]. Function words are a strong venue for this verbal leakage given their different origin in the brain [65] and the fact that people do not pay conscious attention to them. In fact, people have difficulty keeping track of function words even when instructed to do so [51].

It turns out that linguistic style is extremely useful in predicting the mental state of a speaker or author. Pronouns alone reveal a huge variety of personal characteristics. In general, pronouns provide an indication of whom the speaker is focused on: themselves, the person they are speaking to, or others [92]. People in great emotional pain understandably focus on themselves: students with depression use more first-person singular pronouns than control students [81], as do neurotic students [75], while the work of poets who committed suicide contains more first-person singular pronouns and fewer first-person plural pronouns than the work of similar poets who died of other causes [89]. People with Machiavellian personalities also use more first-person singular pronouns [50]. Older people [77] and people with agreeable personalities [75] make fewer references to themselves, and people make fewer references to themselves at work than during leisure time. People use first-person plural pronouns less often when on the phone, and more often when in public places [62].

Since they show focus on the self and others, pronouns reflect relationship quality. Hostile relatives of mentally ill patients use the word "me" more in interactions with their mentally ill family member (compared to non-hostile relatives in the same situation), along with more second-person pronouns and fewer first-person plural pronouns, suggesting that they think of themselves in opposition, not together [83]. Couples who use more first-person plural pronouns are better at solving their marital

problems than others, and those who use more first-person singular pronouns when discussing their problems (consistent with the common therapeutic advice to use "I statements") are more satisfied in their relationships later on [84].

Using more pronouns in every category is a sign of increased (self-perceived) rapport with the listener, since listeners must be trusted to understand who each pronoun refers to [69]. Women use more pronouns overall than men [69], heterosexuals use more pronouns than homosexuals [43], and manic patients use more pronouns than normal controls [61].

LIWC also counts words related to various emotions. Kahn *et al.* [56] tested the construct validity of LIWC's emotional word count and found that the word categories associated with various emotions do, in fact, correspond to those emotions. People use more sadness words when instructed to tell a sad story, more positive emotion words when instructed to tell an amusing one, and so on.

Perhaps obviously, people use more negative emotion words when unhappy—if they are undergoing personal upheaval [26], if they are depressed [81], or if they are high in neuroticism [75]. Even using positive emotion words in a negative way, such as by saying "I am not happy," is indicative of better outcomes than saying "I am sad." Many of the Big Five personality traits (agreeableness, conscentiousness, extraversion, neuroticism, and openness) are associated with the use of positive and negative emotion words: neurotic people use fewer positive emotion words as well as more negative emotion words, while conscientious people use fewer negative emotion words and extroverts use more positive ones [75]. Older people use a greater number of positive emotion words [77], and women also use more positive emotion words than men, while men use more words associated with anger [43].

Other interesting dimensions of LIWC include cognitive complexity and distancing. Distancing is a composite measure related to a person's detachment from what they are talking about. It includes an increase in articles and long words (defined as words of more than 6 letters) along with a decrease in first-person pronouns, modal verbs, and present tense verbs. People under personal stress use more of this distancing language [26]. People telling stories that they have not told before, as opposed to stories they had rehearsed with others, use less-distanced language and are more tentative [73]. Distancing is more common in people with open personalities, while neurotic and agreeable people distance themselves less [75].

Cognitive complexity involves longer words and words relating to thinking, causation, and insight. It also involves a category called "exclusive" words (e.g. "but", "or", "whereas") which make fine-grained distinctions between concepts. Older people exhibit more cognitive complexity in their language [77], while extroverted and conscientious people exhibit less [75]. Increased cognitive complexity over time may be a sign of learning and growth: people who increased complexity over time in their writing reported better outcomes than people who stayed the same, including better grades, fewer negative health symptoms, a quicker uptake of new jobs after a layoff, and a more positive state of mind a year after bereavement [76].

Aside from this, women [43] and extroverts [75] use more "social" words (words other than pronouns denoting people and relationships); younger people [4] and women [62] use more "filler" words ("like", "um"); people with open personalities use more tentative words, while extroverts use fewer [75]; men use more swear words than women [62]; and so on.

Many of these effect sizes are fairly small and probabilistic, holding only on average

across a large sample of people. Nevertheless, it should be clear by now that the "stop words" text miners often throw out contain a wealth of information. Rather than ignoring them, we should be embracing the things these words can tell us about the people who use them.

2.4 Deception Detection

Humans by ourselves are not very good at detecting deception. From an evolutionary perspective, one would not expect us to be good at it: since deception and detection of deception are both adaptive, the likely scenario is an evolutionary arms race between the two abilities, with neither one pulling too far ahead for long. Moreover, it's easier to improve skill at deception than at deception detection, because it's easy to tell when one has failed at telling an effective lie and think about how to make the next one more effective. It is not so easy to tell when one has been fooled by an effective liar. So liars ought to learn their skills faster [54].

Indeed, human deception detection rarely surpasses these low expectations. Individuals have varying levels of proficiency, but when taken as a group, college students and other ordinary subjects rarely perform above chance [33]. Even groups of humans who are employed professionally to detect lies, such as police officers [97], customs inspectors [58], federal polygraphers, judges, and psychiatrists [33], tend to do no better. Moreover, within these groups it is difficult to predict who will perform well: age, sex, experience, self-reported ability to lie, and self-reported proficiency at deception detection do not correlate with performance [33, 97]. Only a few groups have significantly better than chance accuracy: this includes members of the Forensic Services Division of the U.S. Secret Service [33], CIA agents with experience in deception,

clinical psychologists with a special interest in deception [34], and people with a high aptitude at detecting facial microexpressions [40].

Many people have tried to detect deception by analyzing the language of deceptive and non-deceptive research subjects, but the majority of these methods cannot be automated because they require trained analysts to evaluate every transcript. One such model is Criteria-Based Content Analysis (CBCA), which was originally developed to distinguish between true and false reports of sexual abuse by children. CBCA involves a human rater reading a transcript and giving scores for various criteria, such as "unique sensory detail" [2]. Inter-rater reliability can be low [87]. Although CBCA performs reasonably well in some studies (e.g. [95]), it was created for a specific purpose and may not generalize well to other contexts. Most obviously, it contains criteria such as "pardoning the perpetrator" which are meaningless in contexts where no one is reporting any crimes. Other criteria, such as reporting believable details but misunderstanding their significance, may be applicable only to children. A study of deception in college students [79] found that only three of the CBCA criteria—detail, coherence, and "admitted lack of memory"—provided significant results. It should also be noted that expert CBCA raters are easily fooled by liars who understand the CBCA criteria [96].

A related model is Reality Monitoring (RM). RM is based on the idea that memories originating in lived experience and memories originating in imagination have different properties. For instance, real memories are more easily located in space and time and contain more fine-grained sensory detail. Deception detection using RM involves counting all the sensory details described, including spatial and temporal information and descriptions of actions [53]. Like CBCA, RM can have low

inter-rater reliability [87]. RM has had mixed results in practice. It performs well with unintentionally falsified memories [82] but may not be as appropriate for deliberate deception, since a well-prepared liar can invent sensory details or appropriate them from unrelated true memories [79]. In fact, in online communication, deceptive emails [103] and chat messages [45] contain *more* sensory detail than truthful ones. Furthermore, humans in face-to-face communication can detect deception by omission as readily as deception by fabrication, suggesting that the sensory properties of fabricated memories are not central to the nature of deception [79].

Another model, Scientific Content ANalysis (SCAN), is somewhat closer to being automatable. SCAN works on the assumption that truthful stories are coherent and complete in ways that it is difficult for a deceptive person to imitate. It assigns deceptive scores to linguistic cues that suggest a lack of coherence or completeness, such as jumps in time, inconsistency in verb tense, or a lack of pronouns and emotion words. Several of these cues, particularly the latter two, could be automated with LIWC. SCAN has been used to correctly discriminate between innocent and guilty accused criminals [31], but not every study has replicated this effect [79]. Interestingly, SCAN includes some criteria that explicitly contradict those of CBCA: it considers "admitted lack of memory" a sign of deception, while CBCA considers it a sign of truthfulness. In Driscoll's study [31], this was one of the least effective parts of the SCAN model.

Other researchers have searched for cues to deception in other ways with varying degrees of success. Zhou [104] uses a Bayesian model to calculate the probability of various N-grams (that is, groups of N words appearing in order) in a truthful or deceptive statement given the N-grams appearing before them. This method is

vulnerable to data sparsity and overfitting, and works best with 1-grams, suggesting that it functions essentially as another kind of bag-of-words model. Further studies have associated deception with longer sentences [102], shorter sentences [9], more emotion [102], less emotion [9], and similar contradictory things.

Many researchers have attempted meta-analysis of all these models to try to make sense of all the conflicting information [30]. This is made more difficult by a lack of current publically available benchmarks for deception detection systems [104]. No single cue is reliably indicative of deception across studies [15]. One large meta-analysis by DePaulo *et al.* [30] found that deceptive people in general do provide less detail, distance themselves more from their listeners, are more tense, and make more negative statements and complaints. The biggest difference between truth and deception, in this meta-analysis, is that deceptive statements do not make as much sense as truthful ones. Another meta-analysis by Zuckerman *et al.*. [106] found that deception is associated with signs of high arousal and high cognitive complexity.

One potential reason for contradictions between studies is that they study deception in a variety of different situations. For instance, the deceptive people in each study have varying levels of motivation, from criminals trying to avoid conviction to college students who are only trying to be deceptive because the people running the study told them to. Contexts and communicative mediums vary, including spoken and written statements, emails, computer-mediated chat, and face-to-face interrogations. DePaulo *et al.* found that effect sizes increased in situations of greater motivation, in more interactive contexts, and in situations involving a transgression [30]. Deceptive people in situations of greater motivation are also more rigidly controlled in their expression [106]. Computer-mediated communication provides each person with time to

rehearse and compose what they say. Thus, while face-to-face deception may involve less detail [30], computer-mediated deception appears to involve more [45, 105].

If we would like to automate deception detection in spite of these complexities, natural language is a good place to start. Text data for natural language processing is cheap and easy to gather [86]. It has high reprocessibility (a transcript will have the same words in it no matter who counts them) and involves function and content words which have been shown to be psychologically meaningful in other areas. Moreover, natural language processing systems already exist at a relatively high level of sophistication compared to, say, facial expression processing [15].

2.5 The LIWC Deception Model

The model of deception we will be using in this thesis comes from Newman *et al.* [70]. They asked college students to speak or write, either deceptively or truthfully, about a variety of topics (political issues, their feelings about their friends, and a mock theft). Then they analyzed all the truthful and deceptive statements with LIWC. Four linguistic signs of deception emerged across categories:

- First person singular pronouns. These decrease in the deceptive condition. Deceptive people have less personal experience with their subject matter than truthful people, and are less emotionally willing to commit to what they are saying, so they focus on and refer to themselves less.

- Negative emotion words. These increase in the deceptive condition. Deceptive people are thought to do this as a sign of unconscious discomfort. Alternatively, increased emotionality may be used as a persuasive tool: this was what Zhou

Categories	Keywords
First-person pronouns	I, me, my, mine, myself, I'd, I'll, I'm, I've
Exclusive words	but, except, without, although, besides, however, nor, or, rather, unless, whereas
Negative emotion words	hate, anger, enemy, despise, dislike, abandon, afraid, agony, anguish, bastard, bitch, boring, crazy, dumb, disappointed, disappointing, f-word, suspicious, stressed, sorry, jerk, tragedy, weak, worthless, ignorant, inadequate, inferior, jerked, lie, lied, lies, lonely, loss, terrible, hated, hates, greed, fear, devil, lame, vain, wicked
Motion verbs	walk, move, go, carry, run, lead, going, taking, action, arrive, arrives, arrived, bringing, driven, carrying, fled, flew, follow, followed, look, take, moved, goes, drive

Table 2.1: Words used in the LIWC deception model

et al. [105] observed in their study of deceptive emails.

- More motion verbs ("go", "run") and fewer exclusive words ("but", "or", "whereas") as a result of a reduction in cognitive complexity.
- More action verbs ("go", "run", "take") as a result of the same reduction. [70]

The words used in these criteria are summarized in Table 2.1

There is mixed theoretical support for the use of these categories. Lowered self-reference is a feature often mentioned in deception models (e.g. Zhou *et al.* [102]). Hancock *et al.* [45] found that, unlike some features, it persists regardless of the

deceptive person's level of motivation. DePaulo *et al.*'s meta-analysis [30] casts doubt on this, noting that while deceptive people do distance themselves from their listeners, self-references as such are not a reliable cue across studies. Zuckerman *et al.*'s results [106] are also mixed, noting that there is a decrease in self-references across studies, but it is not statistically significant.

Zhou *et al.*'s email study found an increase in both positive and negative emotion words in deceptive email. Truthful people using email were less emotional and less persuasive, feeling that they did not need to prove anything. Expression of negative emotions in general is a cue to deception found in many studies and supported by DePaulo *et al.*'s meta-analysis [30].

DePaulo *et al.* point out that this increase should not generalize to every situation. Most lies in everyday life are "white lies", told in order to smooth over social interactions, and the distress people experience over these lies is minor [29]. Meanwhile, a person confessing truthfully to a wrongdoing would presumably be very distressed about it [30]. For similar reasons, this part of the model should not be applied to psychopaths, who experience no discomfort when breaking moral rules [79]. Toma and Hancock [94] found a *decrease* in negative emotion words in deceptive online dating profiles, suggesting that, at least in computer-mediated contexts, there are some situations where deceptive people will attempt to over-persuade by specifically avoiding such words.

Deception causes a higher cognitive load than truthfulness [30], which produces an increase in visible signs of cognitive difficulty [106], even when the deceptive person has had time to prepare and does not appear to be nervous [97]. This happens because deceptive people must put mental effort into monitoring their performance

and making sure it is convincing [30]. For instance, they must display the appropriate emotions to match their words (which may not be the emotions they are really feeling), and they must make sure not to contradict themselves. Deceptive people in laboratory studies self-report increased anxiety and more concern with self-presentation than truth-tellers [100].

However, this ought not to be true with all forms of deception: some white lies are easy to tell, and some forms of truth (such as a potentially distressing confession) should be as mentally effortful as a lie. While deceptive people on average may have a higher cognitive load, they are not the only ones who must monitor their performance to make sure it has the desired effect [30]. DePaulo *et al.*'s meta-analysis found that, with deception overall, the only reliable cue to higher cognitive load was a reduction in detail. However, when lies were extended or unrehearsed, other cognitive complexity cues emerged [30]. Meanwhile, in computer-mediated communication, Hancock *et al.* [45] did not find a connection between exclusive words and deception.

The LIWC model [70] performed significantly better than human raters in distinguishing truth from deception in the political topics, whether the statements were spoken or written. However, its performance was lower on the other topics, suggesting that the model needs to be fine-tuned to different contexts.

Gupta and Skillicorn [44] suggest that the LIWC model detects not only outright falsehood, but also "spin" or "persona deception", in which a person does not make factually false statements, but consciously projects an image of themselves that they know to be inaccurate. Skillicorn and Leuprecht have used this reasoning to apply LIWC to the speeches of politicians [85].

Further LIWC-based analysis such as Hirschberg *et al.* [35, 47] and Mihalcea and Strapavara [64] have produced slightly different sets of cue word categories. Hauch *et al.* recently meta-analyzed all such studies, as well as a few other studies in which the frequency of categories of words was taken into account, and found support across data sets for the four word categories in Newman *et al.*'s original paper, as well as some additional categories [46]. Since we started this study before Hauch *et al.* published their results, we have limited our analysis to the original four categories.

2.6 Fine-Tuning the LIWC Model

Applying the LIWC model to new contexts has revealed more about its internal structure. Gupta and Skillicorn [44] applied LIWC to a data set consisting of internal emails at Enron in the period just before its fradulent accounting scandal. Performing SVD on all words in the model gave an indication of which were most significant and how closely they corresponded with each other. They found that first-person singular pronouns are the most important part of the model, and that the four parts of the model appear to work separately, not together. That is, a decrease in first-person pronouns is not necessarily associated with a decrease in exclusive words (or an increase in action verbs and negative emotion words), but all these factors taken together are associated with deception. The blandest function words ("I", "but", "go") lie along the axes of greatest variation.

In the same data set, Keila and Skillicorn [57] found that LIWC performed well: the top ranked emails by a measure combining LIWC and SVD were all deceptive. However, in such a large data set, it is difficult to tell how many false negatives there are. Keila and Skillicorn also noted that, in many of these detected deceptive emails,

the expected increase in negative emotion words did not appear. They speculated that, at Enron, deceptive employees did not experience unconscious distress because they did not think they were doing anything wrong.

Skillicorn and Little [86] repeated the SVD-based analysis on transcripts from the Gomery commission, in which former Canadian government officials were examined regarding alleged corruption. They found that the model performs well when collapsed into a few dimensions with SVD, because words within a given category (except exclusive words) were correlated with each other and formed sensible structures in semantic space. They also found that in this data, deception was associated with *increases*—not decreases—in first-person singular pronouns and exclusive words. Altering the model to look for increases in all categories, they produced results in rough agreement with media estimates of who was being deceptive and who was not. (Although ground truth for the Gomery data was not available, the most deceptive people according to the model were people saying they did not remember basic facts about their own employment, such as who they were working for. Meanwhile, the least deceptive people were witnesses called in to explain purely technical matters. The speech of lawyers—since lawyers, in the Gomery commission, were not employed to argue a particular case—was also low in deceptiveness.)

Ott *et al.* [72] built a LIWC-based model to distinguish deceptive "spam" product reviews on the Internet from real ones, and found a higher incidence of first-person singular pronouns in spam, further supporting the idea that first-person singular pronouns in some situations increase with deception. Ott *et al.*'s results are also consistent with Zhou *et al.*'s [104] findings, in which deceptive people in computer-mediated communication feel that they have more to prove, and have more time

to think about their words and persuasion methods, resulting in more detailed and persuasive messages being associated with deception online.

If deception decreases first-person pronoun use in some situations and increases it in others, then this explains DePaulo *et al.* [30] and Zuckerman *et al.*'s [106] inability to find a significant effect for self-references across many studies. However, it raises the more vexing question of what causes these words to behave in this way. Skillicorn and Little [86] suggest that first-person singular pronouns and exclusive words increased with deception in the context of the Gomery commission because it was not an emotionally charged situation.

2.7 Verbal Mimicry

In the case of the Gomery commission, there may be another reason results were different from normal: the Gomery commission was a question and answer setting. The people testifying were not simply extemporizing a true or false statement, but were answering specific questions from specific people (in this case, lawyers). While this was the case in some of the other deception studies cited here, it was not the case in the original LIWC deception study, nor in most of the other research done using that model.

In deception involving two people, it is not only the deceptive person's words that change, but the other person's words as well. A deceptive person does not do something linguistically aberrant by themselves: rather, they and the listener match each other's linguistic patterns, with the result that deception changes the whole character of the discussion for both participants, even if the deceptive person's partner is not aware of their deception [45]. This is a case of a more general phenomenon

known as verbal mimicry. Two people in conversation, even if they are strangers, will imitate each other in everything from facial expression and body language [17] to volume [66] and pitch [55] of the voice, speech rates [99], and length of silences [14]. People speaking to each other also mimic each other's words, from single words and short phrases to entire sentences using the same grammatical structure as a previous sentence [59].

This mimicry is generally not conscious [17]. A striking example of unconscious mimicry is cryptomnesia: accidental plagiarism of phrases one has heard before. Cryptomnesia occurs in studies even when subjects have no motivation for plagiarism and are consciously trying to avoid it [8]. In a similar vein, Levelt and Kelter [59] found that mimicry of single words occurred even when subjects could not be consciously thinking of the words that were used in the question, because their working memory had been filled by a distractor task.

At this point it should not come as a surprise that the function words counted by LIWC are fertile territory for mimicry. They are so easily mimicked, in fact, that mimicry of function words as measured by LIWC has its own name: Linguistic Style Matching (LSM). LSM is measured by comparing LIWC counts on all function word categories between both partners in an interaction, either taking the interaction as a whole or splitting it into parts to track changes in LSM over time. This comparison can be done through product-moment correlation [71] or a weighted difference score [51].

The LSM metric, as measured through weighted difference scores, is internally consistent: if a dyad matches in their use of one function word category, they will probably match to the same degree with all others. It also generalizes well across

different contexts, from online chat conversations [71] and class assignments to the real-life letters and poetry of well-known colleagues and spouses [51] and transcripts of taped conversations between political allies [71].

Just like other function word metrics, LSM appears to a greater or lesser degree in different circumstances. Some linguistic styles are more easily matched than others, and different people will match a given style with more or less ease. In general, the lower-status person in a dyad goes to greater lengths to match the higher-status person than vice versa [51]. Thus, in a working group of several people, those of higher social status do less matching than others [91]. Women match more than men, people high in social skills match more than people low in them, and neurotics do less matching than others [51]. In a task based on a class assignment, students with higher grades and whose parents were more educated matched more [51].

A greater degree of LSM would appear to be a good thing. The correspondence and poetry of famous couples, analyzed over a period of years, consinstently shows more LSM when things are going smoothly and less LSM when there is conflict in the relationship [51]. Higher LSM is associated with future stability of a romantic pairing [52], greater cohesion during group work [91], and even (if the high LSM is stable throughout the interaction) success in hostage negotiations [93].

However, some caveats apply. The causal direction of these associations cannot be unequivocally determined, but mimicry in general is unconscious, and deliberate attempts to increase LSM have no effect [51], implying that LSM is a symptom of social cohesion rather than a tool for increasing it. As well, LSM scores are occasionally high even between strangers in an Internet chat room. Niederhoffer and Pennebaker,

who invented the LSM metric, suggest that it does not signify any particular emotion two people feel for each other, but rather a pattern of mutual influence and engagement [71].

2.8 Bivariate Distributions

We have now described all the psycholinguistic underpinnings of our work. However, we have not yet described all the mathematics we are going to use in these investigations. When looking at the relationships between different word categories, we are going to be using bivariate distributions.

To describe any set of potentially correlated real-valued variables, one often uses a multivariate distribution (bivariate in the case where there are 2 variables). Bivariate distributions can be mathematically complex. The most commonly used one is the bivariate Gaussian or normal distribution. The Gaussian distribution is a "bell curve" shape which (in univariate form) is symmetrical and unipolar, and is described entirely by its mean and standard deviation. Thus, the Gaussian distribution is comparatively easy to work with. The bivariate Gaussian distribution represents two potentially correlated variables whose marginal means are both Gaussian. It is described by a mean in two dimensions and by a covariance matrix which serves as the two-dimensional equivalent of a standard deviation, describing the distribution's length, width, and angle in Cartesian space.

However, there is no reason to suppose that word frequency data from a bag-of-words model should be normally distributed. This data tends to be highly skewed and is almost never symmetrical, as a Gaussian distribution would require. Indeed, most word frequency categories seem at first glance to be closer to an exponential

distribution, with values of 0 very common (but without any negative values) and higher values turning up every so often. However, some words—particularly the most common function words, such as "I"—do have a probability density peak above 0. That is, they appear more often than they fail to appear, which makes an exponential distribution inappropriate.

A promising distribution for words of this kind is the gamma distribution: a generalization of the exponential distribution that does not necessarily peak at 0. Yue [101] uses a distribution of this kind successfully to fit weather data.

2.9 Random Forests

We will also be using random forests in later chapters. Random forests are a state-of-the-art predictive modeling technique developed by Breiman [6].

Random forests function as an ensemble of tree predictors. A tree predictor, or decision tree, is a model that recursively splits the data into a number of categories. The tree is given training data with a number of independent variables and two or more prediction classes. At each split, the tree attempts to make some decision based on one of the independent variables that helps it classify the data into prediction classes. For example, if all the data with value x in variable Y belong to class C, the tree will "split" on value x of variable Y, putting everything with that value into one category. That category does not need to be split anymore, since everything in it belongs to class C. Once a decision tree is trained, it can be tested with new data, and will classify this new data based on the "splits" it learned from the training data.

However, single decision trees are prone to overfitting. They can make splits based on coincidences in the data which will not generalize to other data of the same

type. Random forests present a solution to this problem by making a great number of trees based on different bootstrapped samples of the data and different subsets of the available variables. Coincidental overfitting may still occur in each tree, but the coincidences in each tree will be different due to a slightly different sampling. Once the trees are all trained, they can be tested with new data. The data will be given to every tree, and the trees will "vote" on the category in which the data belongs. The different errors of the different trees cancel each other out and what is left is a "vote" which is much more likely to correspond to a real underlying property of the data.

Besides this robustness to overfitting, random forests also have the advantage that their internal structure is intelligible and they can give useful internal estimates of their own properties. For example, the importance of different variables in the data can be estimated by counting the number of "splits" that are made, throughout the forest, on each variable.

Chapter 3

Method

Because of lingistic style matching, we expect that the word-based signs of deception in a dialogue will be different from those in a monologue. In other words, people do not have a completely free choice of words when responding to another person's questions. In particular, we expect to see a "prompting" effect: a change in frequency in one word category in the question will produce a change in frequency in a grammatical word category in the answer.

Note that the prompting effect, as we have defined it, is a generalization of LSM. The big conceptual change is that we do not think style matching is limited to repeating words in the same category. Rather, some categories of words necessarily prompt the respondent to use *different* categories.

This is particularly obvious with pronouns. When asked a question using the structure, "Did you...?" it is natural to respond with "Yes, I..." or "No, I..."—responding to the presence of second-person pronouns, not with a matched number of second-person pronouns, but with first-person singular pronouns.

This leads us to our first hypothesis:

H1. The function words in an answer to a question are not independent of the question, but are prompted (positively or negatively) by the function words in the question. At least some of the words affected in this way will be relevant words, such as first person singular pronouns, from the LIWC deception model (and since we are studying deception, these are the words we will be testing).

Our second hypothesis follows:

H2. If word frequencies from the LIWC model are affected by the words in the question, then this is an influence that potentially distorts the results of the LIWC model. Removing the effects of the question from the answer before applying the LIWC model will result in greater accuracy.

Our plan of attack has three parts. First, we gather archival question-and-answer data and visualize any relevant changes in frequency that could be caused by H1. Second, we attempt to correct for these changes. Third, although not every answer in our data could be labeled as all true or all a lie, we use some rough and preliminary methods to validate our correction and to ensure that it was changing the data in a way that was consistent with a stronger distinction between truth and deception.

3.1 Data Sets

Instead of conducting our own experiments to gather data, we created three data sets by taking archival transcriptions of real-life, high-stakes question-and-answer interactions.

First, the REPUBLICAN data set comprised transcripts of each of the Republican primary debates leading up to the American presidential election of 2012. There were twenty such debates in various American cities leading up to the primary elections,

involving a total of ten candidates, and each one was televised and transcribed online by various news organizations [3, 10–13, 18–24, 27, 39, 49, 68, 78, 80, 98]. At each debate, one or more moderators (and sometimes volunteers from the audience) asked questions about politics to each candidate. Not all candidates were present at all debates.

We used the REPUBLICAN data set to detect overall patterns of interaction between question and answer words. We did not rate the Republican presidential candidates as deceptive or non-deceptive, nor did we use this data set in validating our methods, since we lack access to ground truth about the candidates' honesty. Fact checking websites may show that specific statements are true or false, but they do not help with the issue of persona deception as defined by Gupta and Skillicorn [44]: canditates will "spin" the facts to present theselves in a favourable light which does not necessarily correspond to their real self-image. Since there is no equivalent of a fact checking website for persona deception, there is no reliable way to judge one candidate overall as more deceptive than another. However, we did judge the debates in general as a forum in which all candidates would be at least moderately motivated towards persona deception. Presenting themselves and the facts in the most favourable possible light is essential to getting elected, and most of the time this favourable light does not correspond exactly with reality.

The full REPUBLICAN data set contained 2118 question-answer pairs and 301,539 total words.

Second, the NUREMBERG data set comprised selected examinations of witnesses from the Nuremberg trials of 1945-1956. Most of these examinations were taken from the Trial of German Major War Criminals, transcribed at the Holocaust memorial website nizkor.org [48]. Two were instead taken from the Nuremberg Medical Trial,

which is partially transcribed online at the website of the Harvard Law Library [67]. We included both direct examinations and cross-examinations—the full list of questions put to every witness in the data set and their answers.

Unlike the REPUBLICAN data set, the NUREMBERG data set contained obvious subgroups with markedly different motivations towards deception.

The first group, DEFENDANTS, contained two Nazi war criminals testifying in their own defense who were eventually found guilty on all counts and executed. These men were highly motivated towards deception: they were guilty, and their lives depended on convincing the tribunal that they were not. Many such defendants were questioned at the Trial of German Major War Criminals, but each defendant's testimony was much longer than the testimony of any other witness. In order to keep the size of the first group at least roughly commensurate with the size of the others, we stopped collecting data for the DEFENDANTS group after recording the testimony of the first two defendants.

The second group, UNTRUSTWORTHY WITNESSES, contained ten lower-ranking Nazis who were not themselves on trial. While this group was not at immediate risk of conviction, it seems reasonable to suppose that their accounts would be moderately deceptive. Most of them would be motivated to absolve themselves either by minimizing Nazi war crimes as a whole or by minimizing their own involvement.

The third group, TRUSTWORTHY WITNESSES, contained nineteen survivors of Nazi war crimes who testified about those crimes. Fourteen of these were Holocaust survivors, while the other five were civilians who reported on more general conditions in Nazi-occupied countries. We did not consider any of these witnesses deceptive.

Some witnesses spoke at great length about their experiences, resulting in very

long answers, so we chose to truncate the answers in all subgroups of the NUREMBERG data set at 500 words. This brought maximum window size into line with the maximum size that occured naturally in the REPUBLICAN data set, and it affected less than 1 percent of the answers.

We used the NUREMBERG data set to validate our model. Once we found patterns in the REPUBLICAN data and worked out a means of correcting for them, we performed this correction on the NUREMBERG data set as well. We hypothesized that the correction would increase the ease of differentiating TRUSTWORTHY WITNESSES from other groups.

The full NUREMBERG data set contained 4159 question-answer pairs (1355 from DEFENDANTS, 1826 from UNTRUSTWORTHY WITNESSES, and 978 from TRUSTWORTHY WITNESSES). It contained a total of 311,099 words.

Finally, we made brief use of a third data set, SIMPSON. This contains depositions from the civil trial of O.J. Simpson for the wrongful death of his wife, Nicole Brown, and another man. Since the Simpson trials received intense press coverage at the dawn of the Internet age, extensive transcripts of his case compiled by interested viewers were available online [90]. SIMPSON contains Simpson's deposition in his own defense, which we considered deceptive, as well as the depositions of family and friends of the deceased, which we considered largely truthful. Other depositions, such as those of Simpson's personal friends, were not included. (We chose the civil trial rather than the criminal trial because it was the first time Simpson testified directly in his own defense; it also had the advantage of being a trial at which he was found guilty.) Since Simpson's own deposition was three times as long as the rest of the data set, we used only every third question-answer pair from him.

We used the SIMPSON data set for further validation after analysis of NUREMBERG. It contained 20,810 question and answer pairs, totalling 412,385 words.

In addition to these three data sets, some of our earlier analysis involved short transcripts of celebrity and politician interviews.

3.2 Data Set Parsing and Model Words

When processing this data, we had to decide on a level of granularity. We were most interested in what occurs on the level of a single question and answer pair. Thus, for us, a "window" was a single question followed by its answer.

To store these windows, except where otherwise noted, we used a pair of data structures. The first structure recorded the names of the questioner and respondent for each window, and the words used in the question and answer. As is typical for bag-of-words models, we removed all punctuation, capitalization, and other information besides the words themselves. We did not perform any stemming. An example is shown in Table 3.1.

The second data structure recorded the length (in words) of each question and answer, along with the count of the number of words belonging to certain categories:

FPS First person singular pronouns. The LIWC model predicts that deceptive people should use a lower rate of first-person pronouns than truthful people. However, in the Gomery commission data, Little and Skillicorn [60] found that deceptive people used a higher rate.

but Lower rates of this and other exclusive words are expected in deception. The LIWC model puts "but" and "or" in the same "exclusive words" category but

WALLACE	thank **you** speaker gingrich one of the ways that **we** judge a candidate is the campaign **they run** in june almost **your** entire national campaign staff resigned along with **your** staff here in iowa **they** said that **you** were undisciplined in campaigning and fundraising and at last report **youre** a million dollars in debt **how** do **you** respond **to** people **who** say that **your** campaign has been a mess so far
GINGRICH	well let **me** say first of all chris that **i took** seriously brets injunction to put aside the talking points and **i** wish you would put aside the gotcha questions like like ronald reagan who had 13 senior staff resign the morning of the new hampshire primary and whose new campaign manager laid off 100 people because he had no money because the consultants had spent it like john mccain who had to **go** and **run** an inexpensive campaign because the consultants spent it **i** intend to run on ideas congress should come back monday they should repeal the doddfrank bill they should repeal sarbanesoxley they should repeal obamacare they should institute lean six sigma across the entire federal government a hard idea for washington reporters to cover **but** an important idea because its the key to american manufacturing success

Table 3.1: Verbal component of a sample window from the REPUBLICAN data set. Words from Newman *et al.*'s deception model in the answer are bolded; all words from the lengths-and-counts data structure are bolded in the question.

Little and Skillicorn's SVD study [60] suggests that the two words have quite different properties in semantic space. We therefore counted "but" and "or" separately.

or Another exclusive word.

excl Other exclusive words, such as "unless" and "whereas".

neg Negative emotion words. Higher rates of these are expected in deception.

action Action verbs. Higher rates of these are expected in deception.

FPP First person plural pronouns. These can be substituted for singular in some circumstances (the "royal we") and a questioner using the "royal we" might encourage an respondent to do likewise. Similarly, if the questioner is not using the "royal we" and is referring to a group that they and the respondent both belong to, this might prompt the respondent to continue talking about this group. Focusing on a group leaves less time to focus on oneself, so we expected higher FPP rates in the question to prompt lower FPS rates in the answer.

SP Second person pronouns. A question containing the word "you" is probably about the respondent, and the respondent should be expected to respond by giving information about themselves. Thus, we expected higher SPP rates to prompt higher FPS rates in the answer.

"Wh" words Who, what, when, where, why, and how. Questions containing these words prompt the respondent for a specific fact. Questions about specific facts should make it harder for the respondent to be evasive. We expected higher

	Len	FPS	but	or	excl	neg	act	FPP	SP	wh	TPS	TPP	these
Q	71	0	0	0	0	0	1	1	7	2	0	2	1
A	141	4	1	0	0	0	4	0	1	2	1	4	5

Table 3.2: Lengths and word counts of the sample window from Table 3.1. Length is in the first column, followed by word counts in each of the relevant word categories.

"wh" word rates to prompt higher rates of exclusive words due to an increase in cognitive complexity.

TPS Third person singular pronouns. These words are references to a person other than the questioner or respondent. Being asked about a third-person should induce the respondent to talk about that third-person, and thus give them less opportunity to talk about themselves. When talking about another person—not oneself, and not the questioner—it should also be easier to make disparaging or negative remarks. We expected higher TPS rates to prompt lower FPS rates and higher negative emotion rates.

TPP Third person plural pronouns. We expected higher TPP rates to prompt lower FPS rates and higher negative emotion rates, for the same reason as above.

"These", "those", and "to" An early analysis with a part-of-speech tagging program showed that rates of these words in the question were weakly correlated with rates of FPS in the answer. This early analysis did not yield other interesting results.

Table 3.2 shows an example of this lengths-and-counts data structure for the same window shown in Table 3.1.

3.2.1 Dealing With Window Sizes

A 500-word answer with five action words is statistically and linguistically different from a 15-word answer with five action words. Therefore, using raw word counts in our statistical analysis would be inappropriate. For all our analyses, we divided the number of words of each type in each window (a "window" here meaning a single question or answer) by the total number of words in the window, giving a rate statistic for each type of word.

We expected the shortest questions and answers to be difficult to analyze. In a five-word answer with one "but", the rate statistic for "but" would be 0.2—unusually large. But it isn't clear that the answer really has all the properties associated with use of the word "but" to a greater degree than, say a 16-word answer with one "but" (which would have a rate of only 0.0625).

For our initial analysis, we chose a minimum window size of 50 and discarded all windows in which the question *or* its answer contained fewer than 50 words. We chose this minimum size because of some interesting results at an earlier, more exploratory stage in the research. At this early stage, we wanted to know if there was a movement through semantic space characteristic of questions, answers, or the transition from a question to its answer. As a result we worked with short overlapping windows within a question and answer. We varied both the size of the windows and the size of the overlaps.

One measurement we made at this early stage was "internal coherence", or the average dot product of the vectors in semantic space leading towards and away from each window within a single question or answer. We found that varying the size of the overlap produced a striking effect: coherence rose rapidly as the overlap increased

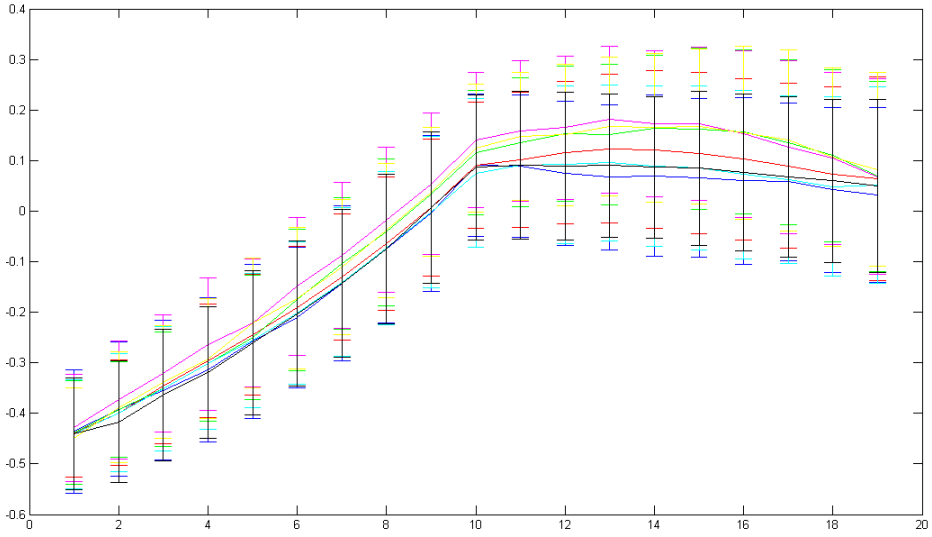


Figure 3.1: Internal coherence (Y-axis) of several small data sets at window size 20 with varying overlap sizes (X-axis). The data sets pictured include politician [38, 63] and celebrity [7, 16] interviews on news shows, excerpts from the NUREMBERG [67] and REPUBLICAN [80] data sets, and the testimony of a witness from the O.J. Simpson criminal trial [90]. All exhibit the same basic pattern.

from 0, then suddenly leveled off when the overlap reached 50% of the current window size. This effect, which we call the 'halfway effect', is visible in Figure 3.1.

The halfway effect is visible at various small window sizes and in many different data sets, but as window size increases, it becomes less pronounced and is finally replaced by a smooth curve at a window size of about 50 (as shown in Figure 3.2).

These figures suggested to us that a window size of 50 was big enough to be "stable" and safe from the sort of conceptual problem that arises with five-word answers. But a model that only works on windows of 50 words or larger is not as

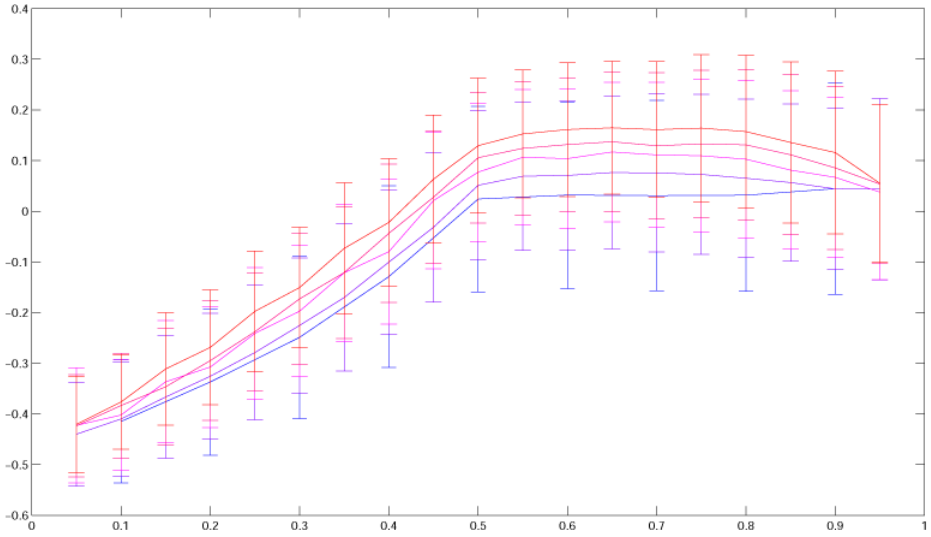


Figure 3.2: Internal coherence (Y-axis) of a single Republican debate [68] at various window sizes. The X-axis is overlap size as a percentage of window size. The window sizes are represented by a continuum of shades from 10 (blue) through 20, 30, 40, and 50 (red).

useful as we would like, because in many settings the average window is much smaller than that. For example, in the SIMPSON depositions, the average question and answer pair contained fewer than 20 words in *both* sides of the window put together. So we looked for a way to apply our model to smaller windows.

After creating the initial model and correction technique with the REPUBLICAN data set, we tried it again on the same data set with smaller minimum window sizes—30, 10, and 1. At each window size, we measured the average change induced at each stage of the correction at each window size and compared it to the average change at a minimum window size of 50. We also measured the percent of words in the original

data set that were represented in windows at each minimum window size.

We also tried lumping adjacent questions and answers together. If a window was smaller than the minimum size, we merged it with an adjacent window with the same respondent, adding the adjacent answer to the answer and the adjacent question to the question, creating a new composite window in place of the old ones. We repeated this until all windows either met the minimum size for both question and answer, or could not be merged with adjacent answers by the same person because there were none. We then removed any windows that still did not meet the minimum size.

With this method, as well, we tried our correction technique, measured the average change induced at each stage of the correction, and compared it to the average change in the original, non-composite data.

As we will further explain in Chapter 4, we found that the best way of dealing with window sizes was with the composite windows, lumping questions and answers together until they reached a minimum size of 50. This provided the best compromise, allowing us to cover large numbers of question-answer pairs while giving results reasonably close to those of the original 50-word-minimum windows.

3.3 Gaussian Distributions

We wished to find differences between prompted and unprompted answers, so we split each pairing of a question word and an answer word into two parts. The "unprompted" part contained all answers to a question that did not have any of the relevant question words. In this data, it could be assumed that a prompting effect related to the question words did not exist. The "prompted" part, in which prompting could potentially exist, contained all answers to questions in which the relevant

question words did appear at least once.

For each pairing of a question word with an answer word, we counted all the rates for each in all windows and fitted a two-dimensional Gaussian distribution to the "prompted" data. We did the fitting using the FitFunc toolbox in MATLAB. We then evaluated the "prompted" data by comparing it to the "unprompted" data.

These Gaussian distributions provided a broad indicator as to the relationship between the question word and the answer word. A distribution with a positive slope and a mean higher than that of the "unprompted" data indicated that the question word prompted higher rates of the answer word. A distribution with a negative slope and a mean lower than that of the "unprompted" data indicated that the question word prompted lower rates of the answer word. A flat or near-spherical distribution with a mean close to that of the "unprompted" data suggested that there was no relationship between the two word categories.

We also used the FitFunc toolbox to produce Gaussian mixture models using Maximum Likelihood Estimation. These contained two or more Gaussians which together provided an explanation for the data. We suspected that, if there were two underlying processes—the respondent's mental state, and the prompting effect of the questions—they might appear as two different Gaussians in a mixture model. We did not create mixture models unless there were enough nonzero points in the data to justify them (at least 30 points per Gaussian)—although the nonzero points were not excluded from the data set, because a non-response to a strong prompt might be significant, and so it was possible for models to appear with distributions that largely represented non-responses. In addition, the FitFunc toolbox sometimes produced one or more distributions explaining a small fraction of the data (below 5%). Mixture

models of these types were discarded.

3.4 Gamma Distributions

In addition to Gaussian distributions, we also tried fitting gamma distributions to the data.

Since MATLAB did not come with functions for two-dimensional gamma distribution mixture models, we built our own. We used the two-dimensional gamma distribution from Yue [101] and created an Expectation Maximization function in MATLAB around the distribution's five variables. The EM function was capable of performing with either a single two-dimensional gamma distribution or a mixture model containing several such distributions.

During the E step, we calculate the probability density for each distribution at each point in the data using the PDF from Yue [101]:

$$f(x,y) = \begin{cases} \frac{K_1}{K_2} \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} c_{jk} (\beta_x x)^j (\eta \beta_y y)^{j+k} & \text{if } \rho > 0 \\ f_x(x) \cdot f_y(y) & \text{if } \rho = 0 \end{cases}$$

$$(x, y \geq 0, 0 < \eta < 1, \gamma_y \geq \gamma_x, \text{ and } 0 \leq \rho < \eta\sqrt{\gamma_x/\gamma_y})$$

where

$$K_1 = (\beta_x x)^{\gamma_x - 1} (\beta_y y)^{\gamma_y - 1} \epsilon(-\frac{\beta_x x + \beta_y y}{1 - \eta})$$

$$K_2 = (1-\eta)^{\gamma_x} \Gamma(\gamma_x) \Gamma(\gamma_y - \gamma_x)$$

$$c_{jk} = \frac{\eta^{j+k} \Gamma(\gamma_y - \gamma_x + k)}{(1-\eta)^{2j+k} \Gamma(\gamma_y + j + k) j! k!}$$

$$\eta = \rho\sqrt{\gamma_y/\gamma_x}$$

in which β_x and β_y are size parameters in each dimension, γ_x and γ_y are shape parameters in each dimension, and ρ is a correlation parameter.

While the infinite double summation in $f(x, y)$ might appear intractable, in practice it always converges. The term c_{jk} contains factorial terms both in its numerator and its denominator, since the gamma function is a modified factorial, but the term $\Gamma(\gamma_y+j+k)j!k!$ in the denominator will always be greater than the term $\Gamma(\gamma_y-\gamma_x+k)$, and as either j or k increases, the ratio between the numerator and denominator increases factorially. The other terms in the summation increase only exponentially. Thus, over time the double summation will reach an asymptotic boundary as the items being summed become incredibly small. Our function stops calculating the double summation when the amount being added at each step, compared to other amounts at previous steps, becomes too small.

This gives us a probability density value for each point and each distribution in the model.

In the M step, we re-estimate the parameters of each distribution. As in Yue [101], size and shape parameters for each dimension are estimated from the parameters of the marginal distributions, as follows:

$$\beta = \frac{M}{S^2}$$
$$\gamma = \frac{M^2}{S^2}$$

where M and S are the mean and standard deviation of the data, respectively. Unlike Yue [101], we are dealing with mixture models involving more than one gamma distribution, so we weight the calculation of both M and S according to the probability density at each point in the given distribution. ρ is likewise re-estimated as a weighted correlation between the two marginal distributions.

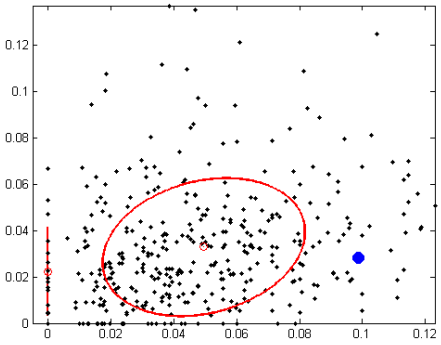
Note that, since the data itself does not change, the only cause for a change in this estimation is a change in the relative weights of each point in each distribution. Therefore, in the case of a model with a single gamma distribution, the EM algorithm converges in a single step, which makes it not really EM. However, this method of estimation is equivalent to Yue's [101] method, which achieved good results. In gamma mixture models, we initialized the algorithm with random parameter values and allowed it to run until it passed either below a minimum change in the weights or above a maximum number of iterations.

We did not run gamma mixture models on data with less than $30n$ nonzero points, where n equals the number of distributions in the model. We also discarded mixture models containing more than one distribution with identical parameters.

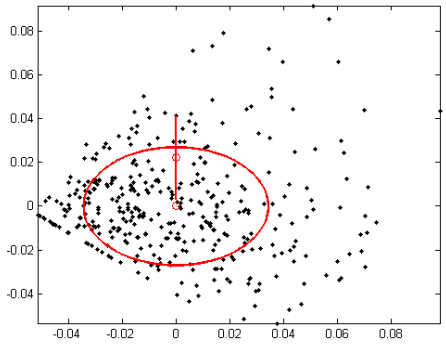
3.5 Data Correction

Of course, we did not want only to model the relationship between question words and answer words: we wanted to remove the influence of questions from answers in order to better analyze the answers. With the Gaussian-based models, we could do this quite easily by performing a series of affine transformations on the points in the model—producing corrected rates for each word category.

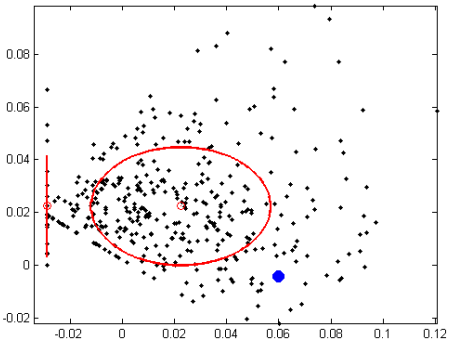
The exact correction method is illustrated in Figure 3.3. For each question-answer pair, we found the single distribution Gaussian model of the prompted data and translated it so the mean was at (0,0). We then rotated the distribution using a standard rotation matrix until it was "flat" (one of its axes lay along X=0). We used the smallest possible angle of rotation that would achieve this end, in either direction. (This angle can be found straightforwardly using eigenvectors and eigenvalues of the



(a) The original data.



(b) Steps 1 and 2: Translated to a mean of (0, 0) and rotated.



(c) Steps 3 and 4: Rescaled and translated back to a mean that matches unprompted data.

Figure 3.3: Steps in the correction process illustrated using second-person pronouns in the question and first-person singular pronouns in the answer. The X-axis shows rates for the question words and the Y-axis shows rates for the answer words. The red line on the Y-axis shows the one-dimensional "unprompted" distribution to a distance of 1 standard deviation in each direction from the mean.

distribution's covariance matrix.)

After that, we rescaled the data on the Y axis so that its standard deviation in the Y direction was equal to the standard deviation of the "unprompted" data, and we translated it again so that its mean was back in its original place in the X direction, and equivalent to the mean of the unprompted data in the Y direction.

We have included a large blue dot in Figure 3.3 representing the example window from Table 3.1, so that the position of this question and answer pair can be traced through each step of the process. Observe that, in the uncorrected data, Gingrich is heavily prompted and responds with a number of first-person singular pronouns that is about average for the data as a whole, but much less than what might be expected given the general trend towards more first-person singular pronouns with more prompting. After the data is rotated and moved back into place, Gingrich's "corrected" first-person singular pronoun rate is near 0.

Note that after this correction, some windows were assigned values slightly less than 0. Obviously it is impossible for a person to say fewer than zero words in response to a question. We treated the negative values as very emphatic zeroes, but did not correct them to 0 until the end of the process, so as to reduce noise from repeated corrections.

We performed these transformations for each question-answer pair in the data, feeding the input from one stage of transformation into the input of the next, so that for each response word, a correction was made for each stimulus word in turn. For question-answer pairs where the question word had negligible effect on the answer word, the effect of this correction would also be negligible, so we performed it for every question-answer pair except those without sufficient nonzero prompted points.

There was a risk of these negligible effects producing slight noise in the data, but we accepted this risk because we did not wish to choose an arbitrary threshold separating effects that "counted" from effects that "did not count"; we were doing exploratory work and wished to work with all possible interactions in the model.

We also separately performed a correction (up to, but not including, the translation back to the new mean) based on the possibility of multiple-Gaussian mixture models. For this, we had to add a step to the algorithm. For each distribution j in the model, we made a copy of the original points p and performed the correction according to the properties of that distribution. This provided points p_{jk}, where k is the number of prompted windows in the data.

We then created points p_k by taking the weighted mean of $w_1 p_{1k}, w_2 p_{2k}, \ldots w_j p_{jk}$ where w_j is the weight of point p in distribution j. In other words, if a point was best explained by a particular distribution, it would be moved mostly based on the properties of that distribution, and if a particular distribution had negligible probability density at a given point, it would not be taken very much into account while processing that point. Finally we translated all points p_k to the new mean just as though they were points from a single-distribution correction, and repeated the process for each question-answer pair.

We began to devise similar methods for performing a correction based on gamma distribution models, but stopped before completing them. This is described in the "Results" section.

For our initial verification of this correction method, we produced figures of the results and inspected each one of them visually. For each question-answer pair in each data set, we drew a scatter plot in MATLAB of all the windows and overlaid the shape

of the Gaussian distribution(s) on them. We did this both before and after each stage of the correction. We looked at each figure and confirmed that the correction method was producing Gaussians parallel to the X-axis and level with the unprompted data, and that it was using the smallest applicable rotation (for example, turning the data 27 degrees instead of 117, or -9 instead of 81).

3.6 Validation

Verifying that our correction method produced Gaussians of the right shape was a good start, but it did not demonstrate that this shape reflected an actual improvement in the data. Ideally, we wanted the corrected data to show a sharper distinction than the original data between truth-tellers and liars.

We first tried a validation method based on Little and Skillicorn's SVD-based method [60]. Before and after performing the correction on the NUREMBERG data set, we performed singular value decomposition on the answer data, reducing the 6-category deception model to 3 dimensions. As Little and Skillicorn did, we zero-centred the data for normalization.

We then made a scatter plot showing each window in the data set semantic space. As in Little and Skillicorn [60], we drew a line corresponding to the square roots of the first three singular values, which could be considered roughly as an "axis of deception". We colour coded each point in the scatter plot, with a different colour for each group within the data set. While not every statement by an UNTRUSTWORTHY WITNESS or DEFENDANT is a lie, we suspected that, in general, these groups' statements would appear higher on the scale than those of TRUSTWORTHY

WITNESSES. We hypothesized that performing our correction would produce a visible increase in separation between the three groups. We also performed a similar visualization with the SIMPSON data set.

When this validation method gave unexpected results we performed a more detailed review of the corrected and uncorrected data, but this is difficult to explain in detail without referring to the unexpected results. It is therefore described in Chapter 4.

Chapter 4

Results

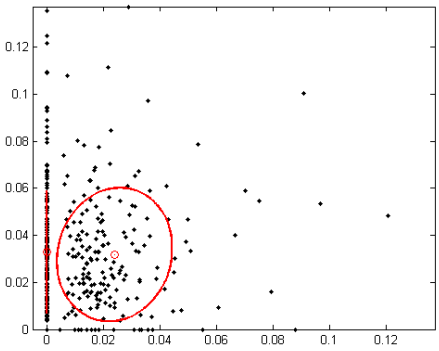
4.1 Gaussian Distributions

Gaussian distributions for the REPUBLICAN data showed a discernible prompting relationship for many question-answer pairs. Our initial guesses about the question word categories had mixed success.

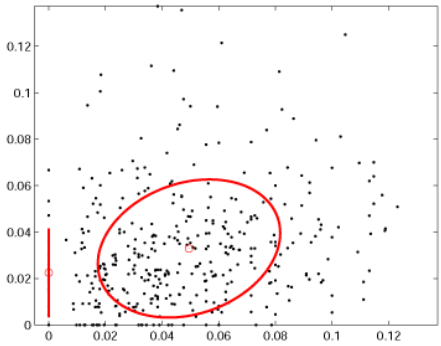
- We predicted that higher first-person plural pronoun (Figure 4.1a and third-person singular pronoun rates in the question (Figure 4.1d would prompt lower first-person singular pronoun rates in the answer. We also predicted that higher "wh" word rates would prompt higher exclusive word rates (Figure 4.1c). None of these actually occurred at any significant level.

- We predicted that higher third-person pronoun rates would prompt higher negative emotion rates. Due to a generally low level of negative emotion words in the data, we lack results on this prediction.

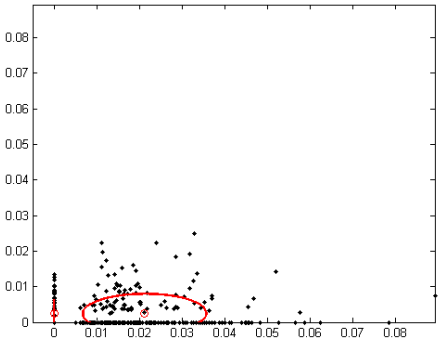
- We predicted that higher second-person pronoun rates in the question would



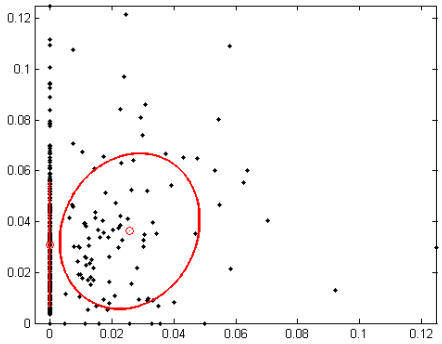
(a) First person plural pronouns prompting first-person singular pronouns



(b) Second person pronouns prompting first-person singular pronouns

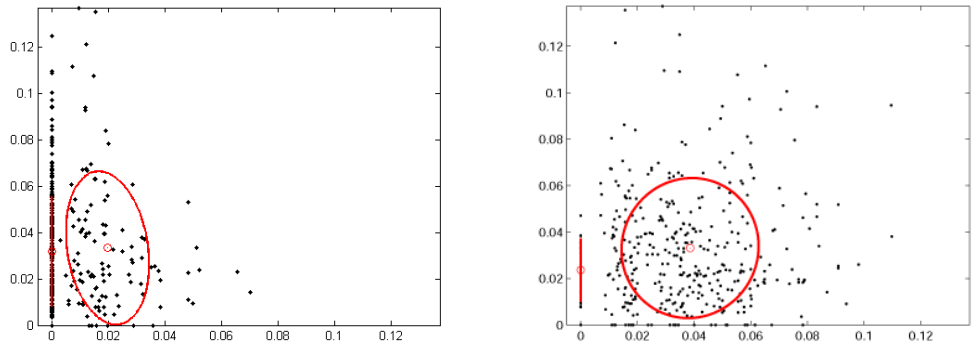


(c) "Wh" words prompting "or". The other exclusive word categories looked similar.



(d) Third person singular pronouns prompting first-person singular pronouns.

Figure 4.1: Gaussian distributions from the REPUBLICAN data set with a minimum window size of 50 words. The X-axis is rates for the answer words and the Y-axis is rates for the question words. The red line on the Y-axis shows the one-dimensional "unprompted" distribution to a distance of 1 standard deviation in each direction from the mean.



(a) Third person plural pronouns prompting first-person singular pronouns.

(b) "These", "those", and "to" prompting first-person singular pronouns.

Figure 4.2: Gaussian distributions from the REPUBLICAN data set with a minimum window size of 50 words. The X-axis is rates for the answer words and the Y-axis is rates for the question words. The red line on the Y-axis shows the one-dimensional "unprompted" distribution to a distance of 1 standard deviation in each direction from the mean.

prompt higher first-person singular pronoun rates in the answer. This did occur, leading to a prompted Gaussian distribution with a positive slope and a higher mean than the unprompted distribution (Figure 4.1b).

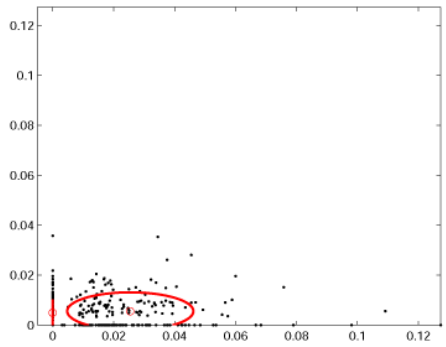
- We predicted that higher rates of "these", "those", and "to" in the question would prompt higher first-person singular pronoun rates in the answer. This was also correct: the prompted mean and range were higher than the unprompted mean and range (Figure 4.2b).

- We predicted that higher third-person plural pronoun rates would prompt lower

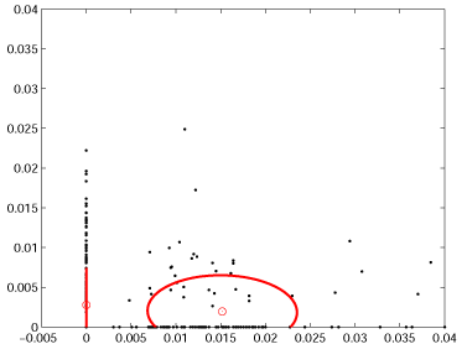
first-person singular pronoun rates. Instead, we found a weak effect in the opposite direction: higher third-person plural pronoun rates prompted very slightly higher first-person singular pronoun rates (Figure 4.2a).

Most other pairs of question and answer words also showed no relationship, but some prompting effects, such as third-person plural pronouns prompting action words, appeared even though we did not expect them (Figure 4.3). Miscellaneous exclusive words (other than "but" and "or") and negative emotion words were generally not numerous enough for distributions to be created. Note that, due to space limitations, not all relationships between pairs of words are shown here.

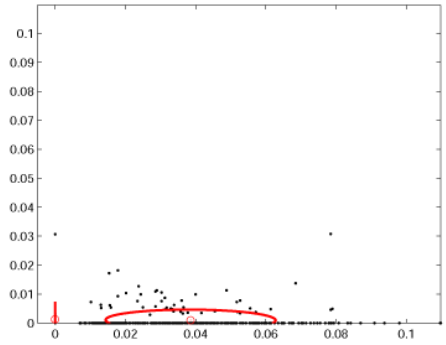
With mixture models, we had speculated that two-distribution models might be plausible: there could be one distribution describing non-responses, for example, and one describing responses, or there could be two different levels of prompting which produced different results. However, the actual function fitting did not favor two-distribution models. Instead, the question-answer pairs produced models with as many distributions as they were allowed, given the constraints of the number of windows. There were few telltale changes in shape or density to support these mixture models on a theoretical level; they simply seemed to be arbitrarily partitioning the data. Figure 4.4 shows an example of this process. Five question-answer pairs had enough points to be partitioned into four or more Gaussian distributions. Most of these distributions did not look plausible to us. The 1-Gaussian and 2-Gaussian distributions each looked somewhat plausible, but there was no clear reason to choose the 2-distribution model over the 1-distribution model. Using our correction method on mixture models also changed the data in a more unpredictable way. Since we could not justify this change, we did not continue serious work with Gaussian mixture



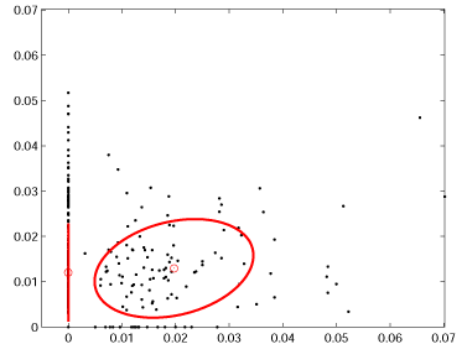
(a) First person singular pronouns prompting "but"



(b) "But" prompting "or"



(c) "These", "those", and "to" prompting negative emotion words



(d) Third person plural pronouns prompting action words

Figure 4.3: Further Gaussian distributions from the REPUBLICAN data set with a minimum window size of 50 words. The X-axis is rates for the answer words and the Y-axis is rates for the question words. The red line on the Y-axis shows the distribution of the one-dimensional "unprompted" distribution to a distance of 1 standard deviation in each direction from the mean.

models.

4.2 Gamma Distributions

In scatter plots of the data from many of the most interesting windows, a roughly triangular shape emerged in which the highest values of Y occurred at moderately low, but not extremely low values of X. Gamma distributions fit this shape quite well (Figure 4.5). However, working with these distributions proved difficult.

First, not all question-answer pairs had a properly fitted gamma distribution. Of the "response" words, only first-person pronouns and action words had enough nonzero prompted points (compared to the number of prompted non-responses) to produce a distribution like this. In all other groups, the number of prompted non-responses was relatively large and caused the shape parameter to slip below 1. With a shape parameter below 1, these gamma distributions are no longer useful: the probability density at one axis rises many orders of magnitude above the probability density at any of the points we are actually interested in. This results in essentially a linear distribution which is useless for explaining anything (Figure 4.6).

Gamma mixture models, meanwhile, worked somewhat better than Gaussian mixture models. They had the same size limit as Gaussian mixture models (at least $30n$ points, where n is the number of distributions), but our EM program did not continue to make mixture models all the way up to this limit. They almost universally stopped at two distributions. Unfortunately, these two-distribution models were not very interesting: the majority simply combined a distribution for non-responses and a distribution for responses (Figure 4.7).

It was also unclear how to correct gamma distributions properly. Because of their

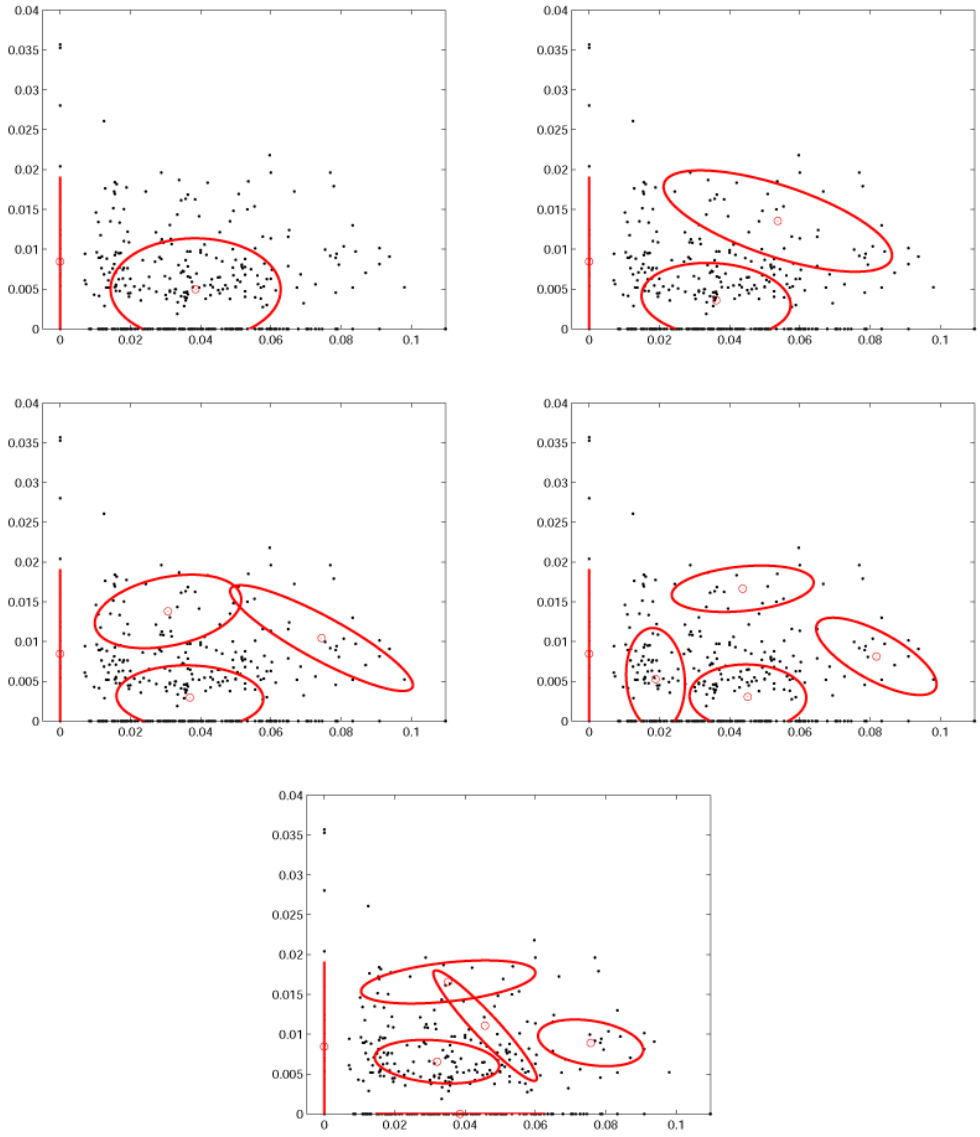
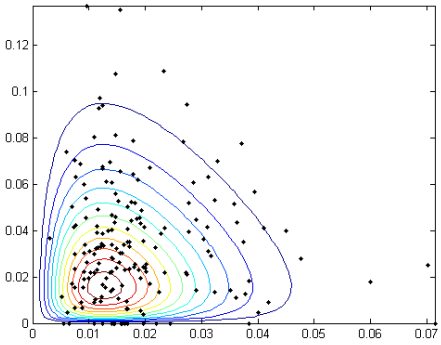
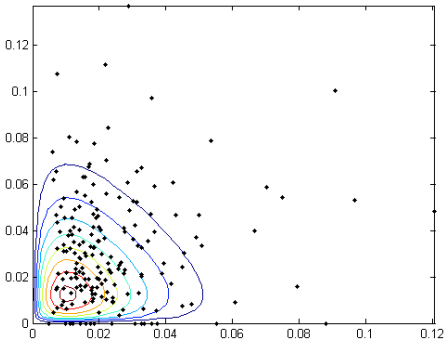


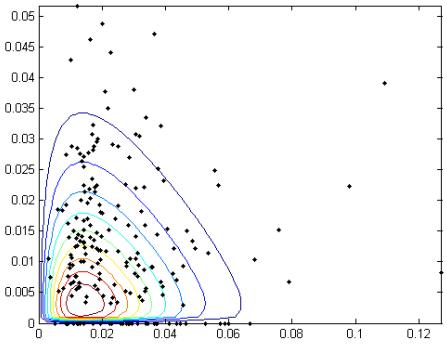
Figure 4.4: Five different Gaussian mixture models on the same question-answer pair ("these", "those", and "to" prompting "but"). The X-axis is rates for the answer words and the Y-axis is rates for the question words. The scale and underlying data points are exactly the same in each model; the only difference is the number of Gaussian distributions (1, 2, 3, 4, and 5). Note that the 5-distribution model contains a distribution that lies along the X-axis.



(a) Action words prompting first-person singular pronouns



(b) First person plural pronouns prompting first-person singular pronouns



(c) First person singular pronouns prompting action words

Figure 4.5: Contour plots of sample gamma distributions from the REPUBLICAN data set. The X-axis is rates for the question word and the Y-axis is rates for the answer word.

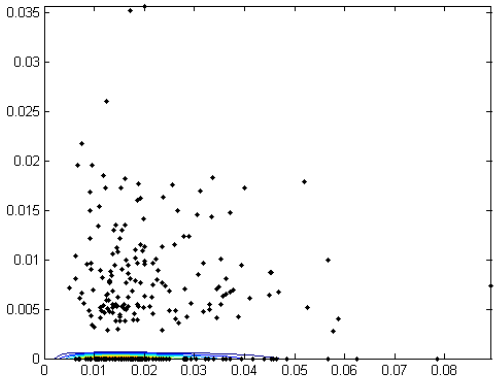


Figure 4.6: Contour plot of a "bad" gamma distribution ("wh" words prompting "but"). The X-axis is rates for the question word and the Y-axis is rates for the answer word. Note the contour lines clustered along the X-axis and failing to explain the visible nonzero points. For this distribution, $\gamma_x = 1.19$ and $\gamma_y = 0.763$.

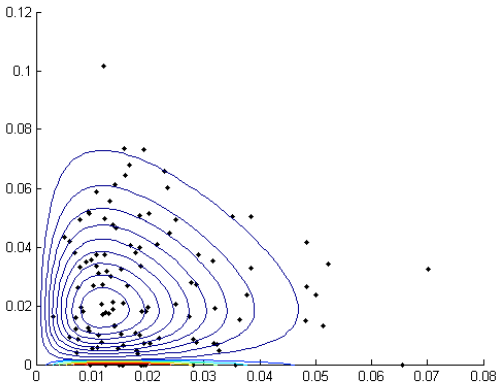


Figure 4.7: An example of a gamma distributed mixture model (first-person plural pronouns prompting first-person singular pronouns). The X-axis is rates for the question word and the Y-axis is rates for the answer word. One distribution describes responses while the other clusters along the X-axis and represents only non-responses.

triangular shape, we could not use simple affine transformations to make them "flat" as we did with Gaussians. We had the idea of making distributions for all of the data (both prompted and unprompted points), then sliding points along a contour in the contour plot—moving them from their current place in the distribution to an "unprompted" place with the same probability density. In theory, we could do this by solving the probability distrubition function $f(x, y)$ for y with $x = 0$ and f (the probability density value) held constant.

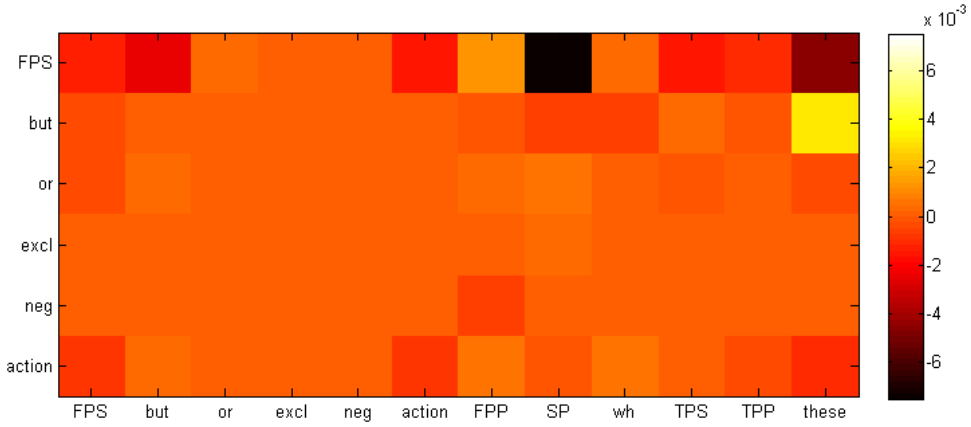
However, attempts to solve for y were stymied because y appears in $f(x, y)$ in different forms—as a term, as an exponent, and otherwise. We isolated terms containing y and reduced the equation to

$$y^{\gamma_y - 1}\epsilon(\frac{-\beta_y}{1-\eta})^y \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} y^{j+k} = \frac{K2f}{(\beta_x x)^{\gamma_x - 1}\beta_y^{\gamma_y - 1}\epsilon(\frac{-\beta_x x)}{1-\eta}) \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} c_{jk}(\beta_x x)^j(\eta\beta_y)^{j+k}}$$

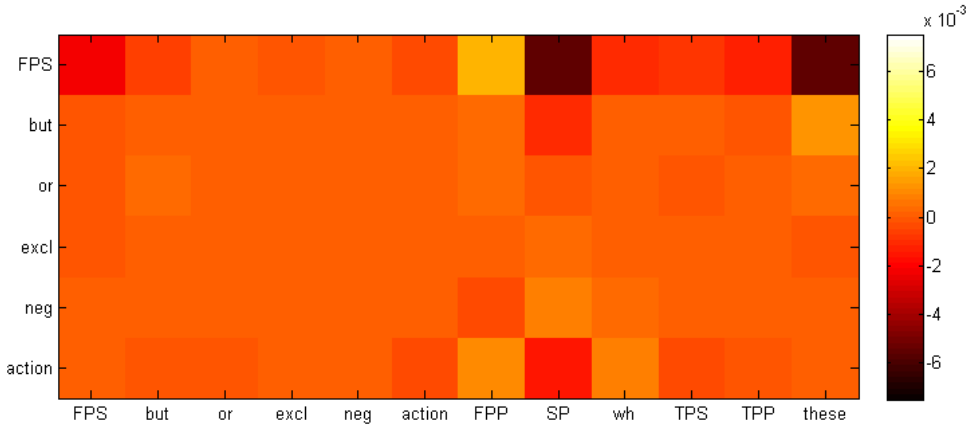
We suspect that this is not analytically solvable, particularly given the double summation. In any case, we could not reduce the equation further. An alternate approach would be to perform gradient descent and find a spot on the Y-axis with a probability density close to f, but such spots do not necessarily exist for all relevant values of f. We decided that, while gamma distributions showed certain kinds of promise, the cost of their unwieldy and complicated structure outweighed the benefit they might otherwise have over Gaussians.

4.3 Window Sizes

Recall that we achieved the above results with the REPUBLICAN data set at a minimum window size of 50. The next step was to check whether the results were the same at other window sizes. In order to see all the prompting effects at a glance, we

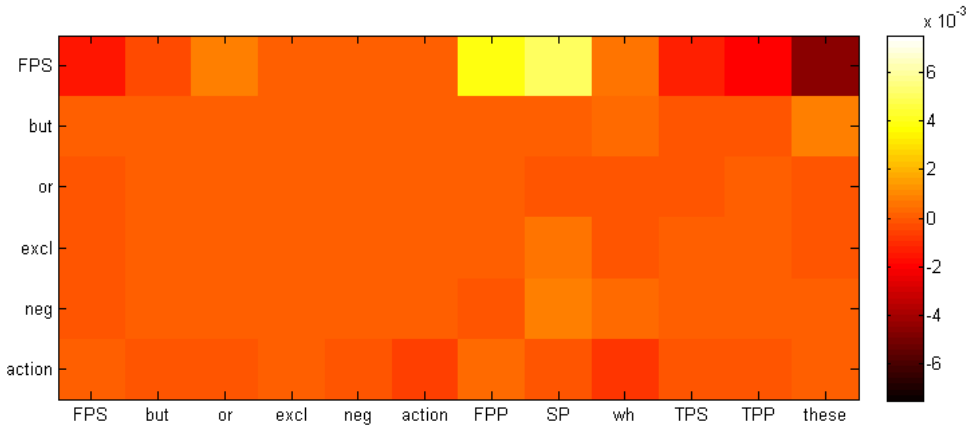


(a) Minimum window size 50

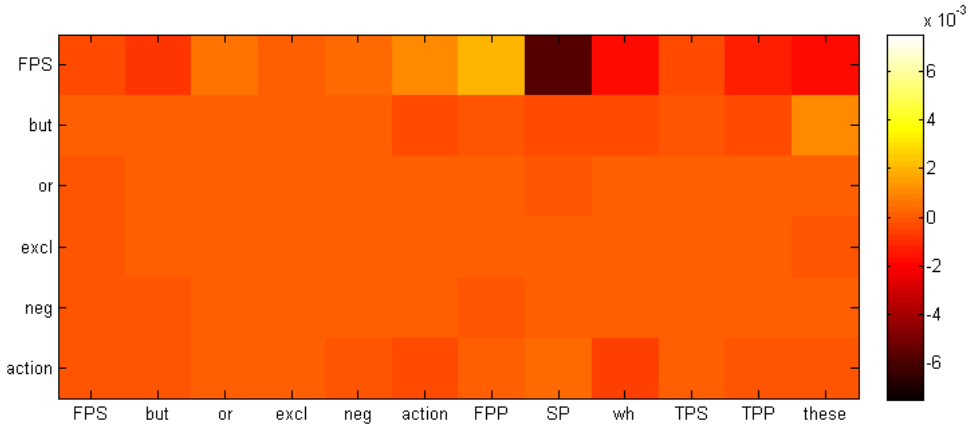


(b) Minimum window size 30

Figure 4.8: Color maps showing the average change in answer rates as the result of corrections for each question-and-answer pair at each minimum window size (with question words on the X-axis and answer words on the Y-axis). Bright colors indicate that higher rates of the question word prompted lower rates of the answer word, and thus, the answer word rates were increased during corrections. Dark colors indicate the opposite. All of these maps are on the same color-based scale.



(a) Minimum window size 10



(b) All windows (no minimum size)

Figure 4.9: Continued from Figure 4.8

made color maps showing the average change in each question-answer pair during the correction (Figure 4.8 and 4.9). We measured the average change on a rate scale. In other words, if the average change of a question-answer pair is -0.02, then on average, one in every fifty words in each answer is an answer word of the relevant type, prompted by a question word of the relevant type, and must be removed from the analysis. (In practice, average changes were much smaller than this!)

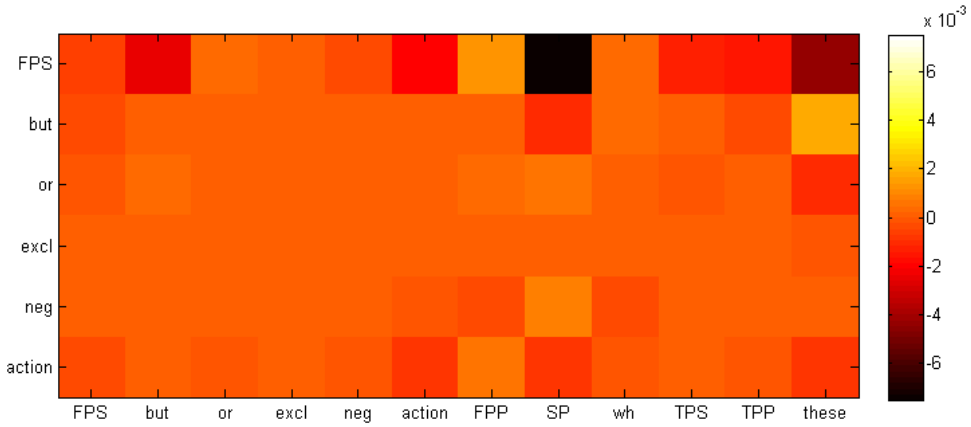


Figure 4.10: Color map showing the average change of each question-and-answer pair for the composite windows. Note that this pattern is close to the patterns of the 50 and 30-word windows.

We also made a color map for the corrections applied to composite windows (Figure 4.10). We found that the patterns in 50-word windows degrade as the minimum window size is lowered, particularly to below 30. The most notable degradation happened in our most promising question-answer pair—second-person pronouns prompting first-person singular pronouns. The composite windows also showed slight degradation, comparable to that in the 30-word windows.

Table 4.1 shows how much of each data set we could represent and process at each window size. For NUREMBERG, small windows predominated, and there were many long stretches where one person answered many questions in a row, making this (and other courtroom data) particularly amenable to the use of composite windows. With NUREMBERG the composite window technique allowed analysis of much more data than the 30-word window. Furthermore, it caused less degradation than the small window sizes which would otherwise have been needed to cover this much data. In

	Republican		Nuremberg		Simpson	
	Q	A	Q	A	Q	A
Full data set	65,480	236,411	109,551	201,548	219,262	193,123
Minimum 10	55,161	159,450	75,313	155,395	45,162	71,357
Minimum 30	43,876	107,836	31,452	51,735	3,280	4,967
Minimum 50	33,193	72,078	15,394	22,366	323	350
Composite	44,759	101,211	105,423	170,706	219,197	192,537

Table 4.1: Number of words that could be taken into account during processing at each minimum window size

Republican, windows tended to be larger, so the effect was less pronounced, but the composite window technique still gave coverage and degradation comparable to that of the 30-word windows. For these reasons, we performed the rest of our analysis (in all three data sets) with composite windows.

4.4 Checks of the Correction Method

A few things need to be said about our correction method. First, as is apparent from Figure 3.3, we confirmed that the affine transformations work as specified: they rotate the prompted distribution by the correct number of degrees and move it so that it is parallel to the prompted distribution.

We should note the size of the effects produced by the correction method. The changes to rate statistics are shown in Figures 4.8 and 4.9, but it would also be good to know what those statistics mean in terms of changes to the actual word counts in

each window. We found that, on average, three first person singular pronouns were taken away from each window. (This is at an average window size of 188 words, about six of which on average were FPS—so about half of these FPS pronouns, according to our model, are caused by prompting.) The average change in other categories was less than one word.

When we averaged the absolute values of the changes, rather than the net changes, FPS stayed about the same. However, the rates of change in "but" and action words rose to about one word per window. This means that some windows were changed in a positive direction and some in a negative direction, which produces a net change about zero, but is a distinct type of process from the other word categories, which remained well below half of a word even when absolute values were taken.

These values were basically the same in the other two data sets for FPS and action words. However, the effect on "but" in the NUREMBERG and SIMPSON data sets fell below half a word.

We also had some minor concerns about the sensitivity of the correction method to distortion by outliers. We re-ran the corrections for each category of answer words after taking out the 2.5% highest rates and the 2.5% lowest rates for that category (removing 5% of the data in total). In this restricted analysis, the average absolute value of the difference after correction was still about three words for FPS and about one word for action words. However, the average change in FPS without taking an absolute value was reduced to only two words, and the effect on "but" disappeared. This suggests that the correction method is somewhat sensitive to outliers, but that its results for FPS and action words are largely valid.

By this point we had made several important discoveries. We had discovered that

a prompting effect did exist. We had developed a method for removing the effects of prompting and verified that it did so, transforming the original data distribution into one in which the question and answer word rates were independent of each other. We had also worked out the best way in which to apply these methods to data with short windows. Now it was time for validation: discovering whether or not our techniques helped distinguish deceptive and truthful subgroups.

4.5 Validation: Nuremberg and SVD

To validate our work, we used the NUREMBERG data. Recall that NUREMBERG was divided up into DEFENDANTS, TRUSTWORTHY WITNESSES, and UNTRUSTWORTHY WITNESSES, and we predicted that a working correction method would increase the distance in semantic space between these groups.

Figure 4.11 shows the result of singular value decomposition on the NUREMBERG data before correction, and Figure 4.12 shows the result after correction. The effect is exactly the opposite of what we predicted—performing the correction erased any spatial distinction that existed between the groups.

It appeared that, rather than being something to do away with, the prompting effect actually contained important information about deception.

To better understand these results, we looked more closely at the prompting effects in NUREMBERG's subgroups.

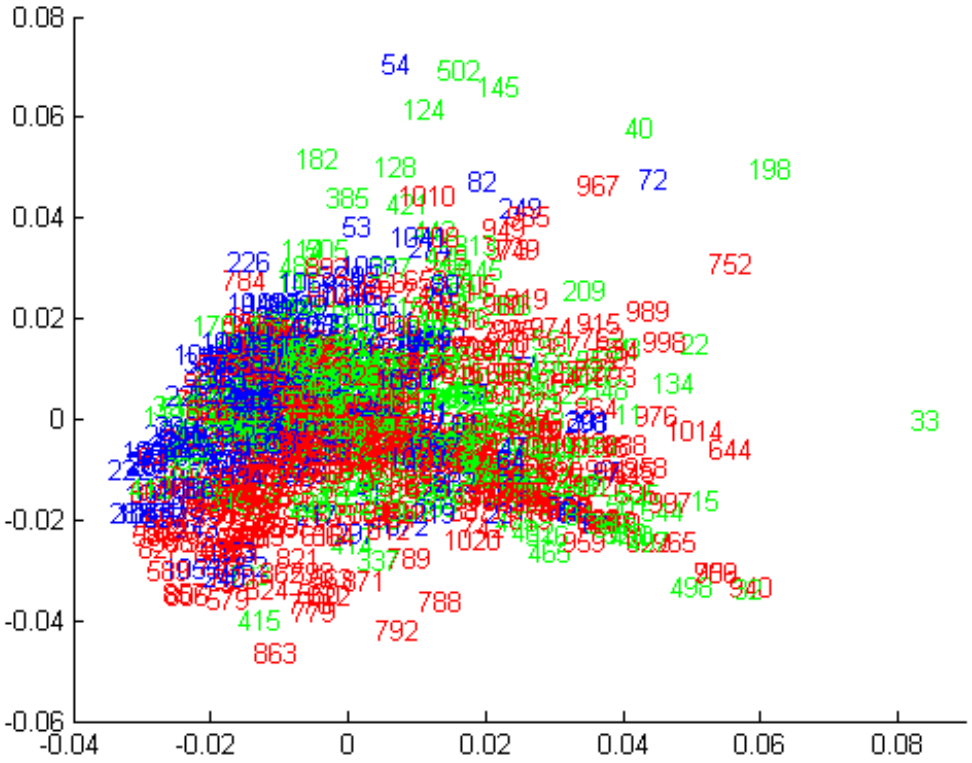


Figure 4.11: Singular value decomposition of the "response" words in the NUREMBERG data set prior to correction. DEFENDANTS are marked in red, TRUSTWORTHY WITNESSES in blue, and UNTRUSTWORTHY WITNESSES in green. Note that the TRUSTWORTHY WITNESSES are concentrated on one side of the semantic space.

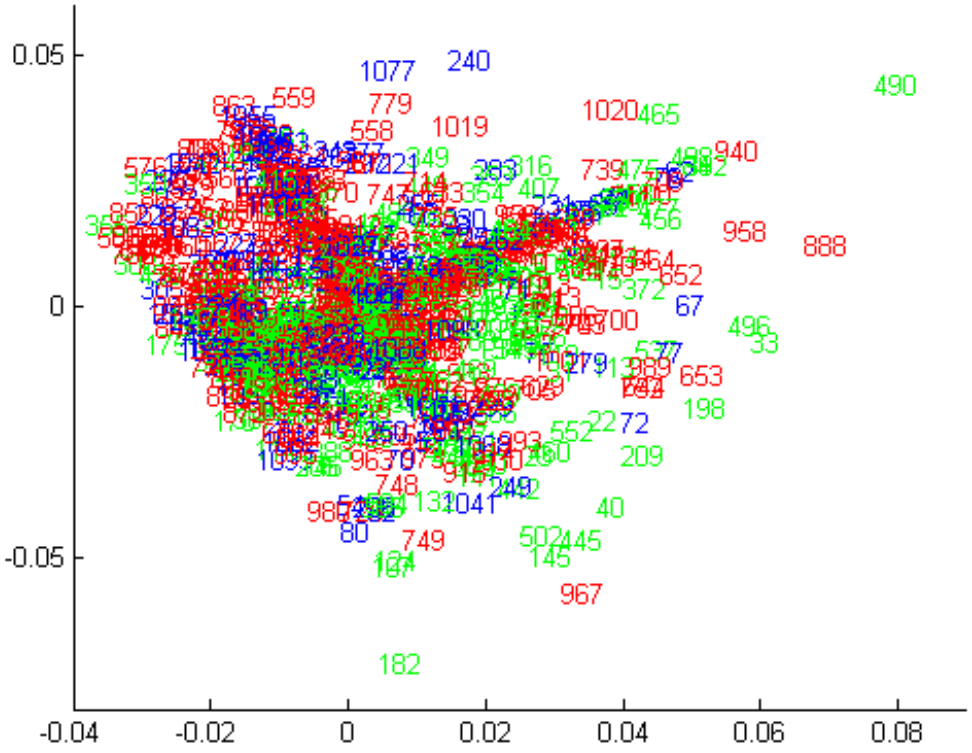
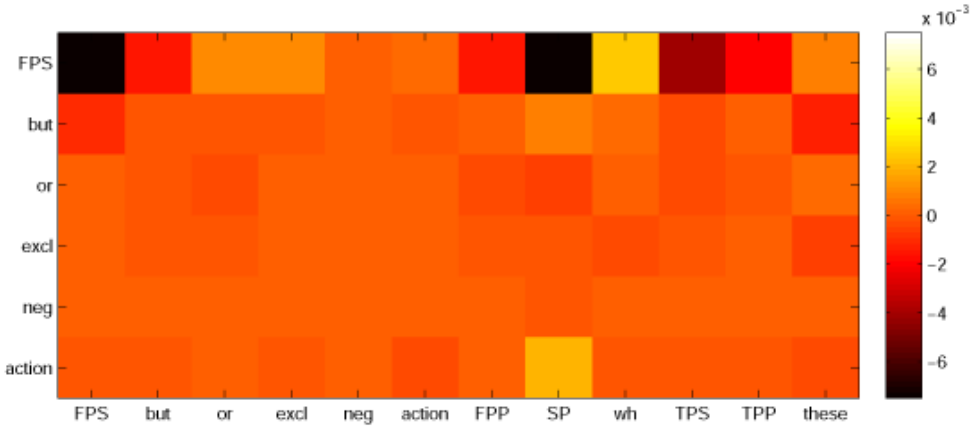
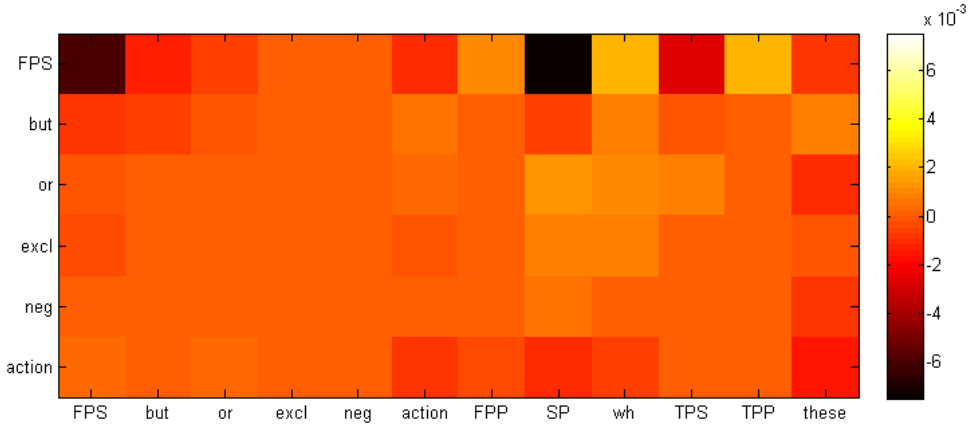


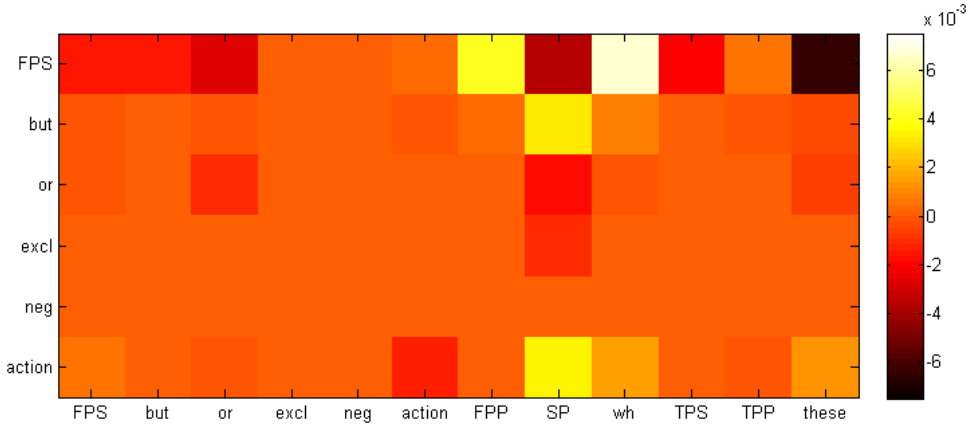
Figure 4.12: Singular value decomposition of the "response" words in the NUREMBERG data set after correction. DEFENDANTS are marked in red, TRUSTWORTHY WITNESSES in blue, and UNTRUSTWORTHY WITNESSES in green. Note that the TRUSTWORTHY WITNESSES are no longer as highly concentrated on one side.



(a) Defendants



(b) Untrustworthy



(c) Trustworthy

Figure 4.13: Corrections for the three Nuremberg subgroups analyzed separately.

4.6 Validation: Subgroups in the Nuremberg Data

We analyzed each of the three NUREMBERG subgroups separately and ran a separate correction on each of them, to see if there was a difference between more and less deceptive groups. Figure 4.13 shows color maps of this data. The differences between DEFENDANTS and UNTRUSTWORTHY WITNESSES (both of the Nazi subgroups) are not large, but the patterns of the TRUSTWORTHY WITNESSES subgroup appear to be quite different.

Note that these color maps are on the same scale as the earlier Republican color maps. We chose to present them this way in order to make comparisons easy across the different color maps. However, by using a scale that works for the REPUBLICAN data, we have obscured an important fact about the NUREMBERG data—namely, the average change in first-person singular pronouns prompted by second-person pronouns, for both the DEFENDANTS and UNTRUSTWORTHY WITNESSES, is literally off the scale. Both of these are more than twice the maximum amount shown by the color map; the DEFENDANTS' change is somewhat larger than the UNTRUSTWORTHY WITNESSES. (Since we were using composite windows for this, the REPUBLICAN average change is also slightly higher than it was in the previous colormaps, which were made with 50-word minimum windows.) Figure 4.14 shows this more clearly.

Seeing these large differences prompted us to look at individual distributions more closely. We overlaid the distribution from one subgroup onto the corresponding distributions for the other subgroups. This confirmed that different subgroups responded to prompts differently. They were not simply being prompted at different rates, or showing different rates of response independently of the prompt: many question-answer pairs showed two completely different distributions with differences in slope and other

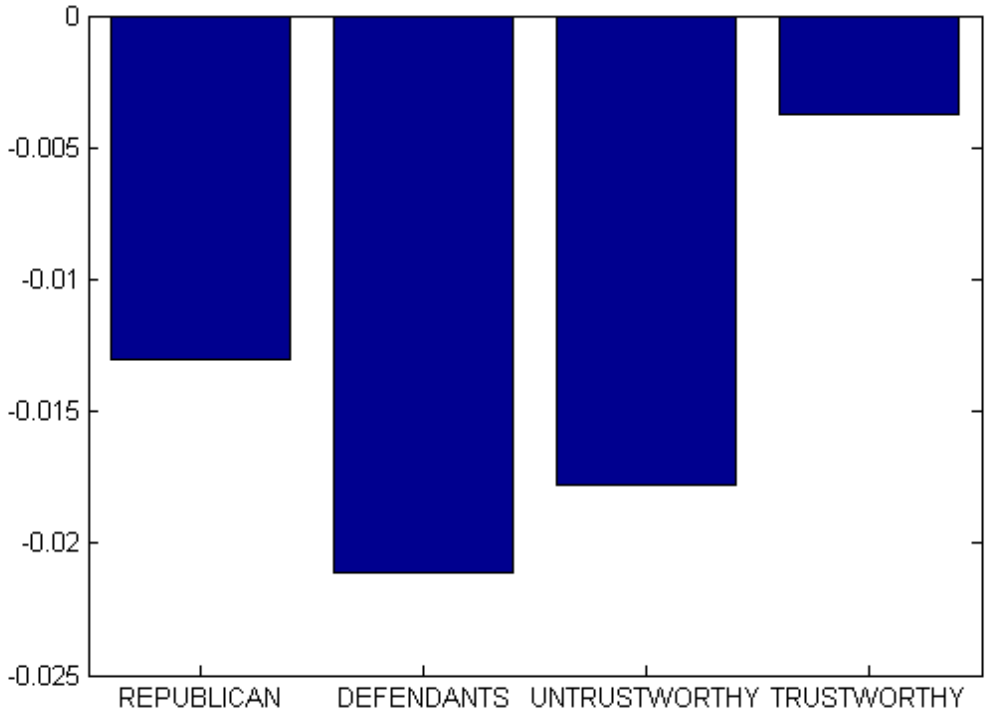
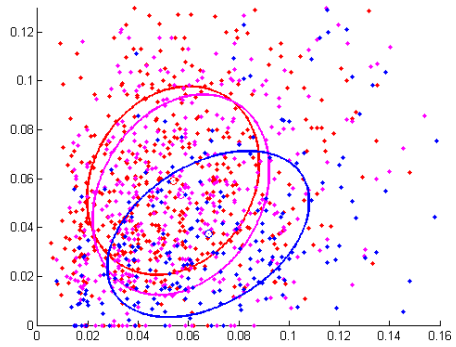


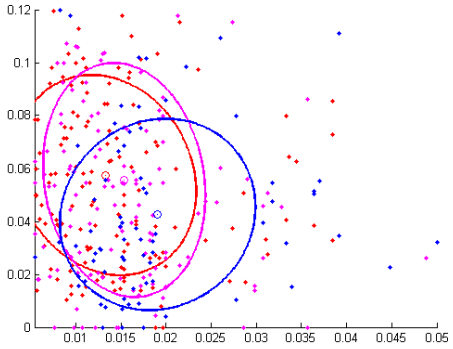
Figure 4.14: Average change in rates of first-person singular pronouns prompted by second-person pronouns in the Republican and Nuremberg data sets, all using composite windows.

properties. In general, Defendants and Untrustworthy witnesses—the two deceptive groups—were basically the same, but the Trustworthy witnesses were different. The strongest effects tended to involve first-person pronouns or action words. Four of the most notable sets of distributions are shown in Figure 4.15.

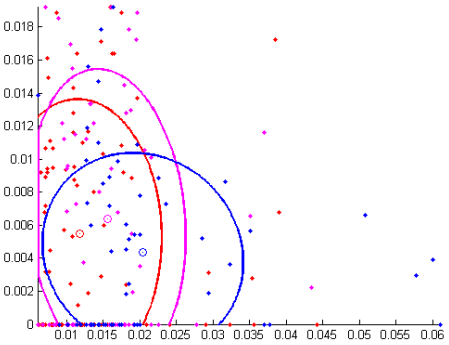
What this means is that the prompting effect is receiver-state-dependent. The deceptive groups respond to prompting in one way, and the truthful group in another—usually more weakly, though in a few cases they respond more strongly. Note that



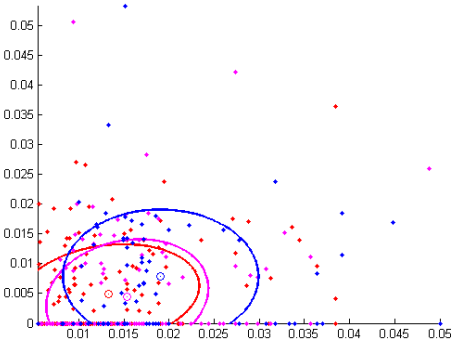
(a) Second person pronouns prompting first-person singular pronouns



(b) "Or" prompting first-person singular pronouns



(c) First person plural pronouns prompting "but"



(d) "Or" prompting action words

Figure 4.15: Gaussian distributions for the NUREMBERG data set, separated by subgroup. DEFENDANTS are red, UNTRUSTWORTHY WITNESSES are magenta, and TRUSTWORTHY WITNESSES are blue. The X-axis is rates for the answer words and the Y-axis is rates for the question words. Not all of the figures are on the same scale.

one of the particularly strong differences involves second-person pronouns prompting first-person singular pronouns, which was already one of the strongest prompting effects in our data. Correcting for prompting reduced, rather than increased, the difference between these groups, because the prompting itself is a factor distinguishing them from each other. Rather than having a base rate of word use (due to deception or lack thereof) which is modulated by prompting, the different subgroups actually experience different *kinds* of prompting—which means paying attention to the way in which prompting occurs ought to further elucidate differences between them.

4.7 Validation: Random Forests

At this stage we were satisfied that we saw differences between subgroups, but we were concerned about whether the interaction between questions and answers, *per se*, had any value. It was possible that the differences we saw were mostly detectable without reference to question words, and could already be detected by models like the LIWC model, which use only answer words. We wanted to know if the question words really improved the analysis.

To get a rough estimate of the significance of question words, we trained a pair of random forests: one with the rates of only the six response word categories in the answers, and one with these plus all the stimulus words in the questions. We created these random forests using the MATLAB Statistics Toolbox. For simplicity, we included only the Defendants and Trustworthy witnesses subgroups.

Because not every statement by Defendant is a lie, we did not expect high accuracy in either of these forests. Rather, we wanted to compare them to each other to see if the question words made a difference. If interactions between questions and

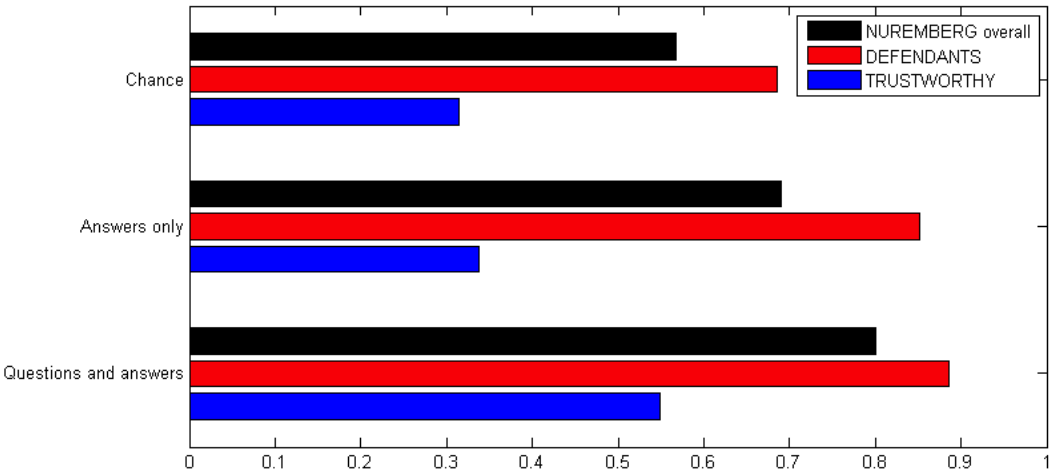


Figure 4.16: Prediction accuracy of random forests trained on the NUREMBERG data

answers were predictive of deception, then the model with both question and answer words should make better predictions, and it should focus on the words involved in the most promising interactions (i.e. first-person singular pronouns in the answer and second-person pronouns in the question). If only the words in the answers were relevant to deception, then both models would perform about the same and would focus on answer words.

Figure 4.16 shows the performance of these random forests. While the answer-words-only random forest performed above chance, it showed a strong bias: its accuracy with TRUSTWORTHY WITNESSES was much lower than its accuracy with DEFENDANTS. (That is to say, it was much too quick to classify windows of both kinds as DEFENDANTS. This was probably because there were more windows belonging to DEFENDANTS than to TRUSTWORTHY WITNESSES.) Adding the question words improved overall accuracy by more than 10 percentage points—an even more dramatic

result than we expected. Moreover, it reduced bias by an even larger amount, producing a huge improvement on the TRUSTWORTHY WITNESSES without reducing the accuracy on DEFENDANTS.

Recall that random forests can provide an estimate of the importance of each variable by counting how many times a tree within the forest "split" on a particular variable. Table 4.2 shows the results of this estimation with our best-performing random forest. The most influential words in the question-and-answer random forest were the same as we predicted: first-person singular pronouns in the answer and second-person pronouns in the question took the top two spots, with similar frequencies, which means it is likely that an interaction between both categories drove many of the decision trees in the forest.

There were some other interesting and somewhat surprising results as well. After these top two spots, the three next most important word categories were all question words, suggesting that the forest not only supplemented its reasoning with question words, but actually made more decisions based on question words than on answer words. However, a small overrepresentation of question words is to be expected given that we are counting a larger number of word categories in the question than we are in the answer. Also, in some cases a question word may give context to an answer word and thus need to be considered first.

4.8 Validation: the Simpson data set

Our final task was to verify that these validations generalized: that is, that they did not simply reflect properties of the NUREMBERG data. So we looked at the SIMPSON data set and performed each validation again.

Word category		Splits
A	first-person singular pronouns	10771
Q	second-person pronouns	10563
Q	"wh" words	9581
Q	"these", "those", and "to"	9277
Q	first-person singular pronouns	6839
A	"but"	6584
A	action words	6581
Q	"or"	4934
A	"or"	4692
Q	action words	4165
Q	third-person plural pronouns	3973
Q	first-person plural pronouns	3641
A	misc exclusive words	3570
Q	third-person singular pronouns	3383
Q	"but"	3119
A	negative emotion words	2254
Q	misc exclusive words	1538
Q	negative emotion words	871

Table 4.2: Word categories in the random forest trained with question-and-answer data, ranked by the number of splits in the model using each word.

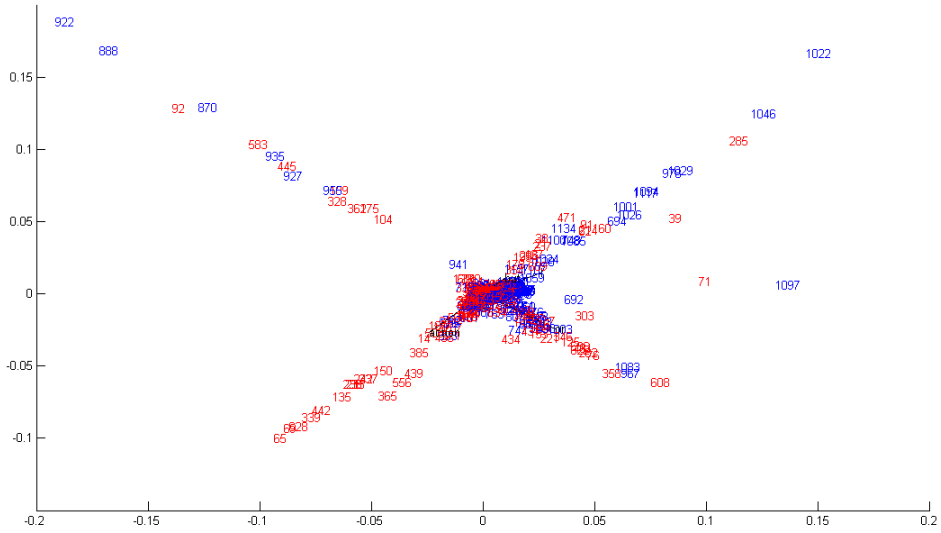


Figure 4.17: Singular value decomposition of the "response" words in the SIMPSON data set prior to correction. O.J. Simpson's answers are marked in red and PLAINTIFFS in blue.

The SVD results from NUREMBERG generalized to the SIMPSON data: while a modest visual distinction existed in semantic space between Simpson and the PLAINTIFFS, applying our correction method did nothing to increase this distinction (Figures 4.17 and 4.18).

The closer look at subgroups did not generalize quite as well. While Simpson's deposition was somewhat different from those of the PLAINTIFFS (Figure 4.19), the two color maps did not differ in the same ways that the NUREMBERG color maps differed. Looking at the overlaid subgroups for individual question-answer pairs (Figure 4.20) was also perplexing: there was evidence of differences between the two groups, as there had been with NUREMBERG, but not quite in the same way. They tended

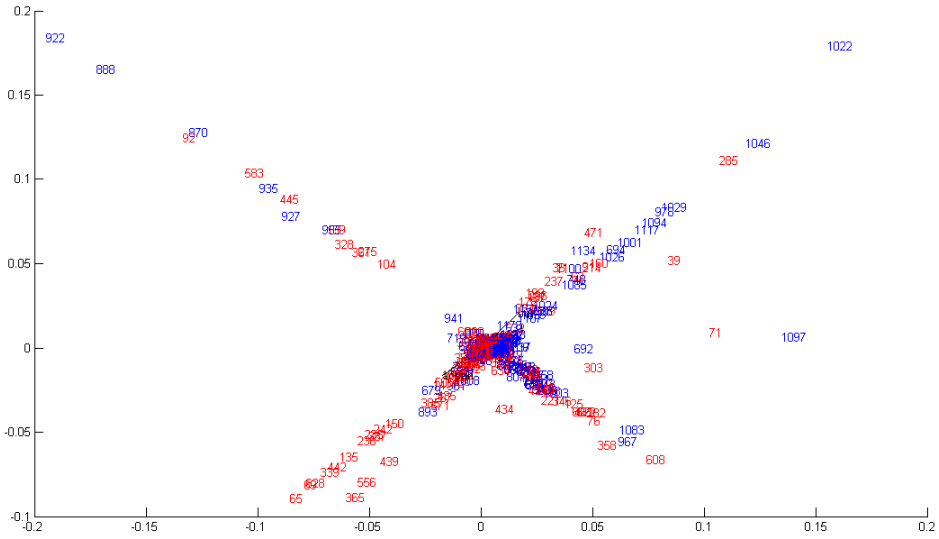
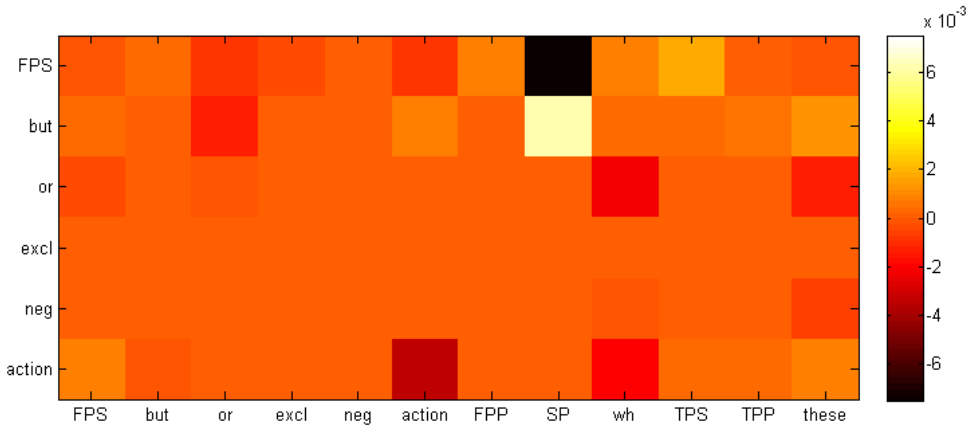


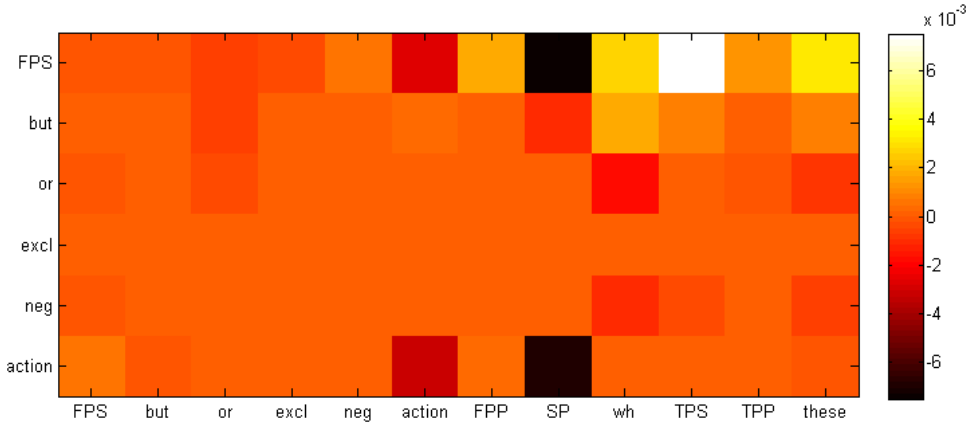
Figure 4.18: Singular value decomposition of the "response" words in the SIMPSON data set after correction. O.J. Simpson's answers are marked in red and PLAINTIFFS in blue.

to be slightly weaker than the differences in NUREMBERG. Moreover, the strongest differences between subgroups in the SIMPSON data did not usually correspond with the strongest differences in the NUREMBERG data. In particular, the relationship between second-person pronouns in the question and first-person singular pronouns in the answer did not appear to be different between subgroups. The two distributions appeared to be extensions of each other, with SIMPSON using more first-person singular pronouns simply because he was prompted more.

Finally we ran the SIMPSON data through the same random forest process that we used with NUREMBERG. Recall that, if there were no interesting differences in interaction between questions and answers, the two random forests would perform



(a) Simpson

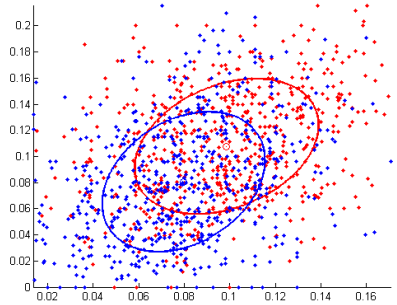


(b) Plaintiffs

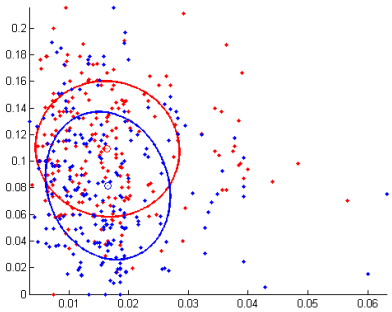
Figure 4.19: Corrections for the two O.J. Simpson subgroups analyzed separately.

about the same. The Nuremberg data showed a large increase in accuracy in the second forest, meaning interactions with question words were important. When we trained both forests on the Simpson data, they showed an increase in accuracy as well, but a smaller one (Figure 4.21).

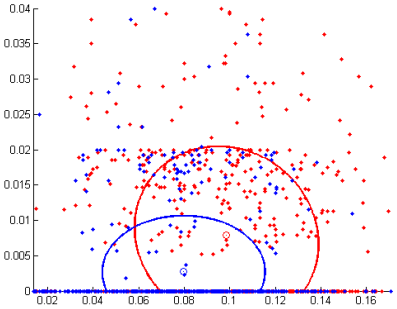
In general, the Simpson data supports the view that prompting effects exist, but it muddies the waters in terms of where the prompting effects actually are and whether



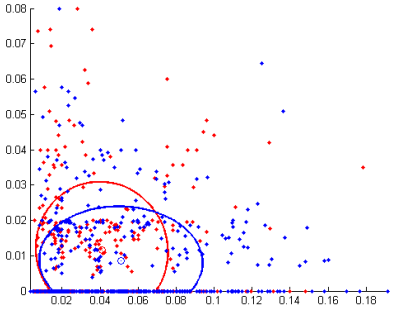
(a) Second person pronouns prompting first-person singular pronouns



(b) "Or" prompting first-person singular pronouns



(c) Second person pronouns prompting "but"



(d) Third person singular pronouns prompting action words

Figure 4.20: Gaussian distributions for the SIMPSON data set, separated by subgroup. SIMPSON is red and PLAINTIFFS are blue. The X-axis is rates for the answer words and the Y-axis is rates for the question words. Not all of the figures are on the same scale.

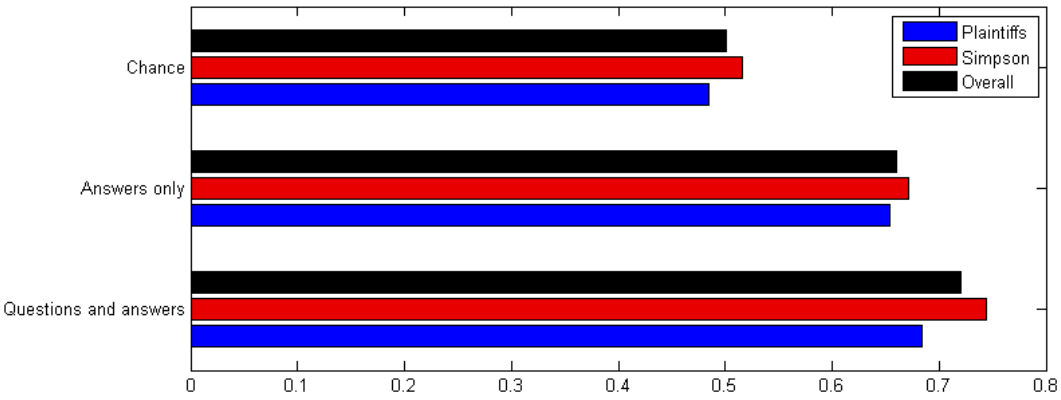


Figure 4.21: Prediction accuracy of random forests trained on the SIMPSON data

they can be predicted across different contexts. We discuss this, and some possible problems with the SIMPSON data, in Chapter 5

4.9 Summary of Contributions

This has been an exploratory analysis geared more towards uncovering patterns than creating definite answers. However, we have made the following contributions to the field of deception detection.

First, we have shown that a prompting effect exists: some word categories in a question do influence the frequency of other word categories words in the answer, including the words used in Newman *et al.*'s deception model. Moreover, we have proposed and verified a method for removing these effects. While this method by itself did not prove useful for our purposes, it does what it was meant to do in terms of removing correlations between question and answer words, replacing the answer words with a flattened distribution. It thus may be useful for other purposes. For

example, observing a pattern in a response, we can evaluate whether it is more likely to have arisen spontaneously or as a reaction to a prompt.

Second, we have discovered that prompting effects are receiver-state-dependent. Different groups respond to the same prompting differently and may require different corrections. This is more complicated to work with than a straightforward correction for everyone. If an appropriately sensitive correction method is found, it needs to be a more nuanced, and therefore more accurate, analysis.

Finally, we have shown that Newman *et al.*'s model—and any deception model dependent on only the words in the answer—is missing something. Looking at only the words in the answer gives us only a partial picture. Our validation with random forests supports this conclusion: deception models can improve markedly if they consider not only the respondent, but the full interaction between both questioner and respondent.

It should also be noted that our results support those of Skillicorn and Little [60]. The original LIWC deception model associates lowered rates of first-person singular pronouns with deception. But in the Gomery commission, Skillicorn and Little found that *higher* rates of first-person singular pronouns were associated with deception. Our data explains this discrepancy: the strongest and most noticeable prompting effect occurs with first-person singular pronouns, which are raised by most forms of prompting, particularly second-person pronouns. In an unprompted condition, deceptive people may use fewer first-person singular pronouns, but prompting raises the first-person singular pronoun rate, and does so to a greater degree when people are more deceptive. This also suggests why DePaulo *et al.* [30] and other meta-analyses did not find a significant association between first-person pronoun rates and

deception: they combined many different studies which included both monologues and dialogues.

However, our results do not explain Ott *et al.*'s [72] results, in which higher rates of first-person pronouns were found in "spam" reviews online, compared to genuine ones. Spam is not a question and answer setting. Ott *et al.*'s results most likely have more to do with factors unique to persuasive computer-mediated communications than with prompting, such as Zhou *et al.*'s [104]; since deceptive reviews ostensibly describe a positive personal experience with a product, perhaps deceptive reviewers refer to themselves excessively as a form of over-persuasion.

Chapter 5

Discussion

At present, neither humans nor computers are particularly good at detecting deception. A computer model is considered promising if it performs significantly better than humans—but all this means is that it performs significantly better than chance. When models perform with 70% accuracy (as in Mihalcea and Strapavara's LIWC-based study [64]), they are considered promising. But we can hardly afford for three in ten accused criminals to be falsely classified as deceptive—or, as Zhou *et al.* [104] suggest, for three in ten online job applicants to be erroneously screened out. State-of-the-art computerized deception models, at present, are simply not good enough. We also cannot rely on our own human ability to detect deception, since for the most part that ability does not exist. Therefore, improvements and refinements to existing deception models are desperately needed.

We have made one such refinement through the research described here. We have found that a prompting effect—similar to the effects measured by LSM [71], but generalized to include the fact that some classes of words prompt other classes—influences the rates of word usage in answers to questions. Moreover, the effect

is receiver-state-dependent. We developed a method to correct for the prompting effect and remove its influence, but this also removed a great deal of the distinction between deceptive and truthful subgroups. The way in which respondents respond to the prompting effect is, in itself, a cue to deception. We supported this conclusion by creating random forests to classify deceptive and truthful subgroups in our data, and found that the random forests performed best when they were given both question data and answer data. In this optimal condition, our random forests performed at 70-80% overall accuracy, which is very good by deception model standards.

This leads us to a number of take-home lessons for other deception researchers. First, as our random forest results show, paying attention to both questions and answers should improve word-based models—even if we are not sure quite what we are looking for. Researchers who have created deception classifiers using other machine learning methods, such as naive Bayes classifiers, should try adding question words as well as answer words to the classifier's input. This alone is likely to result in a measurable improvement.

But beyond this, our research suggests what it is in the questions that we ought to look for. We know that the prompting effect is receiver-state-dependent. So tracking the nature of the prompting effect across different groups of respondents in similar situations should illuminate differences between respondents, including their level of deception. We have made a good start at showing which relationships between words are most useful in such analysis—in particular, second-person pronouns prompting first-person singular pronouns.

Furthermore, we wish to note that deception is not the only setting for word-based analysis in which first-person pronouns are relevant. Research with LIWC [74] has

associated these pronouns with everything from personality [75], sex [69], and age [77] to relationship quality in families [83] and marriages [84]. Any of these variables, particularly the latter two, might be assessed in a dialogue setting. In any such setting, we may suppose that paying attention to both question words and answer words will improve accuracy. It may be the case that people of different personalities, sexes, or ages respond to prompting differently—which means taking question words into account will improve our ability to profile people in dialogues, for example, if a participant's identity in a computer-mediated chat setting is uncertain. Meanwhile, in the study of relationships—useful, for example, in analysis of social networks—there may be rich and interesting effects to uncover with regards to the association between relationship style or quality and the way the people involved respond to each other's prompting. Perhaps people in poorer relationships respond to prompting less (as we might expect, since their LSM scores are generally lower [51]) or perhaps they are prompted by different kinds of words. Or perhaps changes in the kind of prompting that occur over time reflect changes in the relationship. Also, since the more dominant person in an interaction tends to exhibit less LSM than their subordinate(s) [51], we would expect that person to respond less to prompting as well. On the other hand, there may be settings of this nature in which all the interesting subgroups respond to prompting in the same way. In that case, our method for removing the influence of the question may prove useful.

What we ultimately hope that we have done is to introduce the prompting effect as an additional tool in the arsenal of text analysts. We must use all the tools at our disposal if we wish to make any computational sense of the huge complexity of human interaction, and allowing for prompting in our models brings us one step closer to

sense.

5.1 Limitations

5.1.1 Problems and potential biases in the courtroom data

Recall that the prompting effect exists in all the data sets we tested, and that both the NUREMBERG and SIMPSON data show receiver-state-dependent differences in prompting. However, the differences in the SIMPSON data set were somewhat different from those in the NUREMBERG data set and generally smaller. Consistent with smaller differences in prompting, the random forest model also showed a smaller improvement when question data was added. A question arises as to why this is so. Intuitively, either the SIMPSON data set underrepresents the difference between truthful and deceptive testimony, or the NUREMBERG data set overrepresents it—or both.

It is plausible that the SIMPSON data set might underrepresent this difference. The most striking feature of the SIMPSON data set is that both Simpson and the PLAINTIFFS show a high level of second-person pronouns prompting first-person singular pronouns. This is in sharp contrast to the NUREMBERG data set, in which TRUSTWORTHY WITNESSES show a great deal less prompting in this domain than any other group, and DEFENDANTS have the highest level, with UNTRUSTWORTHY WITNESSES not too far behind. (Meanwhile, as one might expect, the average prompting for these words in the REPUBLICAN data set is midway between the TRUSTWORTHY WITNESSES and the Nazis.)

We used the same scale for all of our color maps, across data sets, for two reasons. First, it allowed easy comparisons across data sets, and second, it kept relatively minor

differences visible, rather than washing them all out in favor of a single very large result in one cell. However, this scaling choice obscured the extent of the prompting effect in the NUREMBERG data set. There is actually a second obscured effect of this nature: in both sides of the SIMPSON data set, the prompting of first-person singular pronouns by second-person pronouns is much larger than with any of the NUREMBERG subgroups (about -0.035 for both groups). Care should be taken in interpreting this result. We obviously do *not* suggest that all parties in this civil suit were more deceptive than Nazis! Instead, recall that these data sets were analyzed using composite windows containing several adjacent questions and several adjacent answers. While the average question length in the NUREMBERG data set is fairly short compared to that of the REPUBLICAN data, the average question length in the SIMPSON data set is much shorter. Recall as well that the prompting of first-person singular pronouns by second-person pronouns appears slightly higher in the REPUBLICAN data set when analyzed using composite windows. It appears that clumping short windows together can increase the size of perceived prompting effects in those windows. Therefore, the high numbers in this part of the SIMPSON data set are probably an artifact of having many short questions and answers.

This problem illustrates the difficulty of comparing deceptive and truthful people across different contexts. Just as different levels of motivation and modes of communication can produce different word patterns in deception, different ways of delineating windows during the analysis appear to have an effect. To avoid hopelessly erroneous results, care should always be taken to compare deceptive people to truthful people in as close to the same situation as possible.

However, Simpson and the PLAINTIFFS can still be compared to each other, since

both groups were in identical situations: they were giving videotaped depositions immediately prior to a civil trial, presided over by the same group of lawyers. When the two subgroups are compared only to each other, their lack of differentiation in first-person pronouns remains a mystery.

One possible explanation is that the PLAINTIFFS are not an appropriate control group for a deception study: that is, that at some level they were being deceptive. This does not necessarily imply that they perjured themselves. Recall Gupta and Skillicorn's [44] use of the LIWC model to measure persona deception or "spin". We excluded some individuals widely considered deceptive, such as Kato Kaelin, from our analysis, but the PLAINTIFFS in Simpson's civil case were still highly motivated to present themselves in the best light possible. Like the plaintiffs in any civil suit, they had staked a great deal of money on the outcome. It is likely that, even when telling the truth, they were careful to say things in a way maximally favourable to themselves. Little research has been done on the communication of people who are truthful, or probably truthful, but who are highly motivated to "spin" the truth. This is despite the recommendation of deception researchers such as DePaulo *et al.* [30] who point out that such technically-truthful statements, as well as truthful statements involving feelings of guilt or personal distress, may look similar to deceptive ones. Until such research is thoroughly done, it is not advisable to apply deception models to civil suits, which present this kind of high stakes for both parties.

It is also plausible that the NUREMBERG data set overrepresents receiver-state-dependent differences in the prompting effect. Two factors other than veracity separate the TRUSTWORTHY WITNESSES from the other subgroups.

First, the DEFENDANTS and UNTRUSTWORTHY WITNESSES were native speakers

of German, while the TRUSTWORTHY WITNESSES came from a variety of other countries including France and the Soviet Union. Many of the witness's statements were translated into English in real time by translators who already knew who was a Nazi and who was not, and whose biases could have distorted their work.

Second, the TRUSTWORTHY WITNESSES and other groups were asked about different topics. TRUSTWORTHY WITNESSES were generally called on to give descriptions of conditions either for inmates of concentration camps or for civilians in Nazi-occupied countries. Cross-examination of TRUSTWORTHY WITNESSES was usually perfunctory, and some were not cross-examined at all. In contrast, DEFENDANTS and UNTRUSTWORTHY WITNESSES were asked largely about internal procedures within Nazi Germany, while the DEFENDANTS were also asked in detail about their own behaviour in relation to these procedures, and both groups were cross-examined thoroughly. These two different lines of questioning may have produced two different internal emotional states quite apart from the deceptive or truthful nature of the witnesses' communication. However, if this were the case, one would also expect to see a greater distinction between DEFENDANTS (who were on trial specifically for their own deeds as individuals) and UNTRUSTWORTHY WITNESSES (who often testified for the prosecution, and who were responsible for accounts of Nazi procedures in general, not only of their own actions). This distinction does not generally occur in the data.

Third, power relationships probably affected the witness's use of pronouns, since they are known to affect LSM. DEFENDANTS were speaking in the presence of judges who had the power to decide if they would live or die. UNTRUSTWORTHY WITNESSES' lives were not at risk, but they were on the losing side of a war and were speaking in the presence of the victors. TRUSTWORTHY WITNESSES generally lacked this dramatic

power imbalance with the lawyers and judges presiding, and were not at risk of punishment or maltreatment based on their testimony. Despite the relatively small difference in the data between DEFENDANTS and UNTRUSTWORTHY WITNESSES, a strong confounding effect of these power relationships on the prompting differences between Nazis and TRUSTWORTHY WITNESSES cannot be ruled out.

As Hancock *et al.* [45] discovered, deception is a process involving both the questioner and respondent. It is not only the respondent who matches the questioner's style: the questioner also matches the respondent's, and changes in both can be indicative of deception, even if the questioner is not consciously aware of being deceived. If truthful and deceptive respondents are questioned differently, it may be that the questioner is biased—or it may be that the questioner is responding naturally and unconsciously to the deceptive respondent's speech patterns. To make matters worse, it may be both.

Moreover, attempting to correct for demographic differences and other suspected biases in archival data can be more trouble than it is worth. This is what Fornaciari *et al.* found when they built a support vector machine based partly on LIWC and small N-grams to distinguish true and false court testimony [37]. When they restricted their sample to only testimony by men, who are known to use pronouns and other LIWC categories differently from women [69], the model's performance did not improve — possibly because any benefit of homogeneity of the data was counteracted by the drawbacks of a smaller data set.

It is difficult to imagine an archival setting in which confounds of this kind do not occur—and it is also difficult to construct a laboratory experiment in which deceptive subjects are motivated as strongly as the people in these archival trials.

These confounds are an inherent problem with word-based psychological models, since function word use is affected by so many different variables. Any deception model used in the field will come into contact with all sorts of people who have different linguistic styles: men and women, people from different cultures, people experiencing different emotions and confronted by different kinds of questioning, even people with mental illnesses or disabilities that affect their word choice. No deception model should be considered usable for practical purposes unless it has been tested widely and performed consistently well across all these different kinds of categories. Otherwise, we use these models at our own risk, not knowing where their biases might lie and who might be hurt by their errors.

5.1.2 Rare words and potential for bias in the model

For some words—particularly negative emotion words, and exclusive words other than "but" and "or"—we did not find meaningful results because there were not enough words from these categories in the data to create valid models.

For an analysis method depending on the creation of bivariate distributions, this is inescapable. However, our analysis is admittedly biased towards common word categories. The "average change" metric, in particular, measures not only the strength of a relationship between two pairs of words but how frequently that relationship actually appears. We would defend this choice of metric by pointing out that, the more a model relies on common words, the more applicable it will be to relatively small windows: if one in ten thousand words is involved in a particular prompting effect, it will be hard to discern that prompting effect in 50-word windows, no matter how it is measured.

However, other metrics exist which are not biased towards common word categories. Ireland and Pennebaker's [51] LSM metric, for instance, uses a weighted difference measure which corrects for the base frequency of a category:

$$LSM = 1 - \frac{\|q - a\|}{q + a + .0001}$$

in which q is the number of words from a particular word category in the question, and a is the number of words from the same word category in the answer. This metric is not directly applicable to situations in which one category prompts another, since the baseline frequency of words in the two categories is not necessarily the same, which could skew the results. Still, future researchers might like to experiment with expansions of an LSM-like metric to this situation.

5.1.3 Missing words in the model

Hauch *et al.*'s meta-analysis suggests that the four word categories of Newman *et al.*'s LIWC model are not the only LIWC-detectable cues to deception. A full LIWC-based model, according to Hauch *et al.*, should contain the following additional categories:

- Words with higher rates in the deceptive condition: positive emotion words, emotion words overall, negations
- Words with lower rates in the deceptive condition: third person pronouns, tentative words, time-related words

A fuller analysis of prompting as applied to deception models should include these categories, as well as whatever categories might prompt them. In addition, there may be other words with a prompting effect on the original four categories which we may not have thought of.

However, the word categories we have used are sufficient to make this study's key points: that a prompting effect exists, that it can be corrected for, and that the relationship between question and answer words is itself a key to deception.

5.1.4 Gaussian and Gamma Distributions

We gave up on using gamma distributions to describe our data because they are mathematically complex and more difficult to work with then Gaussians. In particular, it is much easier to perform a correction using Gaussians than using gamma distributions. Also, gamma distributions only work for distributions containing a relatively large number of responses. However, for those cases in which gamma distributions work, they appear to work well—fitting the data better and more tightly than a Gaussian.

We would welcome some other researcher to try to duplicate our analysis using gamma distributions. It would be challenging to come up with a correction method, or way of measuring the strength of a gamma distribution in order to find significant differences between the distributions of subgroups, but such an analysis might well be more precise than the one we have carried out.

Correcting data according to the properties of a Gaussian model was counterproductive because it erased differences between subgroups. We supported this claim by locating the actual differences between the subgroups. However, the relatively poor fit of Gaussians to the data, compared to gamma distributions, introduced a small amount of approximation into each step of the correction. With several steps of correction for each answer word, this approximation may have snowballed into statistical noise which obscures underlying structures. A hypothetical noiseless Gaussian correction would still reduce differences between subgroups, not increase them. But

a correction based on gamma distributions, if such exists, might contain less of this noise and thus better elucidate exactly what is happening within these structures.

5.1.5 Limitations of Bag of Words Models

We also briefly note that throughout this research, we have been using bag-of-words models, and all their inherent limitations apply. Function words have grammatical meanings as well as psychological ones, but counting them does not imply understanding of grammatical structure. Since people have been shown to mimic large-scale grammatical structures as well as words [59], there will probably be forms of prompting that a bag-of-words model cannot detect. Our model is also not sensitive to effects that derive from word order.

Bag of words models, including LIWC and LSA, are stymied by polysemy—the ability of a word to have multiple meanings. LIWC, for example, counts "mad" as an anger word, but "mad" can also denote other emotions in sentences such as "I'm mad about you" or "we sent him to the hospital because he had gone mad". This is another factor limiting the models' precision.

Finally and obviously, bag-of-words models do not lead to any deep understanding of the meaning of sentences. Meaning-based criteria, such as those used in CBCA, are currently beyond the ability of any computer to detect. Nor can a computer evaluate whether or not a statement makes sense, even though "making less sense" was the most predictive cue to deception in DePaulo *et al.*'s meta-analysis [30]. For the foreseeable future, only humans can do these things.

While humans are not good at deception detection *per se*, we have developed an ability to use our understanding of meaning and reason to help compensate for

this weakness. In a courtroom, we do not merely watch the way the defendant and plaintiff talk: we also gather evidence and test whether their statements match the facts. This is a thing that can currently only be done by humans with human-level reasoning ability. Unless we invent a deception model with perfect accuracy, we must take care not to become too dependent on automated detection. A computer with much higher deception detection accuracy than humans would be extremely useful, but only humans can bring that enhanced ability into context through careful consideration of the facts.

Bibliography

[1] A. Abbasi and H. Chen. Visualizing authorship for identification. In *Proceedings of the 4th IEEE international conference on Intelligence and Security Informatics*, ISI'06, pages 60–71, Berlin, Heidelberg, 2006. Springer-Verlag.

[2] S.H. Adams. *Communication under stress: Indicators of veracity and deception in written narratives*. PhD thesis, Virginia Polytechnic Institute and State University, April 2002.

[3] American Broadcasting Company. Full transcript: ABC News Iowa Republican debate, December 11 2011. Accessed in winter 2012 at `http://abcnews.go.com/Politics/full-transcript-abc-news-iowa-republican-debate/`.

[4] S.A. Beaudreau, M. Storandt, and M.J. Strube. A comparison of narratives told by younger and older adults. *Experimental Aging Research*, 32(1), 2006.

[5] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022, January 2003.

[6] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[7] British Broadcasting Company. JK Rowling interview in full, June 19 2003. Accessed in fall 2011 at http://news.bbc.co.uk/2/hi/entertainment/3004456.stm.

[8] A.S. Brown and D.R. Murphy. Cryptomnesia: Delineating inadvertent plagiarism. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(3):432–442, 1989.

[9] J. Burgoon, J. Blair, T. Qin, and J. Nunamaker. Detecting deception through linguistic analysis. In Hsinchun Chen, Richard Miranda, Daniel Zeng, Chris Demchak, Jenny Schroeder, and Therani Madhusudan, editors, *Intelligence and Security Informatics*, volume 2665 of *Lecture Notes in Computer Science*, pages 958–958. Springer Berlin / Heidelberg, 2003.

[10] Cable News Network. Full transcript of CNN-Tea Party Republican debate, 20:00-22:00, September 12 2011. Accessed in fall 2011 at http://transcripts.cnn.com/TRANSCRIPTS/1109/12/se.06.html.

[11] Cable News Network. Republican debate, June 13 2011. Accessed in fall 2011 at http://transcripts.cnn.com/TRANSCRIPTS/1106/13/se.02.html.

[12] Cable News Network. Full transcript of CNN Arizona Republican presidential debate, Feburary 22 2012. Accessed in winter 2012 at http://archives.cnn.com/TRANSCRIPTS/1202/22/se.05.html.

[13] Cable News Network. Full transcript of CNN Florida Republican Presidential debate, January 26 2012. Accessed in winter 2012 at http://archives.cnn.com/TRANSCRIPTS/1201/26/se.05.html.

[14] J.N. Cappella and S.Planalp. Talk and silence sequences in informal conversations iii: interspeaker influence. *Human Communication Research*, 7(2):117–132, Winter 1981.

[15] J.R. Carlson, J. F. George, J.K. Burgoon, M. Adkins, and C.H. White. Deception in computer-mediated communication. *Group Decision and Negotiation*, 13:5–28, 2004. 10.1023/B:GRUP.0000011942.31158.d8.

[16] CBS Broadcasting Inc. Transcript: Michael J. Fox, June 12 2009. Accessed in fall 2011 at `http://www.cbsnews.com/8301-18563_162-2129702.html`.

[17] T. L. Chartrand and R. van Baaren. Human mimicry. *Advances in Experimental Social Psychology*, 41, 2009.

[18] The Chicago Sun-Times. CNN Republican debate, Nov. 22, 2011. Transcript. Accessed in fall 2011 at `http://blogs.suntimes.com/sweet/2011/11/cnn_republican_debate_nov_22_2.html`.

[19] The Chicago Sun-Times. CBS/National Journal GOP debate. Transcript, video, November 13 2011. Accessed in fall 2011 at `http://blogs.suntimes.com/sweet/2011/11/_cbsnational_journal_gop_debat.html`.

[20] The Chicago Sun-Times. CNBC Republican debate. Transcript, video highlights, November 9 2011. Accessed in fall 2011 at `http://blogs.suntimes.com/sweet/2011/11/cnbc_republican_debate_transcr.html`.

[21] The Chicago Sun-Times. Republican Las Vegas CNN debate: Transcript, October 19 2011. Accessed in fall 2011 at `http://blogs.suntimes.com/sweet/2011/10/republican_las_vegas_cnn_debat.html`.

[22] The Chicago Sun-Times. GOP NH ABC/Yahoo News debate: Transcript, January 8 2012. Accessed in winter 2012 at http://blogs.suntimes.com/sweet/2012/01/gop_nh_abcyahoo_news_debate_tr.html.

[23] The Chicago Sun-Times. GOP NH NBC's Meet the Press/Facebook debate: Transcript, January 8 2012. Accessed in winter 2012 at http://blogs.suntimes.com/sweet/2012/01/gop_nh_nbcs_meet_the_pressface.html.

[24] The Chicago Sun-Times. South Carolina GOP CNN debate, Jan. 19, 2012. Transcript, January 20 2012. Accessed in winter 2012 at http://blogs.suntimes.com/sweet/2012/01/south_carolina_gop_cnn_debate_.html.

[25] C. Chung and J. Pennebaker. The psychological functions of function words. In K. Fiedler, editor, *Social Communication*, pages 343–359. New York: Psychology Press, 2007.

[26] M.A. Cohn, M.R. Mehl, and J.W. Pennebaker. Linguistic markers of psychological change surrounding september 11, 2001. *Psychological Science*, 15(10):687–693, 2004.

[27] Council on Foreign Relations. Republican Debate Transcript, Tampa, Florida, January 2012. Accessed in winter 2012 at http://www.cfr.org/us-election-2012/republican-debate-transcript-tampa-florida-january-2012/p27180.

[28] S. Deerwester, S. T. Dumais, G.W. Furnas, and T.K. Landauer. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990.

[29] B.M. DePaulo, D.A. Kashy, S.E. Kirkendol, M.M. Wyer, and J.A. Epstein. Lying in everyday life. *Journal of Personality and Social Psychology*, 70(5):979–95, 1996.

[30] B.M. DePaulo, J.J. Linday, B.E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper. Cues to deception. *Psychological Bulletin*, 129:74–118, 2003.

[31] L.N. Driscoll. A validity assessment of written statements from suspects in criminal investigations using the scan technique. *Police Studies*, 17(4):77–88, 1994.

[32] P. Ekman and W.V Friesen. Nonverbal leakage and clues to deception. *Psychiatry: Interpersonal and Biological Processes*, 32(1):88–106, 1969.

[33] P. Ekman and M. O'Sullivan. Who can catch a liar? *American Psychologist*, 46(9):913–920, September 1991.

[34] P. Ekman, M. O'Sullivan, and M.G. Frank. A few can catch a liar. *Psychological Science*, 10(3), May 1999.

[35] F. Enos, E. Shriberg, M. Graciarena, J. Hirschberg, and A. Stolcke. Detecting deception using critical segments. In *Interspeech*, 2007.

[36] P.W. Foltz, D.Laham, and T.K. Landauer. The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1992.

[37] T. Fornaciari and M. Poesio. On the use of homogenous sets of subjects in deceptive language analysis. In *Proceedings of the Workshop on Computational*

Approaches to Deception Detection, 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 39–47, April 23 2012.

[38] Fox News. Transcript: Fox News Sunday interview with Sarah Palin, February 7 2010. Accessed in fall 2011 at `http://www.foxnews.com/politics/2010/02/07/transcript-fox-news-sunday-interview-sarah-palin/`.

[39] Fox News. Complete text of the Iowa Republican debate on Fox News channel, August 12 2011. Accessed in fall 2011 at `http://foxnewsinsider.com/2011/08/12/full-transcript-complete-text-of-the-iowa-republican-debate-on-fox-news-channel/`.

[40] M.G. Frank and P. Ekman. The ability to detect deceit generalizes across different types of high-stake lies. *Journal of Personality and Social Psychology*, 72(6):1429–1439, 1997.

[41] S. Freud. *Psychopathology of everyday life*. New York: Basic Books, 1901.

[42] A.C. Graesser, P. Wiemer-Hastings, K. Wiemer-Hastings, D. Harter, N. Person, and Tutoring Research Group. Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Enviornments*, 8(2):129–147, 2000.

[43] C.J. Groom and J.W. Pennebaker. The language of love: Sex, sexual orientation, and language use in online personal advertisements. *Sex Roles*, 52(7/8), April 2005.

[44] S. Gupta and D. B. Skillicorn. Improving a textual deception detection model. In *Proceedings of the 2006 conference of the Center for Advanced Studies on Collaborative research*, CASCON '06, New York, NY, USA, 2006. ACM.

[45] J.T. Hancock, L.E. Curry, S. Goorha, and M. Woodworth. On lying and being lied to: a linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, 45:1–23, 2008.

[46] V. Hauch, I. Blandn-Gitlin, J. Masip, and S.L. Sporer. Linguistic cues to deception assessed by computer programs: A meta-analysis. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 1–4, April 23 2012.

[47] J. Hirschberg, S. Benus, J.M. Brenier, F. Enos, S. Friedman, S. Gilman, C.G., M. Graciarena, A. Kathol, and L. Michaelis. Distinguishing deceptive from non-deceptive speech. In *Interspeech*, 2005.

[48] His Majesty's Stationery Office. The trial of German major war criminals sitting at Nuremberg, Germany, 1946. Accessed in summer/fall 2012 at `http://nizkor.org/hweb/imt/tgmwc/`.

[49] History Musings. Republican candidates debate in Sioux city, Iowa december 15, 2011. Accessed in winter 2012 at `http://historymusings.wordpress.com/2011/12/16/full-text-campaign-buzz-december-15-2011-fox-news-gop-iowa-debate-transcript-republican-presidential-candidates-debate-sioux-city-iowa/`.

[50] W. Ickes, S. Reidhead, and M. Patterson. Machiavellianism and self-monitoring: As different as "me" and "you". *Social Cognition*, 4(1):58–74, 1986.

[51] M.E. Ireland and J.W. Pennebaker. Language style matching in writing: Synchrony in essays, correspondence, and poetry. *Journal of Personality and Social Psychology*, 99(3):549–572, 2010.

[52] M.E. Ireland, R.B. Slatcher, P.W. Eastwick, L.E. Scissors, E.J. Finkel, and J.W. Pennebaker. Language style matching predicts relationship initiation and stability. *Psychological Science*, 20(10):1–6, 2012.

[53] M.K. Johnson and C.L. Raye. Reality monitoring. *Psychological Review*, 88(1):67–85, 1981.

[54] C.F. Bond Jr, K.N. Kahler, and L.M. Paolicelli. The miscommunication of deception: An adaptive perspective. *Journal of Experimental Sociao Psychology*, 21:331–345, July 1985.

[55] S.W. Gregory Jr., K.Dagan, and S.Webster. Evaluating the relation of vocal accommodation in conversation partners' fundamental frequencies to perceptions of communication quality. *Journal of Nonverbal Behavior*, 21(1):23–43, Spring 1997.

[56] J.H. Kahn, R.M. Tobin, A.E. Massey, and J.A. Anderson. Measuring emotional expression with the linguistic inquiry and word count. *The American Journal of Psychology*, 120(2):263–286, 2007.

[57] P. S. Keila and D. B. Skillicorn. Detecting unusual and deceptive communication in email. In *Centers for Advanced Studies Conference*, pages 17–20, 2005.

[58] R.E. Kraut and D. Poe. Behavioral roots of person perception: The deception judgments of customs inspectors and laymen. *Journal of Personality and Social Psychology*, 39(5):784–798, 1980.

[59] W.J.M. Levelt and S. Kelter. Surface form and memory in question answering. *Cognitive Psychology*, 14(1):78–106, January 1982.

[60] A. Little and D.B. Skillicorn. Detecting deception in testimony. In *IEEE International Conference on Intelligence and Security Informatics*, pages 13–18, June 17-20 2008.

[61] M. Lorenz and S. Cobb. Language behavior in manic patients. *Archives of Neurology & Psychiatry*, 67(6):763–770, June 1952.

[62] M.R. Mehl and J.W. Pennebaker. The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of Personality and Social Psychology*, 84(4):857–870, 2003.

[63] Microsoft National Broadcasting Company. 'Meet the Press' transcript for Dec. 7, 2008: President-elect Barack Obama. Accessed in fall 2011 at `http://www.msnbc.msn.com/id/28097635/ns/meet_the_press/t/meet-press-transcript-dec/`.

[64] R. Mihalcea and C. Straparava. The lie detector: Explorations in the automatic recognition of deceptive language. In *ACL-IJCNLP*, pages 309–312, 2009.

[65] G.A. Miller. *The science of words*. New York: Scientific American Library, 1995.

[66] M. Natale. Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *Journal of Personality and Social Psychology*, 52(5):790–804, 1975.

[67] National Archive. Official transcript of the military tribunal in the matter of the United States of America against Karl Brandt *et al.*. Harvard Law School Library: Nuremberg Trials Project: A Digital Document Collection, 1946-1947. Accessed in spring/summer 2012 at `http://nuremberg.law.harvard.edu/`.

[68] The New York Times. The Republican debate at the Reagan Library, September 7 2011. Accessed in fall 2011 at `http://www.nytimes.com/2011/09/08/us/politics/08republican-debate-text.html`.

[69] M.L. Newman, C.J. Groom, L.D. Handelman, and J.W. Pennebaker. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45:211–236, 2008.

[70] M.L. Newman, J.W. Pennebaker, D.S. Berry, and J.M. Richards. Lying words: predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29(5):665–675, May 2003.

[71] K.G. Niederhoffer and J.W. Pennebaker. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360, December 2002.

[72] M. Ott, Y. Choi, C. Cardie, and J.T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Annual Meeting of the Association for Computational Linguistics*, pages 309–319, June 19-24 2011.

[73] M. Pasupathi. Telling and the remembered self: linguistic differences in memories for previously disclosed and previously undisclosed events. *Memory*, 15(3):258–70, April 2007.

[74] J.W. Pennebaker. Linguistic Inquiry and Word Count. Accessed in fall 2012 at `http://www.liwc.net/`.

[75] J.W. Pennebaker and L.A. King. Linguistic styles, language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6):1296–1312, 1999.

[76] J.W. Pennebaker, T.J. Mayne, and M.E. Francis. Linguistic predictors of adaptive bereavement. *Journal of Personality and Social Psychology*, 72(4):863–871, 1997.

[77] J.W. Pennebaker and L.D. Stone. Words of wisdom: Language use over the life span. *Journal of Personality and Social Psychology*, 85:291–301, 2003.

[78] PolitiSite. Transcript - Fox News-Google GOP presidential debate September 22, 2011 Orlando, Florida. Accessed in fall 2011 at `http://www.politisite.com/2011/09/23/transcript-fox-news-google-gop-presidential-debate-september-22-2011-orlando-florida/`.

[79] S. Porter and J.C. Yuille. The language of deceit: an investigation of the verbal clues to deception in the interrogation context. *Law and Human Behavior*, 20(4):443–458, 1996.

[80] RonPaul.com. Fox News debate, Greenville SC, May 5 2011. Accessed in fall 2011 at `http://www.ronpaul.com/2012-ron-paul/debates-2012/previous/may-5-2011-greenville-south-carolina/`.

[81] S.S. Rude, E.M. Gortner, and J.W. Pennebaker. Language use of depressed and depression-vulnerable college students. *Cognition and Emotion*, 18(8):1121–1133, 2004.

[82] J.W. Schooler, D. Gerhard, and E.F. Loftus. Qualities of the unreal. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12(2):171–181, 1986.

[83] R.A. Simmons, D.L. Chambless, and P.C. Gordon. How do hostile and emotionally overinvolved relatives view relationships? What relatives' pronoun use tells us. *Family Process*, 43(3), 2008.

[84] R.A. Simmons, P.C. Gordon, and D.L. Chambless. Pronouns in marital interaction: What do "you" and "I" say about marital health? *Psychological Science*, 16(12), 2005.

[85] D.B. Skillicorn and C. Leuprecht. The mental state of influencers. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Workshop on Foundations of Open-Source Intelligence*, pages 922–929, August 2012.

[86] D.B. Skillicorn and A. Little. Patterns of word use for deception in testimony. In Christopher C. Yang, Michael Chau, Jau-Hwang Wang, and Hsinchun Chen, editors, *Security Informatics*, volume 9 of *Annals of Information Systems*, pages 25–39. Springer US, 2010.

[87] S.L. Sporer. The less travelled road to truth: verbal cues in deception detection in accounts of fabricated and self-experienced events. *Applied Cognitive Psychology*, 11:373–397, 1997.

[88] G.W. Stewart. On the early history of the singular value decomposition. *SIAM Review*, 35:551–566, December 1993.

[89] S.W. Stirman and J.W. Pennebaker. Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic Medicine*, 63:517–522, 2001.

[90] Superior Court of the State of California. The Simpson trial transcripts, 1996. Accessed in fall 2012 at `http://walraven.org/simpson/`.

[91] Y.R. Tausczik. Linguistic analysis of workplace computer-mediated communication. Master's thesis, University of Texas at Austin, August 2009.

[92] Y.R. Tausczik and J.W. Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–52, 2010.

[93] P.J. Taylor and S. Thomas. Linguistic style matching and negotiation outcome. *Negotiation and Conflict Management Research*, 1(3):263–281, August 2008.

[94] C.L. Toma and J.T. Hancock. What lies beneath: The linguistic traces of deception in online dating profiles. *Journal of Communication*, 62:78–97, 2012.

[95] A. Vrij, K. Edward, K.P. Roberts, and R. Bull. Detecting deceit via analysis of verbal and nonverbal behavior. *Journal of Nonverbal Behavior*, 24(4):239–264, Winter 2000.

[96] A. Vrij, W. Kneller, and S. Mann. The effect of informing liars about criteria-based content analysis on their ability to deceive CBCA-raters. *Legal and Criminological Psychology*, 5:57–70, 2000.

[97] A. Vrij and S. Mann. Telling and detecting lies in a high-stake situation: The case of a convicted murderer. *Applied Cognitive Psychology*, 15:187–203, 2001.

[98] The Washington Post. Republican presidential debate (full transcript), October 11 2011. Accessed in fall 2011 at `http://www.washingtonpost.com/politics/republican-debate-transcript/2011/10/11/gIQATu8vdL_story.html`.

[99] J.T. Webb. Subject speech rates as a function of interviewer behaviour. *Language & Speech*, 12:54–67, Jan-Mar 1969.

[100] C.H. White and J.K. Burgoon. Adaptation and communicative design: Patterns of interaction in truthful and deceptive conversations. *Human Communication Research*, 27(1):9–37, January 2001.

[101] S. Yue. A bivariate gamma distribution for use in multivariate flood frequency analysis. *Hydrological Processes*, 15:1033–1045, 2001.

[102] L. Zhou, J.K. Burgoon, Jr. J.F. Nunamaker, and D. Twitchell. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communication. *Group Decision and Negotiation*, 13:81–106, 2004.

[103] L. Zhou, J.K. Burgoon, D.P. Twitchell, T. Qin, and J.F. Nunamaker Jr. A comparison of classification methods for predicting deception in computer-mediated

communication. *Journal of Management Information Systems*, 20(4):139–165, Spring 2004.

[104] L. Zhou, Y. Shi, and D. Zhang. A statistical language modeling approach to online deception detection. *IEEE Transactions on Knowledge and Data Engineering*, 20(8):1077–81, August 2008.

[105] L. Zhou, D.P. Twitchell, T. Qin, J.K. Burgoon, and J.F. Nunamaker Jr. An exploratory study into deception detection in text-based computer-mediated communication. In *Proceedings of the 36th Hawaii International Conference on System Sciences*. IEEE, 2003.

[106] M. Zuckerman, B.M. DePaulo, and R. Rosenthal. Verbal and nonverbal communication of deception. *Advances in experimental social psychology*, 14(1):59, 1981.