LESLI
Linguistic Evidence in Security, Law and Intelligence

Volume 1, No. 1 (2013) | ISSN 2327-5596 (online) | DOI 10.5195/lesli.2013.1 | http://lesli-journal.org

Extending Textual Models of Deception to Interrogation Settings

David Skillicorn PhD
Professor, Computer Science, School of Computing, Queen's University, Canada
skill@cs.queensu.ca

Carolyn Lamb
Graduate Student, Computer Science, School of Computing, Queen's University, Canada
carolyn@cs.queensu.ca

Abstract

Models that detect deception in text typically outperform humans but are limited to single pieces of text created by a single individual. Text from dialogues and wider conversations reflects linguistic influence among the participants, and this intertwining makes it difficult to ascribe deception to any one of them. We address this problem in dialogues, particularly interrogations, by seeking to detect and remove the influence of the language of a question from the language of the response. Surprisingly, this does not work as expected: the response by a deceptive person to certain categories of words in questions is qualitatively different from that of a truthful person. Successful prediction of deception in responses, therefore, requires analysis using the words of both questions and answers. We show that such prediction is indeed effective.

Keywords: deception detection, dialogue, discourse structure

Introduction

Many real-world situations, ranging from court cases to airport security, rely on humans to determine who is being truthful and who is not. Unfortunately, human abilities to detect deception are inherently weak and even law-enforcement personnel are often trained to look for signals that are actually irrelevant (and sometimes contradictory – deceivers make too little eye contact, or they make too much).

Models that detect deception in text are well-developed but they assume that the text is free-form, that is, produced by a single individual with complete freedom to use whatever words are desired. Unfortunately, many situations where detecting deception from text would be useful (court cases, law enforcement interviews, job interviews,

Articles in this journal are licensed under a Creative Commons Attribution 3.0 United States License.

This journal is published by the University Library System, University of Pittsburgh as part of its D-Scribe Digital Publishing Program and is cosponsored by the University of Pittsburgh Press.

refugee claims) are not free-form. Language used by one participant in a dialogue or conversation "leaks" into the language used by the others. The result is language that intertwines the mental state of the speaker or writer, including their intent to deceive, and the mental states of the others. In this situation, determining deception is much more difficult.

We approach this problem by restricting our attention to dialogues, in particular interrogations, and attempt to determine, and then remove, the effect of question language on the language of each response. This technique is quite effective and has a number of other potential applications. However, it does not work as we expected to for detecting deception – responses to the same question language are qualitatively different from those trying to deceive and those being truthful. In other words, not only is the response language affected by mental state, the feedback loop between interrogator and respondent is also affected. This suggests that determining deception requires access to the language both of questions and of answers. We show that, given this, prediction accuracies for deception increases markedly.

Background

Function Words

In information retrieval, it is common to remove articles, pronouns, auxiliary verbs, prepositions, and determiners that are collectively known as function words. These words occur with high frequency but do not mark document content strongly – rather they are the framework in which content is placed (Hu & Liu, 2012). However, function words are particularly useful for the discovery of a speaker or author's mental state for two reasons. First, they are plentiful: although normal English speakers have only 500 function words in their vocabulary (out of perhaps 100,000 total English words), 55% of all the words they speak or write are function words (Tausczik & Pennebaker, 2010). Second, these words are produced in a different part of the brain than content words are (Miller, 1995) and can "leak" information about mental state, even outside of awareness (Chung & Pennebaker, 2007).

Social psychologists in the past twenty years have analyzed speaking and writing styles by counting both function words and common content words. One important tool is Pennebaker's (Pennebaker, 2013) Linguistic Inquiry and Word Count program (LIWC). With this program, researchers have discovered correlations between function word use and depression (Rude *et al.*, 2004); sexual orientation (Groom & Pennebaker, 2005); quality of a relationship (Simmons *et al.*, 2005); and, our focus here, deception (Newman *et al.*, 2003). Function words have also been shown to distinguish among different kinds of Islamist language (Koppel *et al.*, 2009) and to be predictive of future aggressive attacks by violent extremist groups (Pennebaker, 2011).

Deception Detection

Unaided humans are poor at detecting deception. Although individual proficiency varies, even groups of trained humans such as police officers rarely perform substantially better than chance (Ekman & O'Sullivan, 1991).

Microexpressions, fleetingly appearing expressions on human faces, have been shown to reveal aspects of mental state, including deception (Ekman, 2002). However, training humans to detect microexpressions is difficult. A large amount of effort has been made to automate the detection of microexpressions with, so far, limited success (Polikovsky *et al.*, 2012).

Textual detection of deception has obvious practical advantages. It does not require a human in the loop, at least initially, and it is both practical and cheap. Much material is, of course, already in the form of text, and speech-to-text software has almost reached the point of operating without per-speaker training.

Empirical approaches to deception detection from text have shown promise, but different studies have produced conflicting results. In fact, no single cue is reliably indicative of deception across studies (Burgoon *et al.*, 2012, Carlson *et al.*, 2004). 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 (DePaulo *et al.*, 2003). Meanwhile, computer-mediated communication appears to function along different lines from face-to-face communication, perhaps because participants have time to rehearse what they will say (Zhou *et al.*, 2003).

The Pennebaker Deception Model

The model of deception we use was developed by Pennebaker's group (Newman *et al.*, 2003). 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:

- First person singular pronouns decrease. Those being deceptive have less personal experience with their subject matter than those being truthful, and are less emotionally willing to commit to what they are saying, so they focus on and refer to themselves less (Newman *et al.*, 2003).
- Exclusive words decrease. Such words introduce increased complexity into sentences. Deception causes a higher cognitive load than truthfulness because of the need to construct situations that did not occur, and the increased difficulty of monitoring deceptive performance (DePaulo *et al.*, 2003). These produce an increase in visible signs of effort (Zuckerman *et al.*, 1981), even when the deceptive person has had time to prepare (Vrij & Mann, 2001).
- Negative emotion words increase. This is thought to be a sign of unconscious discomfort (Newman *et al.*, 2003), which is consistent with DePaulo *et al.*'s (2003) meta-analysis. Alternatively, increased emotion can be a sign of trying too hard to persuade the listener, as in Zhou *et al.* (2003).
- Action verbs ("go", "run") increase. 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.

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.

Lower self-reference is consistent with other models (e.g. Zhou *et al.* (2004)) and persists regardless of motivation (Hancock *et al.*, 2008) but DePaulo *et al.* (2003) and Zuckerman *et al.* (1981), in their meta-analyses, did not find a reliable, statistically significant decrease in self-references across studies.

DePaulo *et al.* (2003) 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 (DePaulo *et al.*, 1996). 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 (Porter & Yuille, 1996). The effect of cognitive load should

Volume 1, No. 1 2013 | DOI 10.5195/lesli.2013.1 | http:// lesli-journal.org

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 (DePaulo *et al.*, 2003).

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 Pennebaker deception model

The Pennebaker model performed significantly better than human raters in distinguishing truth from deception (Newman *et al.*, 2003), up to 70% accuracy in one experiment (Mihalcea & Straparava, 2009). Gupta and Skillicorn (2006) suggest that the Pennebaker model detects not only outright falsehood, but also "spin" or "persona deception", in which people do 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 the Pennebaker model to the speeches of politicians (Skillicorn & Leuprecht, 2012).

Fine-Tuning the Pennebaker Model

In the original Pennebaker model, words are taken as equivalent potential markers of deception. Generally speaking, documents are scored by counting, in each, the number of negative-emotion words and action verbs (positively related to deception) and subtracting the number of first-person singular pronouns and exclusive words (negatively related to deception). A deceptive document is then one that has a high score.

Counting ignores the relevance of base word frequency. For example, in business writing, first-person singular pronouns are rare, so the presence of only a small number might be a very strong indicator of veracity, whereas in a blog the opposite would be true. Scoring models of this kind cannot really be used to determine the absolute veracity or deceptiveness of a single document, but rather to compare the relative veracity or deceptiveness of the documents in a set or corpus. In other words, the model allows a set of documents to be ranked from most to least deceptive, with the implicit assumption that the documents are of the same general kind.

An alternative to scoring is to construct a matrix with one row corresponding to each document, one column to each model word, with the entries representing the frequencies of occurrence of each word in each document. In such a space, each document has a natural representation as a point in the space spanned by the columns, and points that lie close to one another correspond to documents that are similar. This space can be projected into a much lower dimensional space (two dimensions in our case) using a singular value decomposition (Golub & van Loan, 1996). If the model is accurate then the points should form an almost linear structure with the most deceptive documents at

one end and the least deceptive at the other. The deception score for a document is then determined from the position of its point along this structure. The use of singular value decomposition infers the importance of each individual word from its use across the entire corpus and so tends to produce more meaningful scores.

In practice, the resulting structure tends to contain two linear substructures, because individuals tend to differ idiosyncratically in the relative rates at which they use first-person singular pronouns and exclusive words. It is still the case that documents at one end of the structure are most deceptive and those at the other least deceptive (Skillicorn, 2012, Skillicorn, 2010, Keila & Skillicorn, 2005a).

Skillicorn and Little (2010) applied SVD-based analysis to transcripts of 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 (as well as the expected increases in negative-emotion words and action verbs). 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 people ranked most deceptive by the model were people who claimed not to remember basic facts about their own employment, such as who they were working for. Meanwhile, the people ranked least deceptive were witnesses called in to explain purely technical matters. Hence the standard Pennebaker model clearly fails in dialogue settings.

If deception decreases first-person pronoun use in some situations and increases it in others, this explains DePaulo *et al.* (2003) and Zuckerman *et al.*'s (1981) 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 (2010) 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.

Relationships between 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, word patterns in questions force some corresponding structure into a responsive answer. For example, in English, 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 choice 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 (Chartrand & van Baaren, 2009) to volume (Natale, 1975), pitch (Gregory Jr. *et al.*, 1997), and speech rates (Webb, 1969). People speaking to each other also mimic each other's words and phrases (Levelt & Kelter, 1982). This mimicry is generally not conscious (Chartrand & van Baaren, 2009). Subjects mimic phrases they have heard even when consciously trying not to (Brown & Murphy, 1989) and when their conscious working memory is filled with a distractor task (Levelt & Kelter, 1982).

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 function word categories between both partners in an interaction. This comparison can be done through product-moment correlation (Niederhoffer & Pennebaker, 2002) or a weighted difference score (Ireland & Pennebaker, 2010).

The LSM metric is internally consistent: if a pair match in their use of one function word category, they will probably match to the same degree in all others. The metric also generalizes well across different contexts, from

Volume 1, No. 1 2013 | DOI 10.5195/lesli.2013.1 | http:// lesli-journal.org

online chat conversations (Niederhoffer & Pennebaker, 2002) to letters between well-known colleagues (Ireland & Pennebaker, 2010) and face-to-face discussions (Niederhoffer & Pennebaker, 2002).

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 (Ireland & Pennebaker, 2010). However, while greater LSM is associated with greater social cohesion, the matching occurs even between strangers who dislike each other (Niederhoffer & Pennebaker, 2002). We would thus expect a degree of matching to occur in any context – even one as adversarial as a courtroom interrogation.

These results suggest that the language used in a question will affect the language used in the response. Trying to apply the deception model to answers as if they were free-form cannot be expected to perform well, since the data it is using is some kind of mixture of the language of the questioner and the language of the respondent. Thus we begin with the following hypothesis:

H1: Removing the effect of the language of each question from the language of the response should enable deception to be determined more clearly.

Creating Data

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. We might expect that use of "you" in the question has a strong connection to the use of "I" in the response, and this is indeed the case.

Our plan of attack has three parts. First, we gather archival question-and-answer data and visualize relevant changes in frequencies. Second, we develop a method to correct for these changes, that is to remove those occurrences of words in responses that can be accounted for as responses to prompting words in questions. 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 the responses of truthful and deceptive individuals.

Datasets

Datasets where the individual responses are labelled with truth values do not yet exist. We are therefore forced to use datasets where a propensity for deception can be attributed to some of the respondents, primarily trial transcripts where the outcomes can be taken as implications about who was motivated to be deceptive.

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 (Chicago Sun-Times, 2011c, American Broadcasting Company, 2011, Cable News Network, 2011a, Cable News Network, 2011b, New York Times, 2011, Cable News Network, 2012b, Cable News Network, 2012a, Fox News, 2011, Council on Foreign Relations, 2012, PolitiSite, 2011, Washington Post, 2011, RonPaul.com, 2011, Chicago Sun-Times, 2011a, Chicago Sun-Times, 2012c, Chicago Sun-Times, 2011b, Chicago Sun-Times, 2011d, History Musings, 2011, Chicago Sun-Times, 2012a, Chicago Sun-Times, 2012b).

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 (2006). 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 (1946). Two were instead taken from the Nuremberg Medical Trial, which is partially transcribed online at the website of the Harvard Law Library (National Archive, 1946-1947). 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, 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, 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, and 978 from TRUSTWORTHY). It contained a total of 311,099 words.

We also make 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 (Superior Court of the State of California, 1996). 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 his testimony. The SIMPSON dataset contained 20,810 question and answer pairs, totalling 412,385 words.

The length of responses varied from single words to extremely long discourses. Experimentation suggested that the effect of words in the question dissipated some distance into the response, perhaps starting at about the 50-word mark. On the other hand, rapid exchanges of short questions and short responses seemed to demonstrate an effect over more than one question and answer pair. This suggested that the effect is temporal rather than speech act dependent; in other words, question language effects linger as long as they remain in short-term memory, which is typically about the same time it takes to say 50-100 words. We therefore created windows of text by truncating long responses at 500 words, confident that effects of question language had long gone; and aggregating short questions and responses so that the response windows were no shorter than 50 words. We examine the effect of these decisions later.

Volume 1, No. 1 2013 | DOI 10.5195/lesli.2013.1 | http:// lesli-journal.org

Model Words

We recorded the count of the number of words in these categories in both questions and answers:

FPS First person singular pronouns. The Pennebaker model predicts that deceivers should use a lower rate of first-person pronouns than truthful people.

but, or These are exclusive words, but other results, particularly Little and Skillicorn's results (2008), suggest that these two often occur in ways that are uncorrelated with other exclusive words. Hence we counted them separately.

excl The other exclusive words, for example, "unless" and "whereas".

neg Negative emotion words.

action Action verbs.

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 a respondent to do likewise. A respondent who is forced grammatically to use a first-person pronoun might choose a plural one over a singular as a distancing technique. Similarly, if the questioner is referring to a group to which both she and the respondent belong, this might prompt the respondent to continue talking in the context of the 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 (possibly higher FPP 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.

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 word category.

We expected the shortest questions and answers to be difficult to analyze. In a five-word response with one occurrence of "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. This provided another motivation for aggregating questions and answers into units with a minimum size.

Algorithm

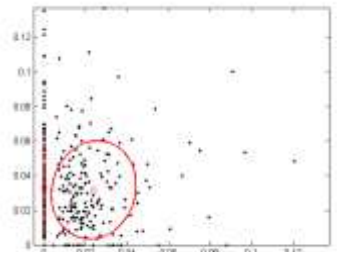
Fitting Distributions to the Data

For each pair of prompt word category and response word category (for example, second-person pronouns in the question and first-person singular pronouns in the response), we separated answers into two sets. The first set consisted of responses where the prompt word category *was not* used in the question but a member of the response word category appeared in the response (the "unprompted" set). The second set consisted of responses where the prompt word category *was* used in the matching question (the "prompted" set).

For the unprompted set, a univariate Gaussian distribution was fitted to the rates of occurrence of the response word across all responses it contains. The mean of this distribution estimates the mean unprompted rate for this word category, and its standard deviation estimates how much variation there is in such use.

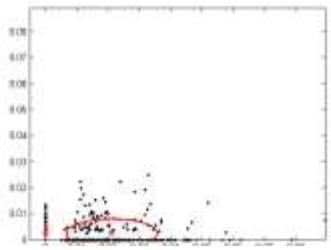
For the prompted set, a bivariate Gaussian distribution was fitted to the rate of occurrence of words from the prompt word category versus the rate of occurrence of words from the response category. This distribution estimates how responses rates depend on prompt rates for the entire set of questions and answers. Its one-standard-deviation contour is an ellipse that provided intuitive information about this dependency. The greater the area enclosed by this ellipse, the more variability in response. If the ellipse is elongated with a positive slope, this indicates that increases in prompt words stimulate increase in response words. If it has a negative slope, this indicates that increases in prompt words reduce the rate of response words. A flat or near-spherical ellipse indicates that there is little relationship between the rate of prompt words and the rate of response words.

Figure 1 shows the fitted distributions for a number of stimulus-response pairs for the REPUBLICAN dataset. In these figures, the height and extent of the bivariate distribution compared to the univariate distribution (the red vertical line on the left) allows the prompted response rate to be compared to the unprompted. For example, the rate of first-person singular pronouns ("I") in response to questions containing "these", "those" and "to" is much higher (but more variable) than in questions that do not contain these words. As expected, second-person pronouns ("you") in the question prompt high rates of first-person singular pronouns ("I") in responses. Third-person plural pronouns ("they") prompt reduced rates of first-person singular pronouns ("I"). Although most pairs of prompt and response word categories produced no effect, a few pairs for which we had no hypothesis also showed a prompting effect.



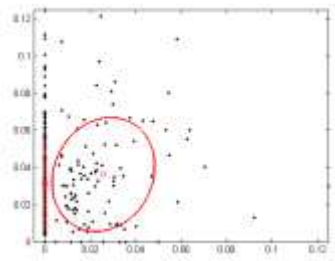
(a) First-person plural pronouns prompting first-person singular pronouns

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

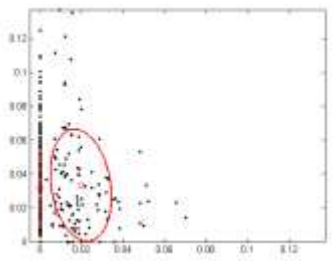


(c) "Wh" words prompting "or"

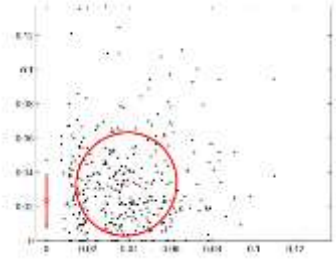
Volume 1, No. 1 2013 | DOI 10.5195/lesli.2013.1 | http:// lesli-journal.org



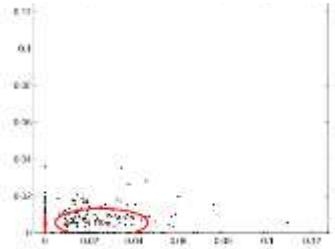
(d) Third-person singular pronouns prompting first-person singular pronouns



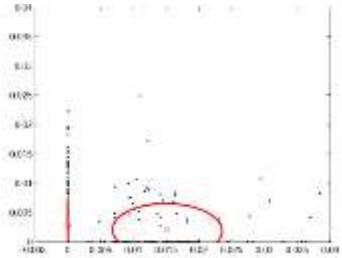
(e) Third-person plural pronouns prompting first-person singular pronouns



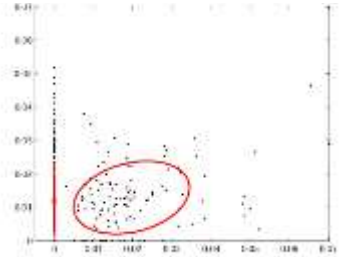
(f) "These", "those", "to" prompting first-person singular pronouns



(g) First-person singular pronouns prompting "but"



(h) "but" prompting "or"



(i) Third-person plural pronouns prompting action words

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

Transformation for 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.

The correction method is illustrated in Figure 2. For each question-response pair, we translate the bivariate Gaussian so that its 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.

Then 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. Thus the bivariate distribution now lies parallel to the *x*-axis with the same mean and vertical extent as the univariate (unprompted response) distribution.

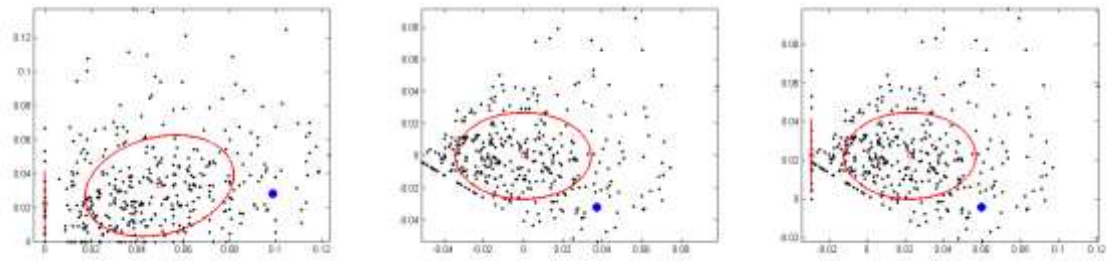


Figure 2: Steps in the correction process illustrated using second-person pronouns in the question and first-person singular pronouns in the answer. The blue point represents a particular speech as it is transformed.

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.

This correction is repeated for every pair of stimulus and response words. The altered bivariate distributions are now used to generate word rates for all responses in the prompted set, word rates that have been altered by the difference in the vertical position of the point corresponding to its initial coordinates and the point corresponding to its position given by the new distribution. In other words, transforming the fitted distribution can also be interpreted as translating each point in a way that reflects the transformation.

During this process, some responses 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 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 matter from effects that don'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 response 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 response, but some responses have changes in the positive direction and some in the negative, so the average change is much less than one word per response. 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,

Volume 1, No. 1 2013 | DOI 10.5195/lesli.2013.1 | http:// lesli-journal.org

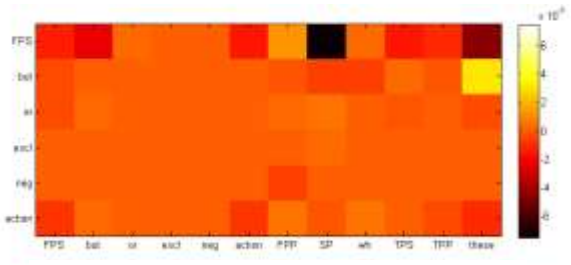
and the effect on "but" disappeared. This suggests that the correction method is somewhat sensitive to outliers, but that its results for FPS and action words are largely valid.

The Effect of Window Sizes

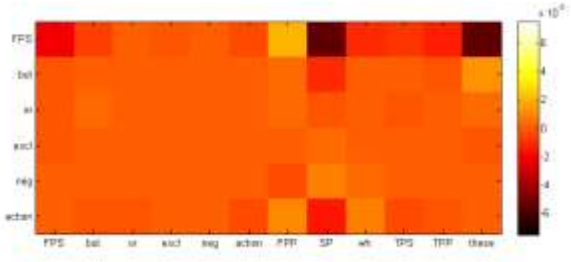
In many settings, the available question or response sizes are much smaller than 50 words. For example, in the Simpson depositions, the average question and answer together contains fewer than 20 words, and vanishingly few met the criterion of having 50 words or more in both the question and the answer.

To assess the impact of aggregating adjacent questions and responses, we experimented with the REPUBLICAN dataset, using versions with several smaller minimum window sizes – 30, 10, and 1; and merged adjacent questions and answers, provided that they involved the same questioner and respondent, until they met the minimum size or could not be merged further. We then removed any questions and responses that still did not meet the minimum size.

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 aggregated windows also showed slight degradation, comparable to that in the 30-word windows.

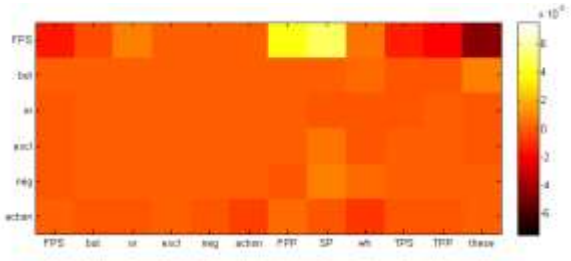


(a) Minimum window size 50

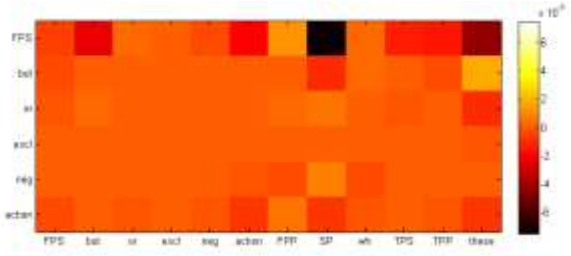


(b) Minimum window size 30

Volume 1, No. 1 2013 | DOI 10.5195/lesli.2013.1 | http:// lesli-journal.org



(c) Minimum window size 10



(d) Aggregates totaling size 50

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.

	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
Aggregated	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

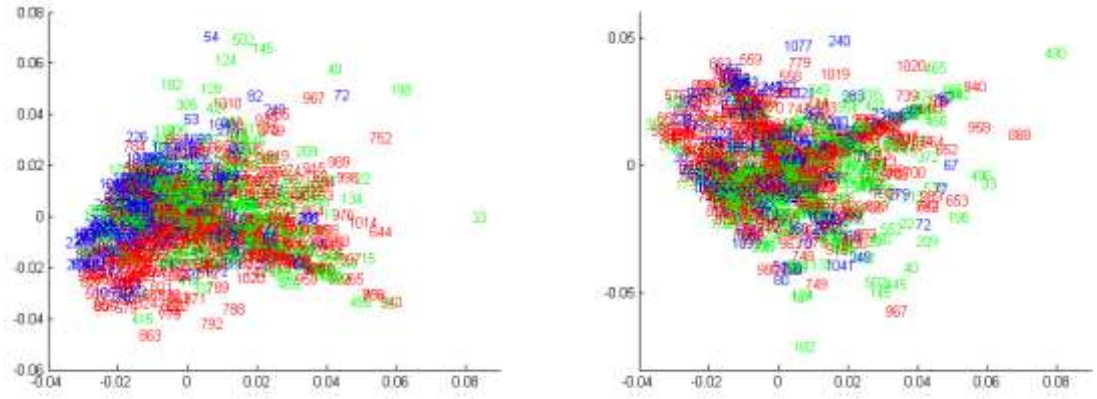
Table 2 shows how much of each dataset is usable at each window size. For NUREMBERG and SIMPSON, small windows predominate, and there are many stretches where one individual answered many questions in a row, making them particularly amenable to the use of aggregated windows. With NUREMBERG and SIMPSON the aggregated 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 the aggregated 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 aggregated windows.

Volume 1, No. 1 2013 | DOI 10.5195/lesli.2013.1 | http:// lesli-journal.org

Validation: NUREMBERG and SVD

We expected the corrected data to show a sharper distinction than the original data between the truthful and the deceptive. Using aggregated 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" in 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 2 dimensions. We then constructed a scatter plot showing each response in the resulting semantic space, expecting an increase in the distance between truthful and deceptive subgroups.



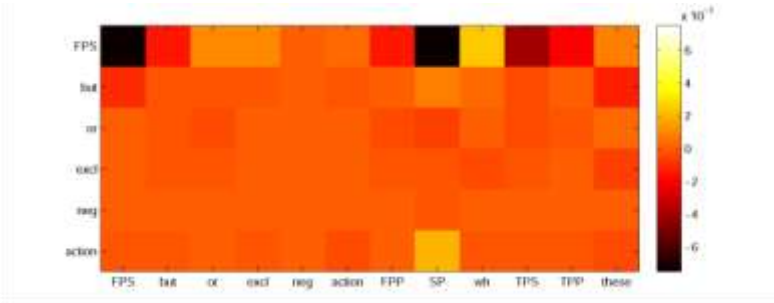
(a) Before correction

(b) After correction

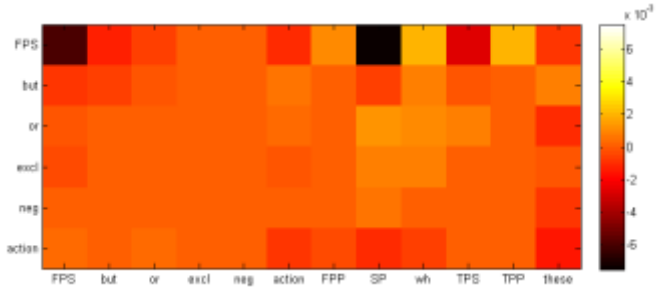
Figure 4: Singular value decomposition of the responses in the NUREMBERG dataset. DEFENDANTS are marked in red, TRUSTWORTHY in blue, and UNTRUSTWORTHY in green. Before correction, the TRUSTWORTHY are concentrated on one side of the semantic space, but this is no longer the case after the correction.

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, prompting reveals important information *about* deception.

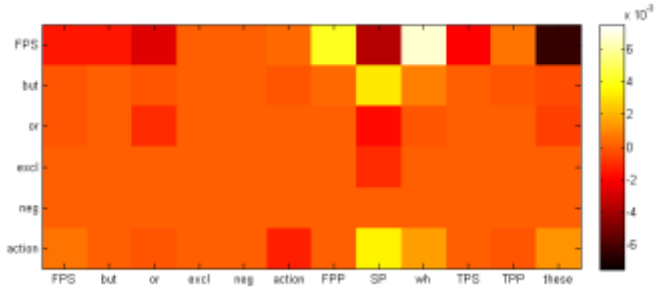
Volume 1, No. 1 2013 | DOI 10.5195/lesli.2013.1 | http:// lesli-journal.org



(a) DEFENDANTS



(b) UNTRUSTWORTHY



(c) TRUSTWORTHY

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

A Revised Hypothesis

We analyzed each of the three NUREMBERG subgroups separately, applying an independent correction to each. Figure 5 shows color maps for this data. The differences between DEFENDANTS and UNTRUSTWORTHY (both the Nazi subgroups) are not large, but the response pattern of the TRUSTWORTHY subgroup is 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 figure. 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

Volume 1, No. 1 2013 | DOI 10.5195/lesli.2013.1 | http:// lesli-journal.org

in first-person singular pronouns prompted by second-person pronouns, for both the DEFENDANTS and UNTRUSTWORTHY, 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. (Since we were using aggregate 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, for first-person pronouns, than the REPUBLICAN data.)

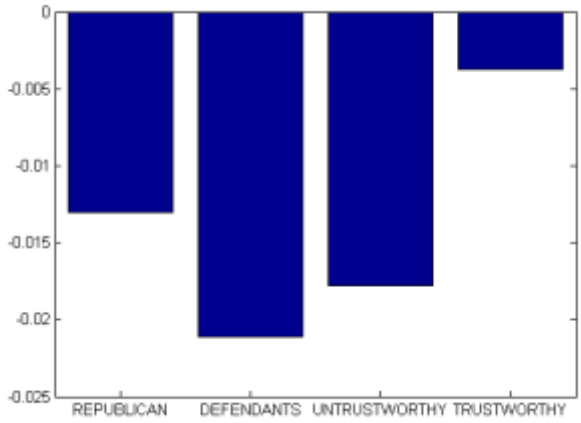
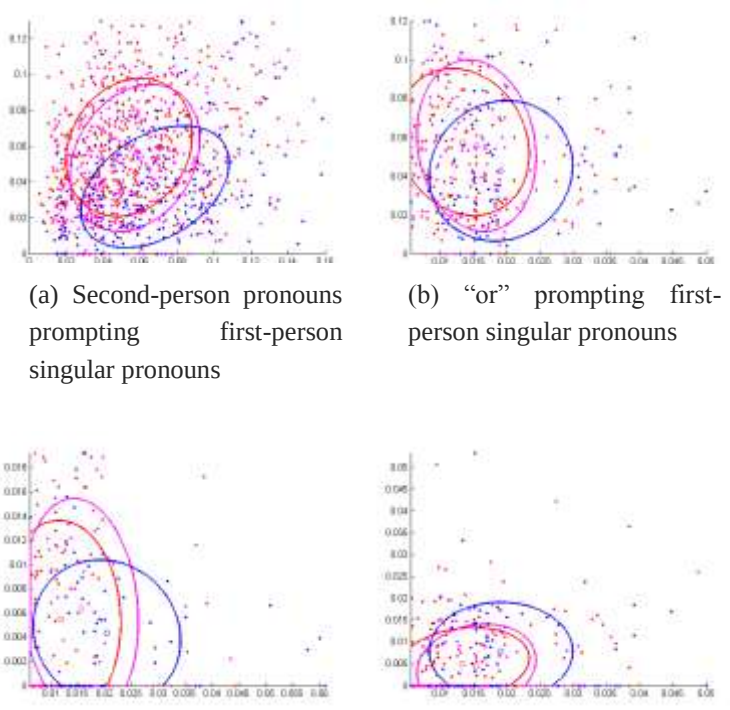


Figure 6: Average change in rates of first-person singular pronouns prompted by second-person pronouns in the REPUBLICAN and NUREMBERG datasets, all using aggregate windows.

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 (they are asked different questions), or showing different rates of response independently of the prompt (they are different kinds of people). Many question word/ response word pairs showed completely different distributions. In general, DEFENDANTS and UNTRUSTWORTHY, the two deceptive groups, were similar, but the TRUSTWORTHY 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.

Volume 1, No. 1 2013 | DOI 10.5195/lesli.2013.1 | http:// lesli-journal.org



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

This suggests a surprising replacement hypothesis:

H2: The response to given question language depends on the mental state of the respondent, in particular whether they are being deceptive or truthful.

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 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. In particular, this suggests that interrogators need to pay attention to the language that they use in questions. For example, using high rates of second-person pronouns in questions makes it easier to distinguish truthful from deceptive responses.

These results also 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. Similar levels of prompting in questions to both groups produces these differentiated responses.

Volume 1, No. 1 2013 | DOI 10.5195/lesli.2013.1 | http:// lesli-journal.org

Prediction Using Question and Answer Language

We used a prediction technique, random forests (Breiman, 2001), to estimate the significance of question words in determining deception. 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 on two datasets, 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 word categories in the questions. For simplicity, we included only the DEFENDANTS and TRUSTWORTHY subgroups.

Because not every statement by a 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.

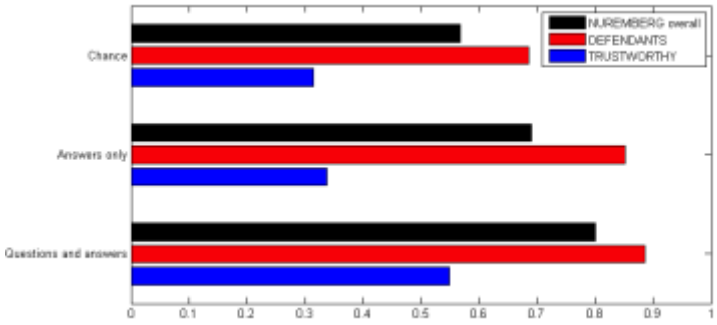


Figure 8: Prediction accuracy of random forests trained on the NUREMBERG data

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 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 without reducing the accuracy on DEFENDANTS.

Table 3 shows the results of attribute ranking with the best-performing random forest. 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 scores. It is likely that the interaction between these two categories drove many of the decision trees in the forest.

Volume 1, No. 1 2013 | DOI 10.5195/lesli.2013.1 | http:// lesli-journal.org

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

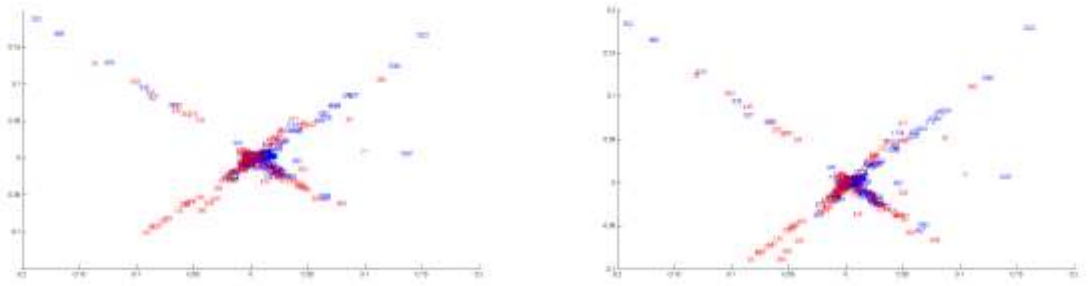
Table 3: Word categories in the random forest trained with question-and-answer data, ranked by the number of splits in the model that use each word.

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.

Validation: the SIMPSON dataset

We now turn to seeing whether these validations generalize. We use the SIMPSON dataset, derived from the transcript of the civil trial, and perform the same analysis.

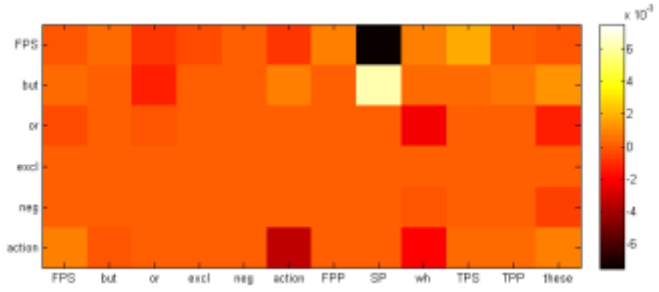
Volume 1, No. 1 2013 | DOI 10.5195/lesli.2013.1 | http:// lesli-journal.org



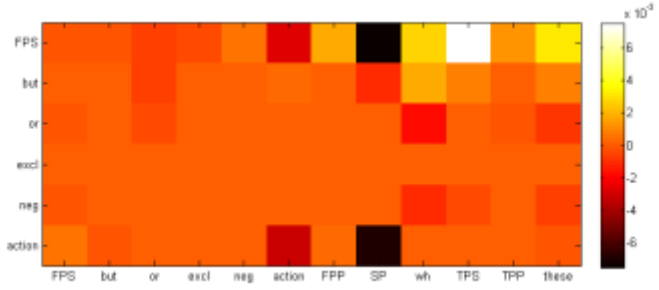
(a) Before correction (b) After correction

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.

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.



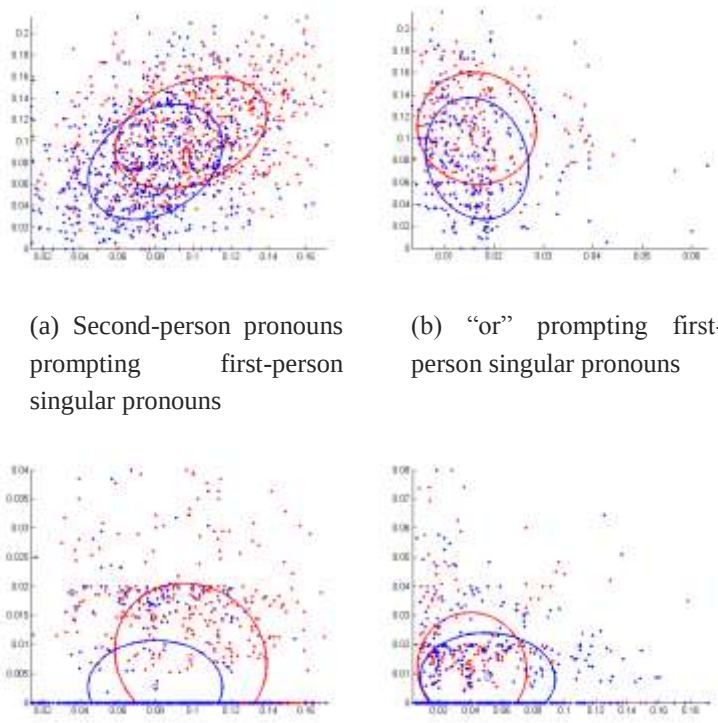
(a) SIMPSON



(b) PLAINTIFFS

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

Volume 1, No. 1 2013 | DOI 10.5195/lesli.2013.1 | http:// lesli-journal.org



(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 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.

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 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 prompted more.

Volume 1, No. 1 2013 | DOI 10.5195/lesli.2013.1 | http:// lesli-journal.org

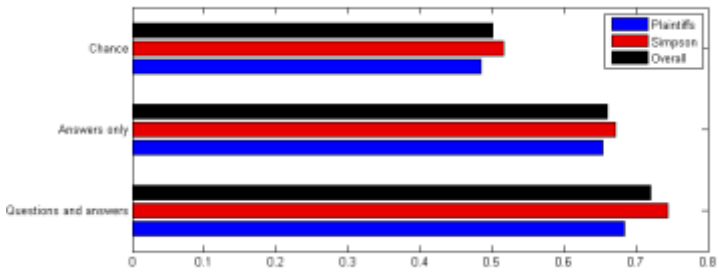


Figure 12: Prediction accuracy of random forests trained on the SIMPSON data

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.

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 of text.

There are almost certainly significant 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 (Hauch *et al.*, 2012) supports the relevance of all the categories of the Pennebaker model, but suggests several other word categories where frequencies might change with deception: increased positive and overall emotion words (as Zhou *et al.* found (2008)), increased negations ("not", "never"), decreased third-person pronouns, and slightly decreased tentative and time-related words.

The use of bag-of-words techniques, that is, models that consider only lexical entities and their frequencies, 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" should be considered to contain an active second-person pronoun or to be a package representing a single, non-pronominal utterance).

The prompting effects in the SIMPSON dataset were somewhat different from those in the NUREMBERG dataset and generally smaller. The random forest model showed a smaller improvement when question data were 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 a civil suit, both plaintiffs and defendants are, in a sense, on trial since the outcome is close to zero-sum, and so both may have motivation to be less than completely truthful, even if only in the sense of presenting themselves as better people than they are. In the SIMPSON dataset, it is plausible that the data underrepresent the difference between deception and truthfulness because of this possibility. 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.

Volume 1, No. 1 2013 | DOI 10.5195/lesli.2013.1 | http:// lesli-journal.org

In the NUREMBERG dataset, the difference between truthfulness and deception may be exaggerated because of other differences between the Nazis and the TRUSTWORTHY group. These groups generally spoke different first languages and were asked different types of questions; the TRUSTWORTHY were often given perfunctory cross-examination. Care must be taken in interpreting these differences in questions. As Hancock *et al.* (2008) showed, 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.* (2012) who tried to restrict their datasets to remove such confounds found that it did not increase the accuracy of their models. The testimony of many of these witnesses was translated in real time; this may have distorted the language patterns, and certainly distorted the timing of questions and responses, which is conceivably relevant.

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 forensic purposes unless it has been tested with respect to all of these issues.

The analysis here has been somewhat broad-brush since we do now have access to ground truth about deception at the level of individual responses; nor have we considered whether response patterns might differ by gender, age, or personality. There is a need for corpora labelled with the truth or deceptiveness of documents (Fitzpatrick & Bachenko, 2012).

Discussion

We began with the observation that, in interrogation, the language of questions naturally affects the language of answers. This is the result both of technical requirement of language, and of verbal mimicry. The Pennebaker model of deception cannot be applied to responses as if they were free-form statements, particularly as many of the words on which it depends are function words, and so strongly affected in dialogues.

Our initial strategy to generalize the Pennebaker model so that it could be applied to interrogations began from hypothesis *H1*, that removing the effects of question words from responses would leave a residue close to the free-form component of each response. We developed an algorithmic technique to estimate how much of the rate of each relevant word was due to prompting words in the question and how much was not.

There are many other settings where language use "flows" from one individual to another. For example, real-world interactions may cause one person to influence the language of another in ways that can then be detected in language use online (an effect which was weakly visible in Enron emails (Keila & Skillicorn, 2005b)). Our technique can be used not only to remove the effect of influence, but also to estimate its strength.

However, analysis of the corrected and uncorrected response data showed that, unexpectedly, removing the effect of question language from responses reduced the differentiation between truthful and deceptive individuals. Further investigation showed that this is because the level of response to a particular prompting word itself depends on the mental state of the respondent – those being deceptive respond differently to those being truthful, even for the same prompts.

This explained an earlier empirical result by Little and Skillicorn, who observed that increases in all four word categories were indicative of deception in responses to interrogation – exactly what would be expected if prompting

Volume 1, No. 1 2013 | DOI 10.5195/lesli.2013.1 | http:// lesli-journal.org

words occurred at the same rates in questions to both the truthful and the deceptive. These results also suggest that interrogators should pay attention to their own word use because of its prompting effect.

If the level of response to a fixed stimulus depends on the mental state of the respondent, then taking into account both the words of questions and the words of responses should make it easier to differentiate truthful from deceptive responses, and indeed this turns out to be the case. Random forest predictors trained on question and response words showed a lift of 10-20% points in accuracy in comparison to predictors trained only on response words.

These results have also clarified exactly which of the pairings of question word category and response word category have significant prompting effects. For example, second-person pronouns in questions have the greatest impact by far on first-person singular pronouns in responses. There could, of course, be other unsuspected but significant word categories in either questions or answers that play a role as markers of deception.

Many of these word categories also play a role in other models of mental state – for example, first-person singular pronouns are important discriminators of power in interactions. The approach taken here can be used, in conversations, to untangle the rates at which such words are generated spontaneously, and the rates that result from prompting by other participants.

References

American Broadcasting Company. 2011 (December 11). *Full Transcript: ABC News Iowa Republican Debate*. abcnews.go.com/Politics/full-transcript-abc-news-iowa-republican-debate/.

Breiman, L. 2001. Random Forests. *Machine Learning*, **45**, 5–32.

Brown, A.S., & Murphy, D.R. 1989. Cryptomnesia: Delineating Inadvertent Plagiarism. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **15**(3), 432–442.

Burgoon, J.K., Hamel, L., & Qin, T. 2012. Predicting Veracity from Linguistic Indicators. *Pages 323–328 of: 2012 European Intelligence and Security Informatics Conference*.

Cable News Network. 2011a (September 12). *Full Transcript of CNN-Tea Party Republican Debate, 20:00-22:00*. transcripts.cnn.com/TRANSCRIPTS/1109/12/se.06.html.

Cable News Network. 2011b (June 13). *Republican Debate*. transcripts.cnn.com/TRANSCRIPTS/1106/13/se.02.html.

Cable News Network. 2012a (Feburary 22). *Full Transcript of CNN Arizona Republican Presidential Debate*. archives.cnn.com/TRANSCRIPTS/1202/22/se.05.html.

Cable News Network. 2012b (January 26). *Full Transcript of CNN Florida Republican Presidential Debate*. archives.cnn.com/TRANSCRIPTS/1201/26/se.05.html.

Carlson, J.R., George, J. F., Burgoon, J.K., Adkins, M., & White, C.H. 2004. Deception in Computer-Mediated Communication. *Group Decision and Negotiation*, **13**, 5–28. 10.1023/B:GRUP.0000011942.31158.d8.

Chartrand, T. L., & van Baaren, R. 2009. Human Mimicry. *Advances in Experimental Social Psychology*, **41**.

Chicago Sun-Times. 2011a (November 13). *CBS/National Journal GOP debate. Transcript, video*. blogs.suntimes.com/sweet/2011/11/_cbsnational_journal_gop_debat.html.

Chicago Sun-Times. 2011b (November 9). *CNBC Republican debate. Transcript, video highlights*. blogs.suntimes.com/sweet/2011/11/cnbc_republican_debate_transcr.html.

Chicago Sun-Times. 2011c. *CNN Republican debate, Nov. 22, 2011. Transcript*. blogs.suntimes.com/sweet/2011/11/cnn_republican_debate_nov_22_2.html.

Chicago Sun-Times. 2011d (October 19). *Republican Las Vegas CNN debate: Transcript*. blogs.suntimes.com/sweet/2011/10/republican_las_vegas_cnn_debat.html.

Chicago Sun-Times. 2012a (January 8). *GOP NH ABC/Yahoo News debate: Transcript*. blogs.suntimes.com/sweet/2012/01/gop_nh_abcyahoo_news_debate_tr.html.

Chicago Sun-Times. 2012b (January 8). *GOP NH NBC's Meet the Press/Facebook debate: Transcript*. blogs.suntimes.com/sweet/2012/01/gop_nh_nbcs_meet_the_pressface.html.

Chicago Sun-Times. 2012c (January 20). *South Carolina GOP CNN debate, Jan. 19, 2012. Transcript*. blogs.suntimes.com/sweet/2012/01/south_carolina_gop_cnn_debate_.html.

Chung, C., & Pennebaker, J. 2007. The Psychological Functions of Function Words. *Pages 343–359 of:* Fiedler, K. (ed), *Social Communication*. New York: Psychology Press.

Council on Foreign Relations. 2012. *Republican Debate Transcript, Tampa, Florida, January 2012*. www.cfr.org/us-election-2012/republican-debate-transcript-tampa-florida-january-2012/p27180.

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

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

Ekman, P. 2002. *Telling Lies: Clues to Deceit in the Marketplace, Marriage, and Politics*. 3rd edn. W.W. Norton.

Ekman, P., & O'Sullivan, M. 1991. Who Can Catch a Liar? *American Psychologist*, **46**(9), 913–920.

Fitzpatrick, E., & Bachenko, J. 2012. Building a Data Collection for Deception Research. *Pages 31–38 of: EACL 2012, Proceedings of the Workshop on Computational Approaches to Deception Detection*.

Fornaciari, T., & Poesio, M. 2012 (April 23). On the Use of Homogenous Sets of Subjects in Deceptive Language Analysis. *Pages 39–47 of: Proceedings of the Workshop on Computational Approaches to Deception Detection*. 13th Conference of the European Chapter of the Association for Computational Linguistics.

Fox News. 2011 (August 12). *Complete Text of the Iowa Republican Debate on Fox News Channel*. foxnewsinsider.com/2011/08/12/full-transcript-complete-text-of-the-iowa-republican-debate-on-fox-news-channel/.

Golub, G.H., & van Loan, C.F. 1996. *Matrix Computations*. 3rd edn. Johns Hopkins University Press.

Gregory Jr., S.W., Dagan, K., & Webster, S. 1997. Evaulating the relation of vocal accomodation in conversation partners' fundamental frequencies to perceptions of communication quality. *Journal of Nonverbal Behavior*, **21**(1).

Groom, C.J., & Pennebaker, J.W. 2005. The Language of Love: Sex, Sexual Orientation, and Language Use in Online Personal Advertisements. *Sex Roles*, **52**(7/8).

Gupta, S., & Skillicorn, D. B. 2006. 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: ACM.

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

Hauch, V., Blandón-Gitlin, I., Masip, J., & Sporer, S.L. 2012 (April 23). Linguistic Cues to Deception Assessed by Computer Programs: A Meta-Analysis. *Pages 1–4 of: Proceedings of the Workshop on Computational Approaches to Deception Detection*. 13th Conference of the European Chapter of the Association for Computational Linguistics.

His Majesty's Stationery Office. 1946. *The Trial of German Major War Criminals Sitting at Nuremberg, Germany*. nizkor.org/hweb/imt/tgmwc/.

History Musings. 2011. *Republican Candidates Debate in Sioux City, Iowa December 15, 2011*. 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/.

Hu, X., & Liu, H. 2012. Text Analytics in Social Media. *Pages 385–414 of:* Aggarwal, C.C., & Zhai, C.X. (eds), *Mining Text Data*. Springer Science+Business Media.

Ireland, M.E., & Pennebaker, J.W. 2010. Language Style Matching in Writing: Synchrony in Essays, Correspondence, and Poetry. *Journal of Personality and Social Psychology*, **99**(3), 549–572.

Keila, P. S., & Skillicorn, D. B. 2005a. Detecting unusual and deceptive communication in email. *Pages 17–20 of: Centers for Advanced Studies Conference*.

Keila, P.S., & Skillicorn, D.B. 2005b. Structure in the Enron Email Dataset. *Computational and Mathematical Organization Theory*, **11**(3), 183–199.

Koppel, M., Akiva, N., Alshech, E., & Bar, K. 2009. Automatically Classifying Documents by Ideological and Organizational Affiliation. *Pages 176–178 of: Proceedings of the IEEE International Conference on Intelligence and Security Informatics (ISI 2009)*.

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

Little, A., & Skillicorn, D.B. 2008 (June 17-20). Detecting deception in testimony. *Pages 13–18 of: IEEE International Conference on Intelligence and Security Informatics*.

Mihalcea, R., & Straparava, C. 2009. The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language. *Pages 309–312 of: ACL-IJCNLP*.

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

Natale, M. 1975. Convergence of Mean Vocal Intensity in Dyadic Communication as a Function of Social Desirability. *Journal of Personality and Social Psychology*, **52**(5), 790–804.

National Archive. 1946-1947. *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. nuremberg.law.harvard.edu/.

New York Times. 2011 (September 7). *The Republican Debate at the Reagan Library*. www.nytimes.com/2011/09/08/us/politics/08republican-debate-text.html.

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

Niederhoffer, K.G., & Pennebaker, J.W. 2002. Linguistic Style Matching in Social Interaction. *Journal of Language and Social Psychology*, **21**(4), 337–360.

Pennebaker, J.W. 2011. Using Computer Analyses to Identify Language Style and Aggressive Intent: The Secret Life of Function Words. *Dynamics of Asymmetric Conflict: Pathways Towards Terrorism and Genocide*, **2**(4), 92–102.

Pennebaker, J.W. 2013. *Linguistic Inquiry and Word Count*. http://www.liwc.net/.

Polikovsky, S., Quiros-Ramirez, M.A., Kameda, Y., Ohta, Y., & Burgoon, J. 2012. Benchmark Driven Framework for Development of Emotion Sensing Support Systems. *Pages 353–355 of: 2012 European Intelligence and Security Informatics Conference (EISIC)*.

PolitiSite. 2011. *Transcript - Fox News-Google GOP Presidential debate September 22, 2011 Orlando, Florida*. www.politisite.com/2011/09/23/transcript-fox-news-google-gop-presidential-debate-september-22-2011-orlando-florida/.

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

RonPaul.com. 2011 (May 5). *Fox News Debate, Greenville SC*. /previous/may-5-2011-greenville-south-carolina/.

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

Simmons, R.A., Gordon, P.C., & Chambless, D.L. 2005. Pronouns in Marital Interaction: What do "You" and "I" Say About Marital Health? *Psychological Science*, **16**(12).

Skillicorn, D.B. 2010. Applying Interestingness Measures to Ansar Forum Texts. *Pages 1–9 of: Proceedings of KDD 2010, Workshop on Intelligence and Security Informatics*.

Skillicorn, D.B. 2012. Lessons from a Jihadi Corpus. *In: Foundations of Open-Source Intelligence FOSINT 2012*.

Skillicorn, D.B., & Leuprecht, C. 2012 (August). The Mental State of Influencers. *Pages 922–929 of: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Workshop on Foundations of Open-Source Intelligence*.

Skillicorn, D.B., & Little, A. 2010. Patterns of word use for deception in testimony. *Pages 25–39 of:* Yang, Christopher C., Chau, Michael, Wang, Jau-Hwang, & Chen, Hsinchun (eds), *Security Informatics*. Annals of Information Systems, vol. 9. Springer US.

Superior Court of the State of California. 1996. *The Simpson Trial Transcripts*. walraven.org/simpson/.

Tausczik, Y.R., & Pennebaker, J.W. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, **29**, 24–54.

Vrij, A., & Mann, S. 2001. Telling and Detecting Lies in a High-Stake Situation: The Case of a Convicted Murderer. *Applied Cognitive Psychology*, **15**, 187–203.

Washington Post, The. 2011 (October 11). *Republican presidential debate (full transcript)*. www.washingtonpost.com/politics/republican-debate-transcript/2011/10/11/gIQATu8vdL_story.html.

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

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

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

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

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