

Detecting Deception in Interrogation Settings

C.E. Lamb and D.B. Skillicorn

December 2012

External Technical Report

ISSN-0836-0227-

2012-600

School of Computing

Queen's University

Kingston, Ontario, Canada K7L 3N6

Document prepared December 18, 2012

Copyright ©2012 C.E. Lamb and D.B. Skillicorn

School of Computing

Queen's University

{carolyn,skill}@cs.queensu.ca

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. Because of the requirements of many languages, including English, as well as the theory of verbal mimicry, such influence ought to exist. We show that it does: words in a question can “prompt” other words in the answer. However, the effect is receiver-state-dependent. Deceptive and truthful subjects respond to prompting in different ways. The accuracy of a bag-of-words deception can be improved by training a predictor on both question words and answer words, allowing it to detect the relationships between these words. This approach should generalize to other bag-of-words models of psychological states in dialogues.

Detecting Deception in Interrogation Settings

C.E. Lamb and D.B. Skillicorn
School of Computing, Queen's University
Technical Report 2012-600

1 Introduction

From court cases to airport security, many high-stakes real-world situations rely on the human ability to distinguish truth from deception. Sadly, humans do not do this well. Automating the detection of deception would thus bring many benefits. If an automated method outperformed humans, it could be used to augment human judgement in high-stakes situations; and it could also be used in lower-stakes situations where expert human judgement is infeasible. For instance, a filter alerting users to potential deception could be used with online job postings.

One obstacle in the way of automation is that deception works differently in different situations. One important factor mediating cues to deception is the difference between monologues and dialogues. Many deception models are based on monologues such as speeches or written statements. In many interesting real-world cases, the deceptive person is in a dialogue, and each participant responds and adjusts to the linguistic style of the other. This makes deception detection in a dialogue complicated. If one participant shows linguistic 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.* [43]. We found that word classes relevant to the model were vulnerable to a prompting effect: using certain words in the question increased the frequency of other words in the answer. Therefore the model cannot be applied to dialogue data in its present form, because the prompting effect is a potential confound.

Our goal, therefore, was to expand the model's applicability by removing the effect of the question from each answer. However, using this correction on real-world data revealed that our assumptions had been overly simplistic. Deceptive and non-deceptive respondents in the data responded to prompting in different ways, and removing the prompting effect also removed some of the signals of difference between these subgroups. Instead of removing the effect of the question, the relationship between question and answer words needs to be taken into account. Tests with a predictor confirmed that a model based on question and answer words, rather than just answer words, is indeed more accurate.

We conclude that, 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, prompting occurs, but the respondent's mental state determines its effect. Fortunately, this second-order behavior is itself detectable. We suspect that this effect will generalize to other word-based psychological models in dialogues, helping reveal other aspects of the relationship between

questioner and respondent.

2 Background

2.1 Stop Words

In text mining it is common practice to remove certain “stop words”: words like “the”, “I”, and “if”, which are thought to be too common to be worth analyzing [31]. 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. However, stop words contain interesting meanings of their own. Aside from their grammatical function, they have contextual, social meanings. (The word “he”, for example, refers to a man, but we need context in order to figure out *which* man.) They also serve a stylistic function. For this reason they are also called “style words” or “function words”.

Function words are particularly useful for the discovery of a speaker’s mental state for two reasons. First, they are more plentiful than content words. Although normal English speakers have 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 [55]. Second, these words are produced in a different part of the brain than content words [39] and can “leak” information about a person’s mental state even without their awareness [17].

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. One important tool for this analysis is Pennebaker’s [45] Linguistic Inquiry and Word Count program (LIWC). With this program, researchers have discovered correlations between function word use and everything from depression [49] to sexual orientation [25] and quality of a relationship [50] and, our focus here, deception [43]. Function words have also been shown to distinguish among different kinds of Islamist language [35].

2.2 Deception Detection

Unaided humans are poor at detecting deception. Although individual proficiency varies, even groups of trained humans such as police officers rarely perform above chance [22]. Many attempts have been made to construct empirical models of deception, relying on objective cues, but this is difficult because different studies produce conflicting results. In fact, no single cue is reliably indicative of deception across studies [8]. One reason for this is that different studies address the issue of deception in different situations, and the relevant cues in these situations may also be different. For example, DePaulo *et al.*’s meta-analysis found that effect sizes were different depending on the deceptive person’s motivation, the amount of interactivity in the setting, and whether a social transgression was involved [21]. Meanwhile, computer-mediated communication appears to function along different lines from face-to-face communication, perhaps because each person has time to rehearse what they will say [61].

2.3 The Pennebaker Deception Model

The model of deception we use comes from Newman *et al.* [43]. The authors asked college students to speak or write, either deceptively or truthfully, about a variety of topics (their opinion on abortion, 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 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 [43].
- Exclusive words decrease in the deceptive condition. Such words introduce increased complexity into sentences. Deception causes a higher cognitive load than truthfulness because of the increased difficulty of monitoring a deceptive performance [21], which produces an increase in visible signs of effort [62], even when the deceptive person has had time to prepare [56].
- Negative emotion words increase in the deceptive condition, consistent with DePaulo *et al.*'s [21] meta-analysis. Deceptive people are thought to do this as a sign of unconscious discomfort [43]. Alternatively, increased emotion can be a sign of trying too hard to persuade the listener, as in Zhou *et al.* [61].
- Motion verbs (“go”, “run”) increase in the deceptive condition. This is the result of increased cognitive load and perhaps a need to keep the story moving and discourage second thoughts on the part of the reader/hearer.

Lowered self-reference is consistent with other models (e.g. Zhou *et al.* [59]) and persists regardless of motivation [27] but DePaulo *et al.* [21] and Zuckerman *et al.* [62], in their meta-analyses, did not find a reliable, statistically significant decrease in self-references across studies.

DePaulo *et al.* [21] point out that increases in emotion should be treated with caution. Most lies in the real world are “white lies”, told to smooth over social interactions, and people telling them experience very little distress [20]. Moreover, some truthful statements, such as confessions of wrongdoing, might lead to high levels of negative emotions. For similar reasons, this part of the model should not be applied to psychopaths, who experience no discomfort when breaking moral rules [47].

The effect of cognitive load should also be treated with caution: some white lies are easy to tell, and some potentially volatile truths may be as mentally effortful as a lie [21].

Since the original work, the Pennebaker model of deception has been extensively validated across a large number of populations. The words used in these criteria are summarized in Table 1.

The LIWC model performed significantly better than human raters in distinguishing truth from deception [43]. Gupta and Skillicorn [26] suggest that the LIWC model detects

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 1: Words used in the LIWC deception model

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 [51].

2.4 Fine-Tuning the LIWC Model

Keila and Skillicorn [34] applied LIWC to a dataset consisting of internal emails at Enron in the period just before its fraudulent accounting scandal. They combined LIWC with singular value decomposition (SVD). SVD is a useful tool for eliciting correlations from multidimensional data. An SVD 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 [53]. The matrices U and V represent the data as vectors in a many-dimensional space. 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 an indication of how much variation is being lost [52].

Reducing the number of dimensions in this manner not only compresses the data but elucidates correlations between different parts of it, because documents or variables that vary together will be represented as closely aligned vectors in the SVD space [19].

Keila and Skillicorn [34] found that LIWC performed well when combined with SVD, because words within a given category (except exclusive words) were correlated with each

other and formed sensible structures in semantic space. The top ranked emails by a geometrical measure in reduced-dimension semantic space were all deceptive. However, in such a large dataset, it is difficult to tell how many false negatives there are.

Skillicorn and Little [52] 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 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.) Hence the standard Pennebaker model does not appear to be effective in dialogue settings.

If deception decreases first-person pronoun use in some situations and increases it in others, this explains DePaulo *et al.* [21] and Zuckerman *et al.*'s [62] 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 [52] 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. As we shall see, there is a better explanation.

2.5 Forcing Word Use in Questions and Answers

There are two processes of interaction between the language of questions and the language of answers that alter the word use in both. The first is the technical requirements of the language itself; the second is the largely unconscious mimicry that participants in a conversation fall into.

In most languages, the function word patterns in questions force some corresponding structure into a responsive answer. For example, a question containing the phrase “Did you . . .” requires either a first-person singular or first-person plural pronoun (“Yes, we did . . .”; “No, I didn’t”) or a passive verb. A respondent therefore does not have the same freedom of word use in a dialogue that they have in an unforced setting.

Two people in conversation, even if they are strangers, will imitate each other in everything from facial expression and body language [9] to volume [40], pitch [33], and speech rates [58]. People speaking to each other also mimic each other’s words and phrases [36]. This mimicry is generally not conscious [9]. Subjects mimic phrases they have heard even when consciously trying not to [3] and when their conscious working memory is filled with a distractor task [36].

LIWC can be used to measure verbal mimicry on the level of categories of words, where it is called Linguistic Style Matching (LSM). LSM is measured by comparing LIWC counts on all function word categories between both partners in an interaction. This comparison can be done through product-moment correlation [44] or a weighted difference score [32].

The LSM metric 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 [44] to letters between well-known colleagues [32] and face-to-face discussions [44].

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 [32]. However, while greater LSM is associated with greater social cohesion, the matching occurs even between strangers who dislike each other [44]. We would thus expect a degree of matching to occur in any context – even a courtroom interrogation.

3 Method and Results

The prompting effect we expect to see is a generalization of style matching because, unlike pure mimicry, we expect that some categories of question words prompt *different* categories of response words. This is particularly obvious in the case of pronouns. Our first hypothesis is therefore:

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 from the Pennebaker deception model.

Our second hypothesis follows:

H2. If word frequencies from the LIWC model are affected by the words in the question, then this potentially distorts the results of the Pennebaker model. Removing the effects of the question from the answer before applying the LIWC model should 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 develop a method to correct for these changes. Third, although not every answer in our data can be labeled as entirely truthful or entirely deceptive, we use some straightforward methods to validate the corrections and to investigate whether they do, in fact, improve the separation between truthful and deceptive responses.

3.1 Datasets

We created three datasets by taking archival transcriptions of real-life, high-stakes question-and-answer interactions.

The REPUBLICAN dataset comprises transcripts of each of the Republican primary debates leading up to the American presidential election of 2012. These involved a total of ten candidates and were televised and transcribed online by various news organizations [1, 4–7, 10–16, 18, 24, 30, 42, 46, 48, 57].

We used the REPUBLICAN dataset to investigate overall patterns of interaction between question and answer words. We did not rate the Republican presidential candidates as

deceptive or non-deceptive, since there is no reliable estimate of 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, and 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 motivated towards persona deception as defined by Gupta and Skillicorn [26]. Presenting themselves and the facts in the most favorable possible light is essential to getting elected, and much of the time this favorable light does not correspond exactly with reality. The full REPUBLICAN dataset contained 2118 question-answer pairs and 301,539 total words.

The NUREMBERG dataset comprised selected examinations and cross-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 [29]. Two were instead taken from the Nuremberg Medical Trial, which is partially transcribed online at the website of the Harvard Law Library [41]. Unlike the REPUBLICAN dataset, the NUREMBERG dataset 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. 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, so we chose to truncate answers in the NUREMBERG dataset at 500 words, matching the maximum size that occurred naturally in the REPUBLICAN dataset. This affected less than 1 percent of the answers. The full NUREMBERG dataset 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.

We also make brief use of a third dataset, SIMPSON. This contains depositions from the civil trial of O.J. Simpson for the wrongful death of his wife, Nicole Brown, and another man. Extensive transcripts of this material were available online [54]. 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. (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 dataset, we used only every third question-answer pair from him. The SIMPSON dataset contained 20,810

question and answer pairs, totalling 412,385 words.

3.2 Model Words

When processing this data, we are most interested in what occurs at the level of a single question and answer pair. Thus, we looked at “windows” in the data, which we defined as a single question followed by its answer. We recorded the length (in words) of each question and answer, along with the count of the number of words in these categories:

FPS First person singular pronouns. The Pennebaker 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 [37] found that deceptive people used a higher rate.

but Lower rates of this and other exclusive words are expected in deception. The Pennebaker model puts “but” and “or” in the same exclusive words category but Little and Skillicorn’s results [37] suggest that the two words often have quite different properties in semantic space. We therefore counted “but” and “or” separately.

or Another exclusive word.

excl The other exclusive words from the model, for example, “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. Alternatively, a respondent who is forced grammatically to use a first person pronoun might choose a plural one instead of a singular as a distancing technique. Similarly, if the questioner is referring to a group to which both they and the respondent belong, 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.

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

“Wh” 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 “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 person, and thus give them less opportunity to talk about themselves. When talking about another person – not oneself, and not the questioner – it might 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 An early analysis with a part-of-speech tagging program showed that rates of “these”, “those”, and “to” in the question were weakly correlated with rates of FPS in the answer. This early analysis did not yield other interesting results.

3.3 Gaussian Distributions

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 single question or answer by the total number of words it contains, giving a rate statistic for each type of word.

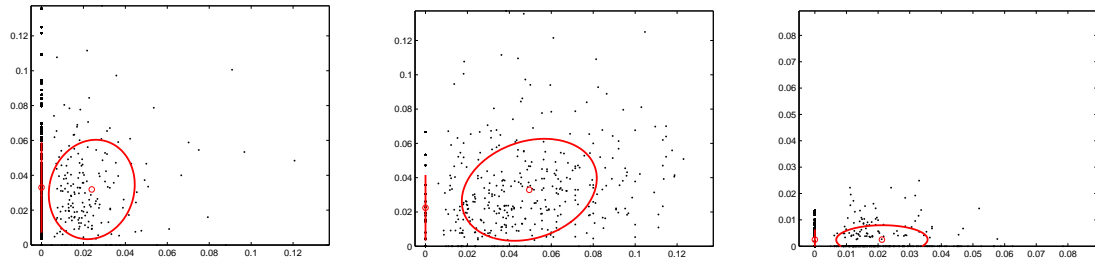
We expected the shortest questions and answers to be difficult to analyze. In a five-word answer with one “but”, the rate statistic for “but” would be 0.2 – unusually large. It isn’t clear that the answer has all the properties associated with use of the word “but” to a greater degree than, say a 16-word answer with one “but” which would have a rate of only 0.0625. Based on this reasoning and some early inspection of the data, we began the analysis with a minimum window size of 50 words in both question and answer.

We separated answers into two categories for each pair of question word and response word. The “unprompted” category contained all answers where the matching question did not contain the relevant question word. We assume that rates in such answers reflect their natural rates without prompting. The “prompted” category contained all answers in which the relevant question word appeared at least once.

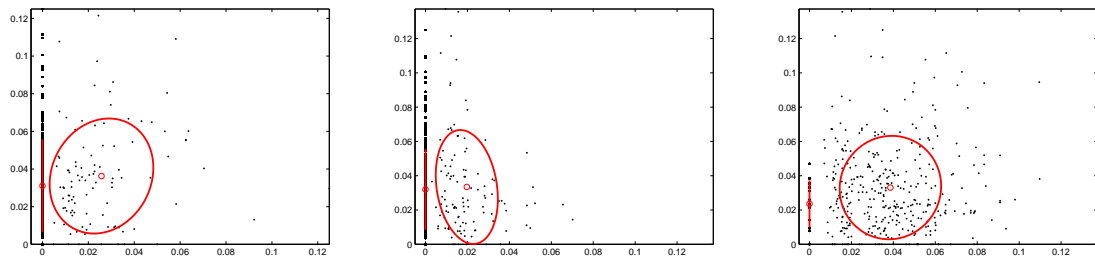
For each pair of question word and answer word in the prompted category, we counted the rates in all windows and fitted a two-dimensional Gaussian distribution to the data, using the FitFunc toolbox for Matlab. For the unprompted category we computed a one-dimensional Gaussian fitting the rates in the answers.

These Gaussian distributions provide an indicator of the relationship between the question word and the answer word pair. A distribution with a positive slope and a mean higher than that of the unprompted data indicates that the question word prompts higher rates of the answer word. A distribution with a negative slope and a mean lower than that of the unprompted data indicates that the question word prompts lower rates of the answer

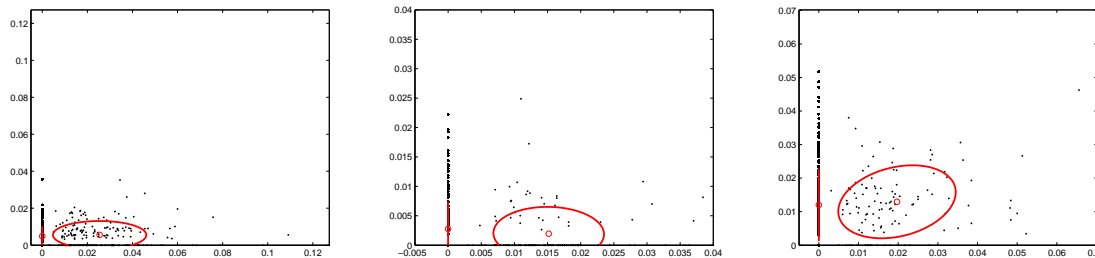
word. A flat or near-spherical distribution with a mean close to that of the unprompted data suggests that there is no relationship between the two word categories.



(a) First-person plural pronouns prompting first-person singular pronouns (b) Second-person pronouns prompt first-person singular pronouns (c) “Wh” words prompting “or”. The other exclusive word categories looked similar.



(d) Third person singular pronouns prompt first-person singular pronouns (e) Third person plural pronouns prompt first-person singular pronouns (f) “These”, “those”, and “to” prompt first-person singular pronouns.



(g) First-person singular pronouns prompt “but” (h) “But” prompting “or” (i) Third person plural pronouns prompt action words

Figure 1: Gaussian distributions from the REPUBLICAN dataset – minimum window size of 50 words. x -axis = rates for answer words, y -axis = rates for question words; the red line parallel to the y -axis shows the one-dimensional unprompted distribution out to 1 standard deviation.

Our initial estimates of the question word categories that would have significant prompting effects had mixed success. We did not find evidence of many of the relationships we had expected (Figures 1). However, we did find that higher rates of second-person pronouns in the question, as expected, were associated with higher rates of first-person singular pronouns in the answer. So were “these”, “those”, and “to”. With third-person plural pronouns, we found a weak effect in the opposite direction from what we expected: higher third-person plural pronoun rates prompted very slightly higher first-person singular pronoun rates. Although most pairs of question and answer word categories produced no effect, a few pairs for which we had no hypothesis also showed a prompting effect.

3.4 Data Correction

To remove the effect of prompting words, we carried out a series of affine transformations on the points in a bivariate Gaussian model in Cartesian space – producing corrected rates for each word category.

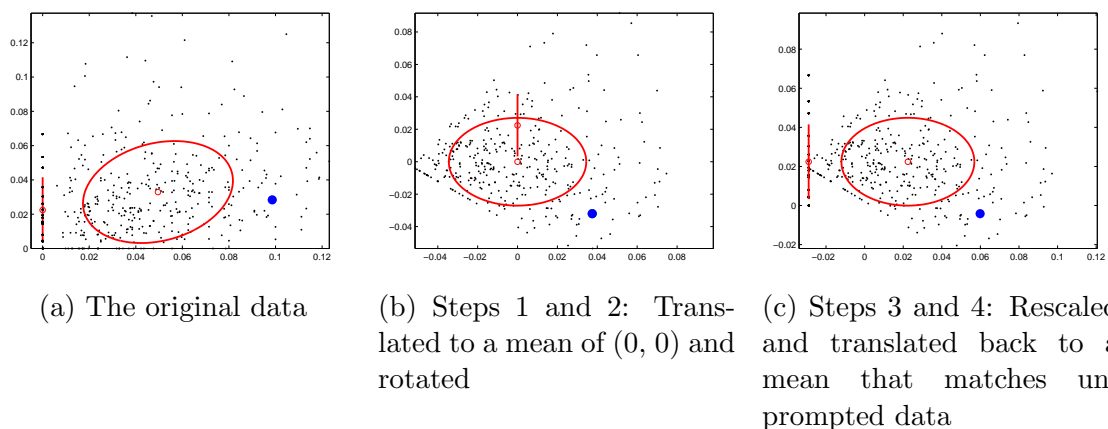


Figure 2: Steps in the correction process illustrated using second-person pronouns in the question and first-person singular pronouns in the answer. The blue point represents a particular speech as it is transformed.

The correction method is illustrated in Figure 2. For each question-answer pair, we fit a bivariate Gaussian to the prompted data and translate it so the mean is at (0,0). We then rotate the distribution using a standard rotation matrix until it is “flat” (one of its axes was parallel to $y = 0$). We use the smallest possible angle of rotation in either direction.

After that, we rescale the data on the y -axis so that its standard deviation in the y direction equals the standard deviation of the unprompted data, and translate it so that its mean returns to its original value in the x direction, and is equal to the mean of the unprompted data in the y direction.

We have included a blue dot in Figure 2 representing an example window (in this case, the respondent is Newt Gingrich) so that it can be traced through each step of the process. 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 increased first-person singular pronouns with more prompting. After the data is corrected, Gingrich’s first-person singular pronoun rate is low, as expected.

During this process, some windows can be assigned slightly negative values. Obviously it is impossible for a person to say fewer than zero words in response to a question. We treat the negative values as very emphatic zeroes, but do not correct them until the end of the process, since a subsequent correction for another word pair might return them to positive values.

We performed these transformations for each question-answer pair in the data, feeding the input from one transformation to the next so that, for each response word, a correction is made for each prompting word in turn. For question-answer pairs where the prompting word had negligible effect on the response word rate, the effect of this correction would also be negligible. There was a risk of these small 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 didn’t.

Examples of the changes to rate statistics are shown in Figure 3. It is useful to see how large a change in word frequencies such a correction represents. An average absolute value of three words were added or taken away from each window in both the FPS categories – 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. “But” and action words changed by an average of one word per window, but some windows change in a positive direction and some in a negative, so the average change is much less than one word per window. The other categories, on average, were barely changed.

To examine the sensitivity of the correction method we reran the corrections for each category of answer words after removing the 2.5% highest rates and the 2.5% lowest rates for each category (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 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.

3.5 Window Sizes

In many settings, the available window sizes are much smaller than 50 words. For example, in the SIMPSON depositions, the average question and answer contains fewer than 20 words in *both* sides of the window put together, and vanishingly few met the criterion of having 50 words or more in both the question and the answer. So we investigated ways to apply our model to smaller windows.

We experimented with the REPUBLICAN dataset, using versions with smaller minimum window sizes – 30, 10, and 1. We also tried merging adjacent questions and answers, provided

that they involved the same questioner and respondent, repeating this until all windows either met the minimum size for both question and answer, or could not be merged adjacent windows because there were none. We then removed any windows that still did not meet the minimum size. We call these composite windows.

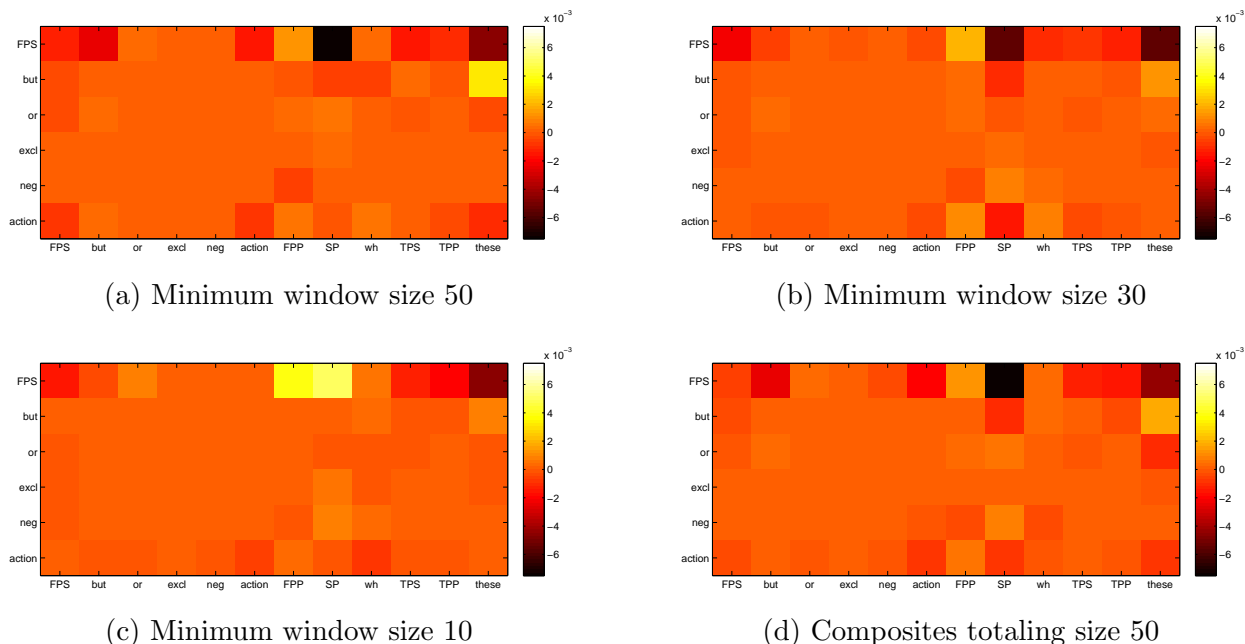


Figure 3: Color maps showing the average change in answer rates as the result of corrections for each question-and-answer pair at each minimum window size. Prompting words are the columns and response words the rows; bright colors indicate that question words force lower rates of answer words, and so answer word rates are corrected to increase; dark colors indicate the opposite. All maps are on the same color-based scale.

We then applied the corrections, measured the average change induced at each stage of the correction, and compared it to the average change in the original data. We created color maps that show the average correction to each question-answer pair (Figure 3). The patterns in 50-word windows degrade as the minimum window size is lowered, particularly to below 30. The most notable degradation happened in the 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 2 shows how much of each dataset is usable at each window size. For NUREMBERG and SIMPSON, small windows predominate, and there are many stretches where one individual answered many questions in a row, making them particularly amenable to the use of composite windows. With NUREMBERG and SIMPSON the composite window technique allowed analysis of much more data than 30-word windows. Furthermore, it caused less degradation than the small window sizes which would otherwise have been needed to cover this much data. In REPUBLICAN, windows tended to be larger, so the effect was less pronounced, but

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

Table 2: Number of words that could be included for each minimum window size

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 datasets) with composite windows.

3.6 Validation: Nuremberg and SVD

We expected corrected data to show a sharper distinction than the original data between the truthful and the deceptive. Using composite windows, we encoded the NUREMBERG data and checked that it had similar properties to the REPUBLICAN data. The magnitude of correction in FPS and action words was quite similar in the two datasets, although the effect on “but” the NUREMBERG data was much smaller.

We performed singular value decomposition on the answer data before and after performing the correction on the NUREMBERG dataset. This reduced the 6-category deception model to 3 dimensions. We then made a scatter plot showing each window in the resulting semantic space, expecting an increase in the distance between truthful and deceptive subgroups.

Figure 4 shows the result of singular value decomposition on the NUREMBERG data before and after correction. The effect is the opposite of what was expected – the correction erases most of the spatial distinction that existed between the groups. This suggests that, rather than being a confounding effect that should be removed to improve the detection of deception, the prompting effect contains important information *about* deception.

3.7 Validation: Subgroups in the Nuremberg Data

We analyzed each of the three NUREMBERG subgroups separately, applying a correction to each. Figure 5 shows color maps of this data. The differences between DEFENDANTS and UNTRUSTWORTHY WITNESSES (both the Nazi subgroups) are not large, but the response patterns of the TRUSTWORTHY WITNESSES subgroup are 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 that the average change in first-person singular pronouns prompted by second-person pronouns, for both the DEFENDANTS

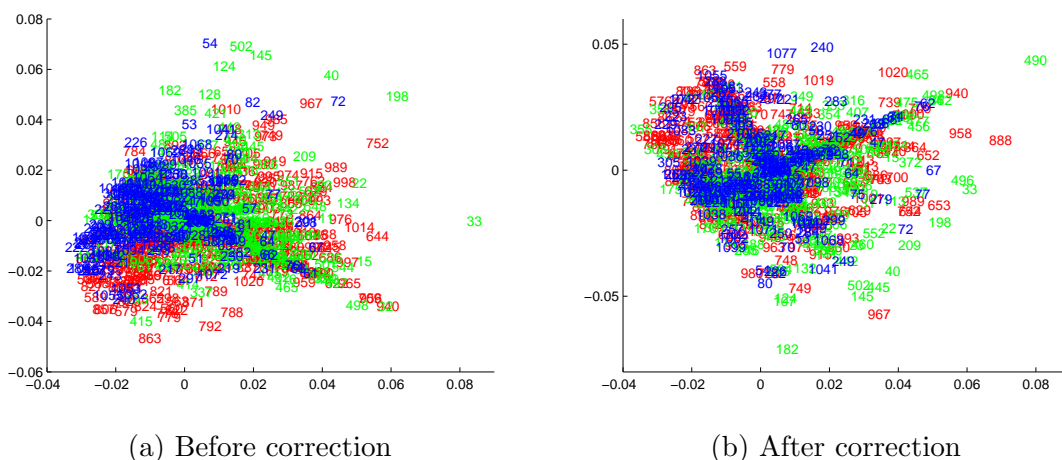
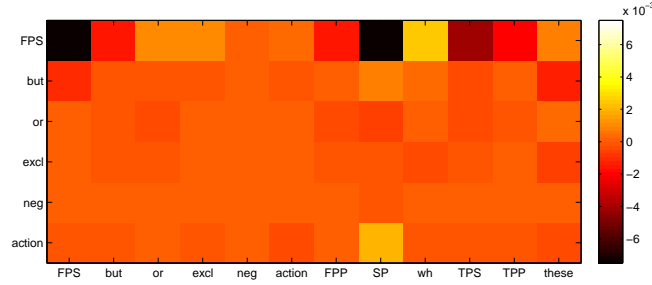


Figure 4: Singular value decomposition of the responses in the NUREMBERG dataset. DEFENDANTS are marked in red, TRUSTWORTHY WITNESSES in blue, and UNTRUSTWORTHY WITNESSES in green. Before correction, the TRUSTWORTHY WITNESSES are concentrated on one side of the semantic space, but this is less the case after the correction.

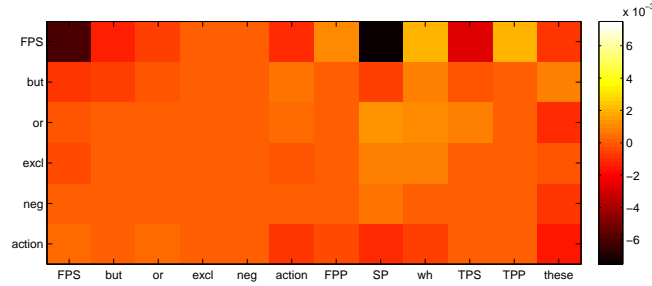
and UNTRUSTWORTHY WITNESSES, is literally off the scale. Both of these are more than twice the maximum amount shown by the color map; the DEFENDANTS' change is somewhat larger than the UNTRUSTWORTHY WITNESSES. (Since we were using composite windows for this, the REPUBLICAN average change is also slightly higher than it was in the previous colormaps, which were made with 50-word minimum windows.) Figure 6 shows this more clearly. (This is not a contradiction of our earlier claim that first-person pronoun corrections, in words per window, were similar in both the NUREMBERG and REPUBLICAN datasets. The NUREMBERG data had smaller windows, and it contained subgroups with both much higher and much lower rate statistics, on first-person pronouns, than the REPUBLICAN data.)

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 prompting differently. They were not simply being prompted at different rates, or showing different rates of response independently of the prompt – many question word/ response word pairs showed completely different distributions. In general, DEFENDANTS and UNTRUSTWORTHY WITNESSES, the two deceptive groups, were similar, but the TRUSTWORTHY WITNESSES group was different. The strongest effects tended to involve first-person pronouns or action words. Four of the most striking sets of distributions are shown in Figure 7.

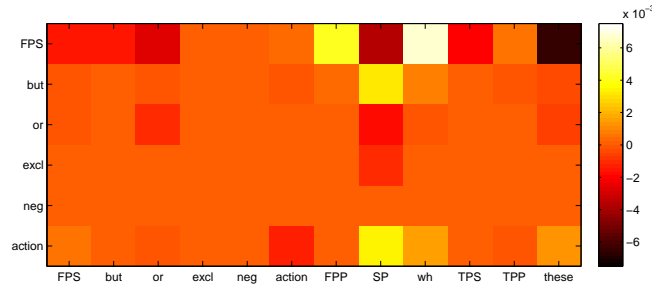
The prompting effect must be receiver-state-dependent – deceptive individuals respond to prompting one way, and truthful individuals another. One of the particularly strong differences involves second-person pronouns prompting first-person singular pronouns, which was already one of the strongest effects. Correcting for prompting reduced, rather than increased, the difference between these groups, because the prompting itself is a factor that



(a) DEFENDANTS



(b) UNTRUSTWORTHY



(c) TRUSTWORTHY

Figure 5: Corrections for the three NUREMBERG subgroups analyzed separately.

distinguishes 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.

These results explain the success of the *ad hoc* coding scheme used by Little and Skillicorn. Without knowledge of the rates of prompting words in questions, deceivers appear to increase their rates of first-person singular pronouns and exclusive words relative to less deceptive respondents. In fact, similar levels of prompting in questions to both groups produces these differentiated responses.

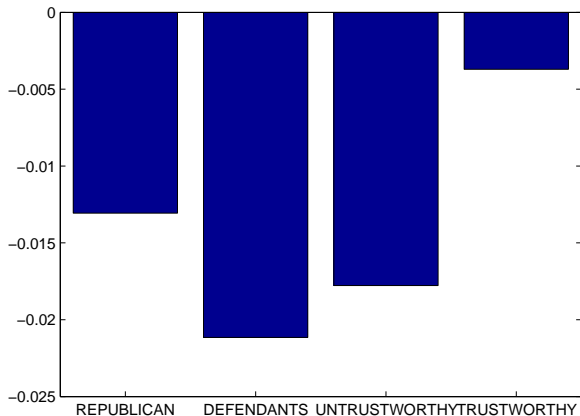


Figure 6: Average change in rates of first-person singular pronouns prompted by second-person pronouns in the REPUBLICAN and NUREMBERG datasets, all using composite windows.

3.8 Validation: Random Forests

Does information in the question actually improve the detection of deception or is that information present in the answer, but in some way not captured by the Pennebaker model? We used random forests [2] to estimate the significance of question words. A random forest not only classifies data but estimates the importance of each attribute in the data by counting how often each attribute is selected to act as the split point for an internal node of a decision tree of the forest.

We trained random forests, one with the rates of only the six response word categories in the answers, and one with these plus the rates of all the stimulus words in the questions. 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 from either of these forests. Rather, we wanted to compare them to each other to see if the question words made a difference. If both questions and answers were predictive of deception, then the model that uses both should make better predictions, and it should select the words that appear the most promising (i.e. first-person singular pronouns in answers and second-person pronouns in questions). If only the words in answers were relevant to deception, then both models should perform about the same.

Figure 8 shows the performance of both random forests. While the answer-words-only random forest performs above chance, it shows a strong bias: its accuracy with TRUSTWORTHY WITNESSES responders is much lower than its accuracy with DEFENDANTS, that is it tends to classify windows of both kinds as DEFENDANTS. Adding the question words improves overall accuracy by more than 10 percentage points – an even more dramatic result than we expected. Moreover, it reduces bias by an even larger amount, producing an improvement on the TRUSTWORTHY WITNESSES without reducing the accuracy on DEFENDANTS.

Table 3 shows the results of attribute ranking with the best-performing random forest.

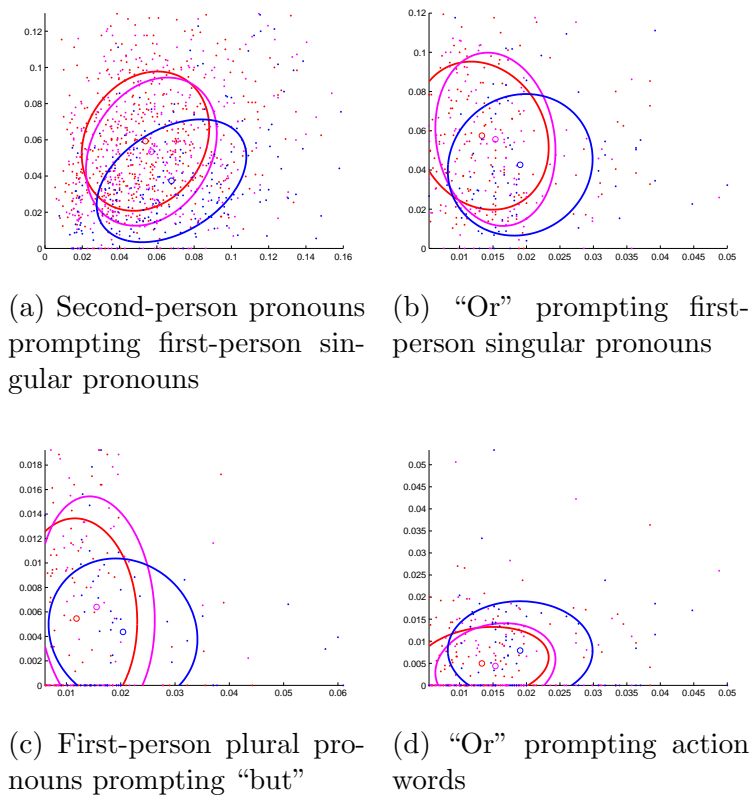


Figure 7: Gaussian distributions for the NUREMBERG dataset, separated by subgroup. DEFENDANTS are red, UNTRUSTWORTHY WITNESSES are magenta, and TRUSTWORTHY WITNESSES are blue. Not all of the figures are on the same scale.

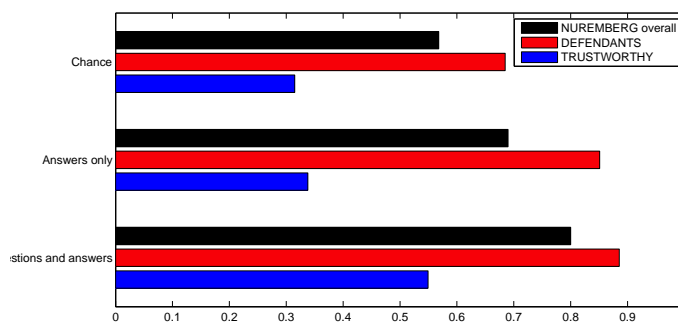


Figure 8: Prediction accuracy of random forests trained on the NUREMBERG data

The most influential words in the question-and-answer random forest were as predicted: first-person singular pronouns in the answer and second-person pronouns in the question, with similar frequencies. It is likely that the interaction between these two categories drove many of the decision trees in the forest.

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

After the 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 in the answer. Also, in some cases a question word may give context to an answer word and thus needs to be considered first.

3.9 Validation: the Simpson dataset

Our final task was to see whether these validations generalized. We looked at the SIMPSON dataset and performed the same analysis.

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 reduced this distinction (Figure 9). The average change in words per window for each answer word category was very similar to that of the NUREMBERG data.

The distinction between subgroups did not generalize as well. While Simpson’s deposition was somewhat different from those of the PLAINTIFFS (Figure 10), 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 11) was also different: there was

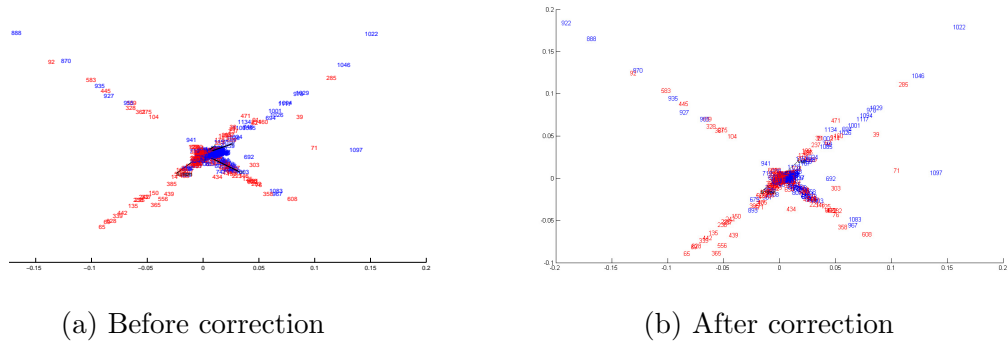
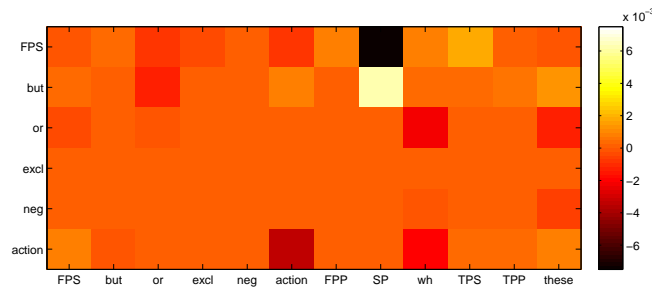
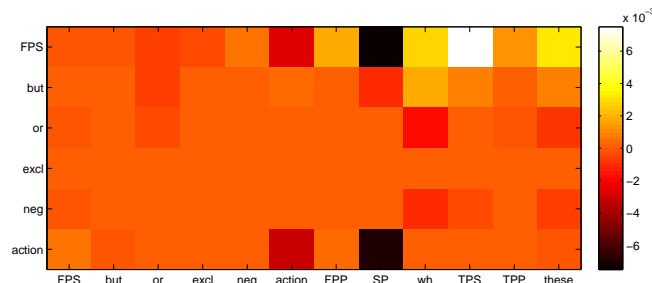


Figure 9: Singular value decomposition of responses in the SIMPSON dataset before and after correction, with O.J. Simpson’s responses in red and PLAINTIFFS in blue. The difference is very small.



(a) SIMPSON



(b) PLAINTIFFS

Figure 10: Corrections for the two O.J. Simpson subgroups analyzed separately.

evidence of differences between the two groups, as there had been with NUREMBERG, but not in quite the same way – they tended to be slightly weaker. 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 appeared much weaker between subgroups – SIMPSON used more first-person singular pronouns but was also

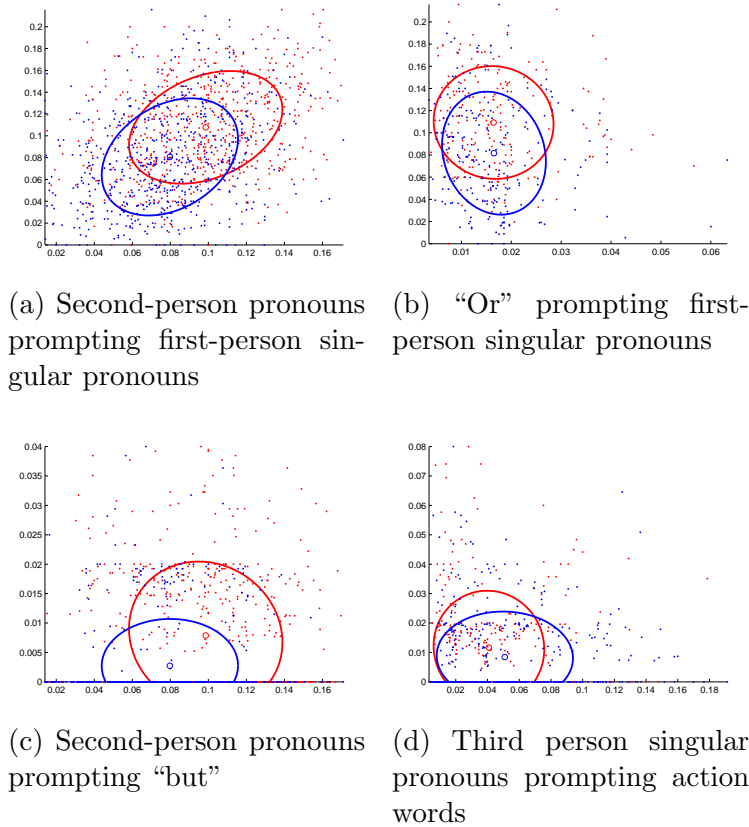


Figure 11: Gaussian distributions for the SIMPSON dataset, separated by subgroup. SIMPSON is red and PLAINTIFFS are blue. Not all of the figures are on the same scale.

prompted more.

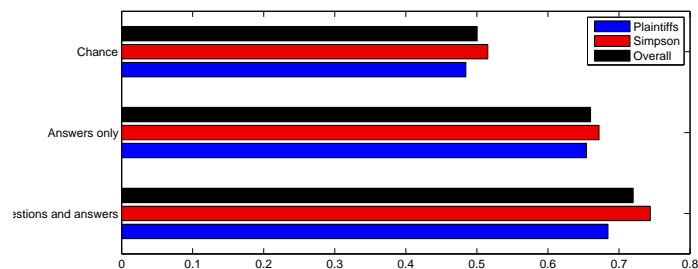


Figure 12: Prediction accuracy of random forests trained on the SIMPSON data

Finally we constructed the same two random forests using the SIMPSON data. The random forest trained with both question and response words showed an increase in accuracy, but a smaller one than that for the NUREMBERG data (Figure 12). In general, the SIMPSON data supports the view that prompting effects exist, but it is less clear where the prompting

effects actually are.

4 Discussion

At present, neither humans nor computers are particularly good at detecting deception. Models that perform with 70% accuracy (as in Mihalcea and Strapavara’s LIWC-based study [38]) are considered promising. But we can hardly afford for three in ten accused criminals to be falsely classified as deceptive. We also cannot rely on our own human ability to detect deception, since that ability is quite weak. Improvements to existing deception models are desperately needed.

We have increased the understanding of the processes that occur when a deceptive person is questioned. A prompting effect influences the rates of word usage in answers to questions, but the effect is receiver-state-dependent.

We developed a method to correct for prompting effects and so remove its influence. This method contains two components: a corrected representation of an utterance, and the size of the corrective change that is made. Each of these components might be useful in different situations. In our setting, the corrected representation was not useful because it removed some of the distinction between deceptive and truthful subgroups. However, tracking the size of the change was useful as it allowed us to see at a glance the similarities and differences in prompting effects between different subgroups.

The way in which respondents respond to the prompting effect is, in itself, a cue to deception. We support this conclusion by building random forests to classify deceptive and truthful subgroups, finding that the random forests performed substantially better when they were given both question data and answer data. In this situation, the random forests performed at 70-80% overall accuracy.

This suggests a number of avenues for ongoing research in deception. First, as the random forest results show, paying attention to both questions and answers will improve word-based models – even without detailed understanding of what the effect of question language is. Other classification approaches to deception will almost certainly benefit from including the words of questions in the data.

Beyond this, these results suggest what we ought to look for in the questions. 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. We have made a start at showing which relationships between words are most useful in such analysis – in particular, second-person pronouns prompting first-person singular pronouns.

Deception is not the only setting in which first-person pronouns are relevant. Any variable studied using bag-of-word approaches might be assessed in dialogue settings. In any such setting, it seems probable 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 that 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 from the association between relationship style or quality and the way the people involved respond to each other’s prompting. 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.

4.1 Limitations

Our analysis is biased towards common word categories. The “average change” metric measures not only the strength of a relationship between two pairs of words but how frequently that relationship actually appears. We defend this choice of metric by pointing out that the more a model relies on common words the more applicable it will be to small windows.

There are almost certainly important words that do not appear in this analysis. The Pennebaker model has been extensively validated, but other bag-of-word models have emerged. Hauch *et al.*’s recent meta-analysis [28] supports all the categories of the Pennebaker model but suggests several other cues to deception: increased positive and overall emotion words (as Zhou *et al.* found [60]), increased negations (“not”, “never”), decreased third-person pronouns, and slightly decreased tentative and time-related words.

The use of bag-of-words implies the treatment of words as forms without any semantics. There are therefore potential confounds associated both with polysemy, and with the use of function words in a stylized, rather than active, way (for example, whether “thank you” is considered to contain an active second-person pronoun or not).

The prompting effects in the SIMPSON dataset were somewhat different from those in the NUREMBERG dataset and generally smaller. The random forest model also showed a smaller improvement when question data was added. Intuitively, either the SIMPSON dataset underrepresents the difference between truthful and deceptive testimony, or the NUREMBERG dataset overrepresents it – or both. This illustrates the difficulty of comparing deceptive and truthful people across different contexts. In the SIMPSON dataset, it is plausible that the data underrepresents the difference between deception and truthfulness because the PLAINTIFFS were possibly not entirely truthful. They had a stake in the outcome, and were possibly motivated towards persona deception. Despite DePaulo *et al.*’s recommendations, little research has been done on the communication of people who are probably truthful, but highly motivated to “spin” the truth. Until such research is done, applying the deception models to civil suits remains problematic.

In the NUREMBERG dataset, the difference between truthfulness and deception may be exaggerated because of other differences between the Nazis and the TRUSTWORTHY WITNESSES. These groups generally spoke different first languages and were asked different types of questions; the TRUSTWORTHY WITNESSES were often given perfunctory cross-examination. Care must be taken in interpreting these differences in questions. As Hancock *et al.* [27] discovered, deception is a process involving both the questioner and respondent. If truthful and deceptive respondents are questioned differently, it may partly be because the questioner is biased – or it may be that the questioner is responding unconsciously to

deceptive speech patterns. To make matters worse, it may be both. As for demographic differences such as language and nationality, these cannot be ruled out as a potential confound, but their influence is probably small: researchers such as Fornaciari *et al.* [23] who tried to restrict their datasets to remove such confounds found that it did not increase the accuracy of their models. Nevertheless, the testimony of many of these witnesses was translated, and this may have distorted the language patterns, and certainly distorted the timing of questions and responses.

Any deception model used in the field will be faced with this type of confound: it will be used on 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 across all these different kinds of categories.

References

- [1] 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/>.
- [2] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.
- [3] A.S. Brown and D.R. Murphy. Cryptomnesia: Delineating inadvertent plagiarism. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(3):432–442, 1989.
- [4] 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>.
- [5] Cable News Network. Republican debate, June 13 2011. Accessed in fall 2011 at <http://transcripts.cnn.com/TRANSCRIPTS/1106/13/se.02.html>.
- [6] Cable News Network. Full transcript of CNN Arizona Republican presidential debate, February 22 2012. Accessed in winter 2012 at <http://archives.cnn.com/TRANSCRIPTS/1202/22/se.05.html>.
- [7] 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>.
- [8] 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.

- [9] T. L. Chartrand and R. van Baaren. Human mimicry. *Advances in Experimental Social Psychology*, 41, 2009.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] 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>.
- [19] 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.
- [20] 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.

- [21] B.M. DePaulo, J.J. Lindsay, B.E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper. Cues to deception. *Psychological Bulletin*, 129:74–118, 2003.
- [22] P. Ekman and M. O’Sullivan. Who can catch a liar? *American Psychologist*, 46(9):913–920, September 1991.
- [23] 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.
- [24] 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/>.
- [25] 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.
- [26] 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.
- [27] 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.
- [28] 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.
- [29] 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/>.
- [30] 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/>.
- [31] X. Hu and H. Liu. Text analytics in social media. In C.C. Aggarwal and C.X. Zhai, editors, *Mining Text Data*, pages 385–414. Springer Science+Business Media, 2012.
- [32] 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.

- [33] S.W. Gregory Jr., K.Dagan, and S.Webster. Evaluating the relation of vocal accomodation in conversation partners' fundamental frequencies to perceptions of communication quality. *Journal of Nonverbal Behavior*, 21(1), Spring 1997.
- [34] P. S. Keila and D. B. Skillicorn. Detecting unusual and deceptive communication in email. In *Centers for Advanced Studies Conference*, pages 17–20, 2005.
- [35] M. Koppel, N. Akiva, E. Alshech, and K. Bar. Automatically classifying documents by ideological and organizational affiliation. In *Proceedings of the IEEE International Conference on Intelligence and Security Informatics (ISI 2009)*, pages 176–178, 2009.
- [36] W.J.M. Levelt and S. Kelter. Surface form and memory in question answering. *Cognitive Psychology*, 14(1):78–106, January 1982.
- [37] 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.
- [38] R. Mihalcea and C. Straparava. The lie detector: Explorations in the automatic recognition of deceptive language. In *ACL-IJCNLP*, pages 309–312, 2009.
- [39] G.A. Miller. *The science of words*. New York: Scientific American Library, 1995.
- [40] 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.
- [41] 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/>.
- [42] 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>.
- [43] 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.
- [44] 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.
- [45] J.W. Pennebaker. Linguistic inquiry and word count. <http://www.liwc.net/>.
- [46] 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/>.

- [47] 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.
- [48] 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/>.
- [49] 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.
- [50] 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.
- [51] 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.
- [52] 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.
- [53] G.W. Stewart. On the early history of the singular value decomposition. *SIAM Review*, 35:551–566, December 1993.
- [54] Superior Court of the State of California. The Simpson trial transcripts, 1996. Accessed in fall 2012 at <http://walraven.org/simpson/>.
- [55] 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.
- [56] 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.
- [57] 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.
- [58] J.T. Webb. Subject speech rates as a function of interviewer behaviour. *Language & Speech*, 12:54–67, Jan-Mar 1969.
- [59] 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.

- [60] 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.
- [61] 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.
- [62] M. Zuckerman, B.M. DePaulo, and R. Rosenthal. Verbal and nonverbal communication of deception. *Advances in experimental social psychology*, 14(1):59, 1981.