Detecting Deception through Linguistic Analysis

Conference Paper *in* Lecture Notes in Computer Science · June 2003

DOI: 10.1007/3-540-44853-5_7 · Source: DBLP

4 authors, including:

Judee K Burgoon
The University of Arizona
289 PUBLICATIONS **10,603** CITATIONS
SEE PROFILE

J. Pete Blair
Texas State University
24 PUBLICATIONS **370** CITATIONS
SEE PROFILE

Jay F. Nunamaker
The University of Arizona
173 PUBLICATIONS **4,508** CITATIONS
SEE PROFILE

Detecting Deception through Linguistic Analysis

Judee K. Burgoon[1], J. P. Blair[2], Tiantian Qin[1], Jay F. Nunamaker, Jr[1],

[1] Center for the Management of Information, University of Arizona

[2] Department of Criminal Justice, Michigan State University

Submitted to the Intelligence and Security Informatics Symposium, Tucson, June 2003.

Portions of this research were supported by funding from the U. S. Air Force Office of Scientific Research under the U. S. Department of Defense University Research Initiative (Grant #F49620-01-1-0394). The views, opinions, and/or findings in this report are those of the authors and should not be construed as an official Department of Defense position, policy, or decision.

Abstract. Tools to detect deceit from language use pose a promising avenue for increasing the ability to distinguish truthful transmissions, transcripts, intercepted messages, informant reports and the like from deceptive ones. This investigation presents preliminary tests of 16 linguistic features that can be automated to return assessments of the likely truthful or deceptiveness of a piece of text. Results from a mock theft experiment demonstrate that deceivers do utilize language differently than truth tellers and that combinations of cues can improve the ability to predict which texts may contain deception.

1 Introduction

One of the most daunting challenges facing intelligence and law enforcement communities in the wake of 9/11 is anticipating and preventing terrorist attacks. Doing so requires sifting through a daily avalanche of information and communications to discriminate good information from bad, and deceptive communications from truthful ones. Much of this task falls to human analysts. Aside from the enormity of the task, the empirical evidence is compelling that humans, even professionally trained ones, are typically very poor at detecting deception and fallacious information [1, 2, 3]. Moreover, evidence is building that detection abilities are even more faulty when communication is computer-mediated, especially if is text-based such as with e-mail [4, 5]. Increasing reliance worldwide on electronic communications thus carries with it the very real likelihood of further decrements in detection accuracy and heightened vulnerabilities to national and international security. Human information processors are in serious need of tools that can assist them in filtering information and alerting them to suspicious information. Aside from the security implications, such tools would also be of great benefit to law enforcement in dealing with a wide array of criminal investigations.

Our program of research, funded by the Department of Defense, has begun to develop such tools by examining linguistic features and content characteristics of texts for reliable indicators of deceit. These indicators can then be incorporated into software tools that automate their detection and subject them to statistical analysis for probabilistic estimates of truthfulness or deceptiveness. Decades of research have confirmed that there are few indicators of deceit that remain invariant across genres of communication, situations, communicators, cultures, and other features of communication contexts. Yet combinations of indicators have shown to have good predictive ability in specified contexts [6]. The research to be reported here was guided by the objectives of identifying those indicators that are (1) the least context-sensitive and (2) the most amenable to automation. We present preliminary results from a mock theft experiment that is still in progress, our purposes being to illustrate the promise of examining text-based linguistic indicators

and of examining such indicators using a particular statistical approach that examines combinations of cues.

2 Background

Two experiments from our research program predate the one to be reported here. One was modeled after the familiar Desert Survival Problem, in which pairs of participants were given a scenario in which their jeep had crashed in the Kuwaiti desert, read material we developed from an Army field manual that we entitled, "Imperative Information for Surviving in the Desert," and then were asked to arrive at a consensus on the rank-ordering of salvageable items in terms of their importance to survival. The task was conducted via email over the course of several days. In half of the pairs, one person was asked to deceive the partner by advocating choices opposite of what the experts recommend (e.g., discarding bulky clothing and protective materials so as to make walking more manageable). Partners discussed the rankings and their recommendations either face-to-face or using a computer-mediated form of communication such as text chat, audioconferencing, or videoconferencing. All discussions were recorded and transcribed then subjected to linguistic analysis of such features as number of words, number of sentences, number of unique words (lexical diversity), emotiveness, and pronoun usage. Of the 27 indicators that were examined, several proved to reliably distinguish truth tellers from deceivers. Deceivers were more likely to use longer messages but with less diversity and complexity, and greater uncertainty and "distancing" in language use than truth tellers. These results revealed that systematic differences in language use could help predict which messages originated from deceivers and which, from those telling the truth.

The second experiment was designed as a pilot effort for the experiment to be reported below. In this experiment, participants staged a mock theft and were subsequently interviewed by untrained and trained interviewers via text chat or face-to-face (FtF) interaction [4, 5]. The FtF interactions were later transcribed, and the transcripts and chats were submitted to linguistic analysis on the same features as noted above, plus several others that are available in the Grammatik tool within WordPerfect. Due to the small sample size, none of the differences between innocents (truth tellers) and thieves (deceivers) were statistically significant, but patterns were suggestive of deceivers tending toward briefer messages (fewer syllables, words, and sentences; shorter and simpler sentences) of greater complexity (e.g., greater vocabulary and sentence complexity, lower readability scores) than truth tellers (higher Flesch-Kincaid, sentence complexity, vocabulary complexity, syllables per word).

The patterns found in these first efforts suggested that we should expect to find many linguistic differences between deceivers and truth tellers with a larger, and well-designed experiment. We therefore hypothesized that *deceptive senders display higher (a) quantity, (b) nonimmediacy, (c) expressiveness, (d) informality, and (e) affect; and less (f) complexity, (g) diversity, and (h) specificity of language in their messages than truthful senders.*

3 Method

Students were recruited from a multi-sectioned communication class by offering them credit for participation and the chance to win money if they were successful at their task. Half of the students were randomly assigned to be "thieves," i.e., those who would be deceiving about a theft, and the other half became "innocents," i.e., those who would be telling the truth.

Interviewees in the deceptive condition were assigned to "steal" a wallet that was left in a classroom. In the truthful condition, interviewees were told that a "theft" would occur in class on an assigned day. All of the interviewees and interviewers then appeared for interviews according to a pre-assigned schedule. We attempted to motivate serious engagement in the task by offering interviewers $10 if they could successfully detect whether their interviewee was innocent or guilty and successfully detect whether they were deceiving or telling the truth on a series of the interview questions. In turn, we offered interviewees $10 if they convinced a trained interviewer that they were innocent and that their answers to several questions were truthful. An additional incentive was a $50 prize to be awarded to the most successful interviewee.

Interviewees were then interviewed by one of three trained interviewers under one of three modalities— Face to Face (FtF), text chat, or audioconferencing. The interviews followed a standardized

Behavioral Analysis Interview format that is taught to criminal investigators [7]. Interviews were subsequently transcribed and submitted to linguistic analysis. Clusters of potential indicators, all of which could be automatically calculated with a shallow parser (Grok or Iskim) or could use a look-up dictionary, were included. The specific classes of cues and respective indicators were as follows:

1. *Quantity* (number of syllables, number of words, number of sentences)
2. *Vocabulary Complexity* (number of big words, number of syllables per word)
3. *Grammatical Complexity* (number of short sentences, number of long sentences, Flesh-Kincaid grade level, average number of words per sentence, sentence complexity, number of conjunctions)
4. *Specificity and Expressiveness* (emotiveness index, rate of adjectives and adverbs, number of affective terms)

4 Results

In the following two subsections, we investigate data from two perspectives: analysis of individual cues and cluster analysis. To analyze how well individual cues distinguish messages of deceivers from those of truth tellers, we conducted multivariate analyses of related groups of cues followed by directional *t*-tests on individual cues to identify which ones contribute most to differentiating deceivers and truthful tellers. The cluster analysis answers the question of whether combinations of cues (in a hierarchy structure) can improve overall ability to differentiate deceivers from truth tellers. Furthermore, unlike the traditional statistical cluster analysis, we used a data-mining algorithm -- C4.5 [8] -- to cluster the cues and obtain a hierarchical tree structure. In this way, we fulfilled the "automatic" requirement of automating deception detection.

4.1 Individual cue analysis

Results were based on data of 41 subjects whose modality was text chat (txt) or audio. (In the future, data for face-to-face (FtF) will be included after those sessions have been completed and all video files are transcribed). Among 41 subjects, 29 interacted via txt and 20, via audio; 26 were "thieves" (i.e., deceivers) and 23 were "innocents (i.e., truth tellers). Table 1 presents descriptive statistics for the 16 cues that were analyzed.

Table 1. Means (Deviations) for 16 cues

Cues	Deceiver		Truthful-teller	
	Txt	Audio	Txt	Audio
Syllables	121.06(82.723)	146.44(75.583)	144.58(86.732)	189.27(86.383)
Words	86.06(58.297)	106.00(53.708)	106.25(64.435)	140.36(62.094)
Sentences	5.88(3.621)	7.11(4.372)	6.00(5.117)	6.91(2.948)
Short sentences	2.82(2.298)	3.11(3.140)	2.58(3.872)	2.09(1.57)
Long sentences	0.24(0.437)	0.44(0.520)	0.92(1.16)	1.09(1.22)
Simple sentences	1.29(1.21)	2.11(2.205)	2.20(2.64)	1.36(0.920)
Big words	6.29(6.64)	6.78(4.32)	6.55(3.77)	8.28(5.71)
Average syllables per word	1.42(0.107)	1.36(0.089	1.36(0.068)	1.34(.096)
Average words per sentence	15.45(6.53)	16.93(9.82)	21.02(7.76)	22.94(13.36)
Flesch-Kincaid grade level	7.28(2.839)	7.32(3.986)	8.91(3.223)	8.82(3.087)
Sentence complexity	41.12(16.86)	43.33(18.66)	52.91(26.44)	49.73(24.08)
Vocabulary complexity	9.88(7.54)	7.11(4.48)	7.27(4.60)	7.27(4.22)
Conjunctions	4.47(3.48)	5.30(4.03)	5.00(3.82)	9.27(7.08)
Rate of Adjectives & Adverbs	0.12(0.014)	0.12(0.016)	0.12(0.014)	0.13(0.014)
Emotiveness	0.29(0.100)	0.29(0.141)	0.29(.073)	0.27(.160)
Affect	0.24(0.562)	0(0)	0.55(0.820)	0.33(0.651)

Results of the multivariate tests and *t*-tests are shown in Table 2. (For plots of means by deception condition and modality, see the figures in the appendix.) The multivariate analysis on indicators of quantity

of language produced a significant multivariate effect for deception (p = .033) and no modality by deception interaction. Deceivers said or wrote less than truth tellers.

The multivariate analysis of indicators of complexity at both the sentence level (simple sentences, long sentences, short sentences, sentence complexity, Flesch-Kincaid grade level, number of conjunctions, average-words-per-sentence (AWS)*)* and vocabulary level (vocabulary complexity, number of big words, average-syllables-per-word (ASW)) did not produce overall multivariate effects, but several individual variables did show the effects of deception condition. Deceivers had significantly fewer *long sentences, AWS* and *sentence complexity* than truth tellers; and a lower *Flesch-Kincaid grade level* than truth tellers. This meant their language was less complex and easier to comprehend. The *t*-tests also provided weak support for deceivers having fewer *ASW* (p = .102) and *conjunctions* (p = .149) in messages than truth tellers. Thus, deceivers used less complex language at both the lexical (vocabulary) and grammatical (sentence and phrase) levels. Modality effects also showed that subjects in text chat used fewer *conjunctions* than in audio chat, indicating that that modality was less likely to exhibit compound and complex sentences.

For the analyses of message specificity and expressiveness (*adjectives and adverbs, emotiveness*, and *affect*), the multivariate test showed a trend toward a main effect for the deception condition (p = .101). There was a significant univariate difference on *affect,* such that deceivers used less language referring to emotions and feelings than did truth tellers.

Table 2. Univariate F-tests (p-values) and t-tests for individual cue analysis

	Test of between-subject effects			
Cues	Modality	Condition	Modality*Condition	Independent Samples t-Test
Syllables	2.054(.159)	1.842(.182)	.156(.695)	1.502(.140)
Words	2.363(.131)	2.407(.128)	.162(.689)	1.702(.096)*
Sentences	.810(.373)	.001(.972)	.018(.894)	.111(.912)
Short sentences	.122(.016)*	.588(.447)	.225(.637)	-.725(.472)
Long sentences	.547(.464)	6.566(.014)*	.005(.947)	2.781(.008)*
Simple sentences	.029(.886)	.002(.969)	1.874(.178)	.061(.951)
Big words	.462(.500)	.288(.594)	.146(.704)	.616(.541)
Average syllables per word	1.949(.17)	1.703(.199)	.413(.524)	-1.668(.102)
Average words per sentence	.374(.544)	4.368(.042)*	.006(.936)	2.414(.021)*
Flesch-Kincaid grade level	.001(.979)	2.690(.108)	.005(.943)	1.958(.056)*
Sentence complexity	.006(.940)	2.055(.159)	.181(.673)	1.779(.082)*
Vocabulary complexity	.657(.422)	.512(.478)	.657(.422)	-.997(.324)
# of Conjunctions	3.393(.072)*	2.569(.116)	1.496(.228)	1.426(.163)
Rate Adjectives and Adverbs	0.150(.700)	.329(.569)	.301(.586)	-.596(.554)
Emotiveness	0.020(.889)	.054(.818)	.060(.808)	-.233(.817)
Affect	1.591(.214)	3.291(.214)	.004(.948)	1.630(.110)

* p < .05, one-tailed.

4.2 Cluster analysis by C4.5

Although many linguistic cues were not significant as shown in section 1, they can form a hierarchy tree that performs relatively well in discriminating deceptive communicators from truthful ones. Among many data-mining algorithms, we chose C4.5 because it provides a clear cluster structure (compared with neural network), as well as satisfactory precision [9]. C4.5 used a pruned tree to cluster the cues. This algorithm cuts off redundant branches while constraining error rates. We used software of *Weka* (University of Waikato in New Zealand; Witten and Frank, 2000) to implement the C4.5. Figure 1 is the output of a pruned tree. "1" stands for truthful condition, "2" stands for deception condition.

The correct prediction rate using 15-fold cross-validation is 60.72%, which is reasonably satisfactory given the small size of the data set.

As shown in the Figure 1, a combination of linguistic cues can well categorize deception behaviors. For example, *sentence level complexity* combined with *vocabulary* or *affect* acted as good classifier. Those significant linguistic cues in section 1 also played important roles in the cluster classification: *number of conjunctions*, *FK grade level*, *AWS*, *affect*. On the other hand, the cluster structure also showed consistency with multivariate tests: not all linguistic cues contribute in identifying

deceptions. There were "unhelpful" cues, such as *emotiveness*, which showed no significance in both the single level structure and cluster analysis (hierarchy structure). However, it is premature to conclude the ineffectiveness of any linguistic cues at this point. Further investigations with larger data sets will give us deeper insight into the intra-relations of cues.

The confusion matrix shows the number of misclassifications: 10 out of 37 true conditions are misclassified as deceptive, and 19 out of 35 deceptive conditions are misclassified. The tree mentioned above produced less misclassifications in the truth condition than in the deceptive condition, which implies a "truth-biased" judgment. In other words, the cluster method is cautious in designating messages as untrustworthy.

```
#conjunction <= 10
|   #conjunction <= 3
|   |   #bigwds <= 7
|   |   |   FK_grade <= 8.383333
|   |   |   |   #bigwds <= 3
|   |   |   |   |   Voc_comp <= 12
|   |   |   |   |   |   number_of_sentences <= 2
|   |   |   |   |   |   |   Average_words_per_sentence <= 7.625: 1 (2.0)
|   |   |   |   |   |   |   Average_words_per_sentence > 7.625: 2 (3.0)
|   |   |   |   |   |   number_of_sentences > 2: 2 (7.0)
|   |   |   |   |   Voc_comp > 12: 1 (2.0)
|   |   |   |   #bigwds > 3: 2 (9.0)
|   |   |   FK_grade > 8.383333: 1 (7.0/1.0)
|   |   #bigwds > 7: 1 (5.0/1.0)
|   #conjunction > 3
|   |   Average_words_per_sentence <= 31.25
|   |   |   Affect <= 1: 1 (23.0/4.0)
|   |   |   Affect > 1
|   |   |   |   Average_words_per_sentence <= 14.166667: 2 (3.0)
|   |   |   |   Average_words_per_sentence > 14.166667: 1 (4.0)
|   |   Average_words_per_sentence > 31.25: 2 (3.0)
#conjunction > 10: 2 (4.0)

=== Confusion Matrix ===
  a  b   <-- classified as
 27 10 |  a = truthful
 19 16 |  b = deceptive
```

Fig. 1. Pruned tree output from C4.5

5 Discussion

This investigation was undertaken largely to demonstrate the efficacy of utilizing linguistic cues, especially ones that can be automated, to flag potentially deceptive discourse, and to use statistical clustering techniques to select the best *set* of cues to reliably distinguish truthful from deceptive communication. This investigation demonstrates the potential of both the general focus on language indicators and the use of hierarchical clustering techniques to improve the ability to predict what texts might be deceptive.

As for the specific indicators that might prove promising, these results provide some evidence for the hypothesis that deceivers behave differently than truth tellers in communications via text chat and/or audio chat. Although many tests were not significant due to the small sample size, there was a trend shown in the profile plots demonstrating that: deceivers' messages were briefer (i.e., lower on quantity of language), were less complex in their choice of vocabulary and sentence structure, and lack specificity or expressiveness in their text-based chats. This is consistent with profiles found in nonverbal deception research showing deceivers tend to adopt, at least initially, a fairly inexpressive, rigid communication style with "flat" affect. It appears that their linguistic behavior follows suit and also demonstrates their inability to create messages rich with the details and complexities that characterize truthful discourse. Over time, deceivers may alter these patterns, more closely approximating normal speech in many respects. But it is

possible that language choice and complexity may fail to show changes because deceivers are not accessing real memories and real details, and thus will not have the same resources in memory upon which to draw.

Unlike asynchronous experiments such as the Desert Survival experiment (DSP), subjects did not have sufficient time to provide detailed lies that contained more quantity and complexities [10]. The differences in synchronicity in these two tasks points to time for planning, rehearsal, and editing as a major factor that may alter the linguistic patterns of deceivers and truth tellers. As a consequence, no single profile of deception language across tasks is likely to emerge. Rather, it is likely that different cue models will be required for different tasks. Consistent with *interpersonal deception theory* [11], deceivers may adapt their language style deliberately according to the task at hand and their interpersonal goals. If the situation does not afford adequate time for more elaborate deceits, one should expect deceivers to say less. But if time permits elaboration, and/or the situation is one in which persuasive efforts may prove beneficial, deceivers may actually produce longer messages. What may not change, however, is their ability to draw upon more complex representations of reality because they are not accessing reality. In this respect, complexity measures may prove less variant across tasks and other contextual features. The issue of context invariance thus becomes an extremely important one to investigate as this line of work proceeds.

Modality also plays a role in communication. Subjects talked more than they wrote, but message complexity did not seem to be much different between the text and audio modalities. Future research will explore the effect of different communication modalities on the characteristics of truthful and deceptive messages.

Although clustering analysis did not consider modality effects, it provided a hierarchy tree structure to capture the combined characteristics of the cues. It also provided an exploratory threshold value to separate deceptive and true messages. It should also be noted that the analysis in this study used the absolute values of linguistic characteristics to classify statements as truthful or deceptive. Because people vary greatly in their usage of the language (e.g. some people naturally use more or less complex language than others), the use of cue values that are relative to the sender of the message may result in greater classification accuracy. This would require building a model of the baseline speech patterns of an individual and then comparing an individual message to this model.

Future research will consider intra-connections among linguistic cues, tasks, and modalities. More data will also enhance reliability of current results, but it is clear from these results alone that linguistic cues that are amenable to automation may prove valuable in the arsenal of tools to detect deceit.

References

1. Burgoon, J. K., Buller, D. B., Ebesu, A., Rockwell, P.: Interpersonal Deception: V: Accuracy in Deception Detection. Communication Monographs **61** (1994) 303-325

2. Levine, T., McCornack, S.: Linking Love and Lies: A Formal Test of the McCornack and Parks Model of Deception Detection. J. of Social and Personal Relationships **9** (1992) 143-154

3. Zuckerman, M., DePaulo, B., Rosenthal, R.: Verbal and Nonverbal Communication of Deception. In: Berkowitz, L. (ed.): Advances in Experimental Social Psychology, Vol. 14. Academic Press, New York (1981) 1-59

4. Burgoon, J., Blair, J. P., Moyer, E.: Effects of Communication Modality on Arousal, Cognitive Complexity, Behavioral Control and Deception Detection during Deceptive Episodes. Paper submitted to the Annual Meeting of the National Communication Association, Miami. (2003, November)

5. Burgoon, J., Marett, K., Blair, J. P.: Detecting Deception in Computer-Mediated Communication. In: George, J. F. (ed.): Computers in Society: Privacy, Ethics & the Internet. Prentice-Hall, Upper Saddle River, NJ (in press)

6. Vrij, A.: Detecting Lies and Deceit. John Wiley and Sons, New York (2000)

7. Inbau, F. E., Reid, J. E., Buckley, J. P., Jayne, B. C.: Criminal Interrogations and Confessions. 4th edn. Aspen, Gaithersburg, MD (2001)

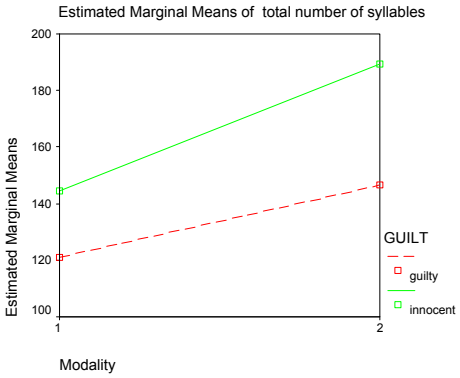
8. Quinlan, J. R.: C4.5. Morgan Kaufmann Publishers, San Mateo, CA (1993)

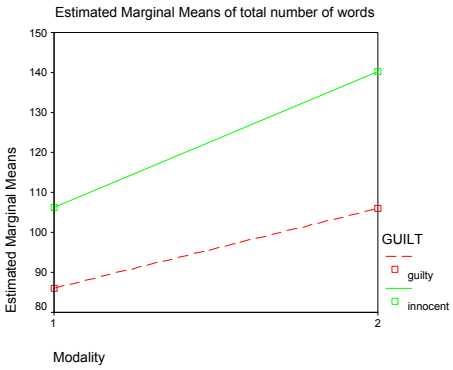
9. Spangler, W., May, J., Vargas, L.: Choosing Data-Mining Methods for Multiple Classification: Representational and Performance Measurement Implications for Decision Support. J. Management Information Systems **16** (1999) 37-62

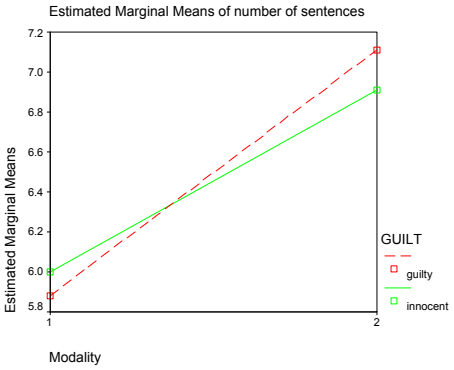
10. Zhou, L. Twitchell, D., Qin, T., Burgoon, J. K., Nunamaker, J. F., Jr.: An Exploratory Study into Deception Detection in Text-based Computer-Mediated Communication. In: Proceedings of the 36th Annual Hawaii International Conference of System Sciences. Big Island, Los Alamitos, CA (2003)

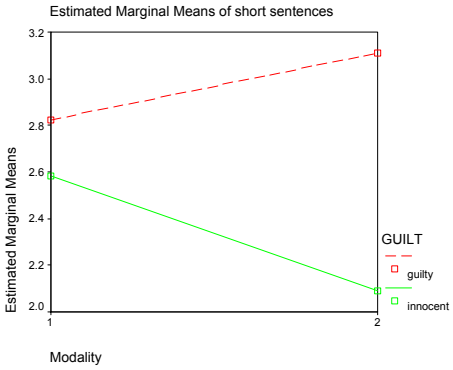
11. Buller, D. B., Burgoon, J. K.: Interpersonal Deception Theory. Communication Theory **6** (1996) 203-242

Appendix: Individual cue comparisons by modality and deception condition*

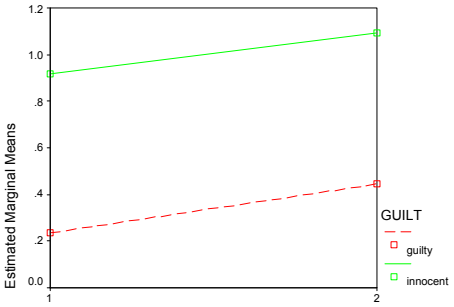




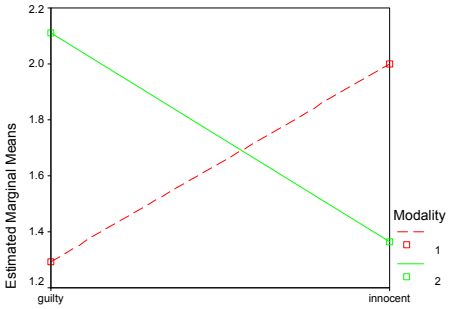




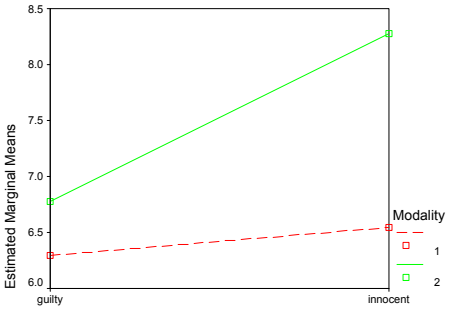
Estimated Marginal Means of long sentences



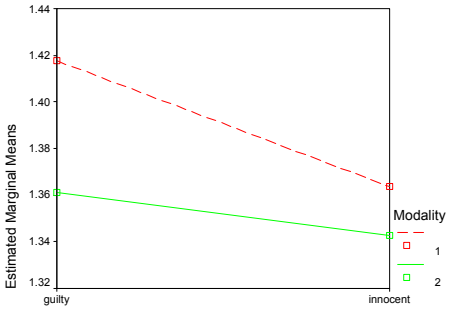
Estimated Marginal Means of simple sentences



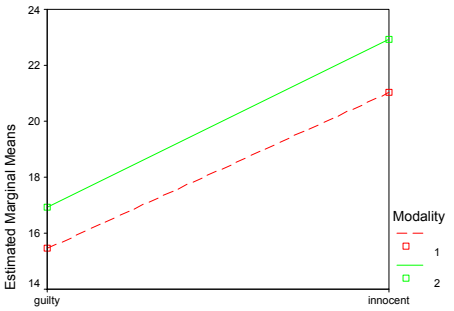
Estimated Marginal Means of big words



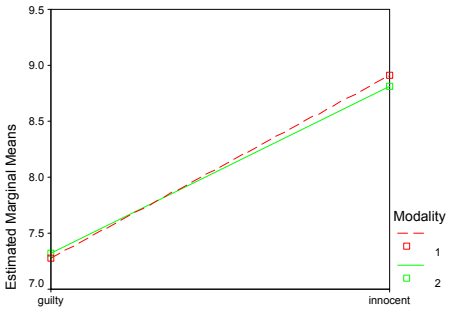
Estimated Marginal Means of average syllables per word



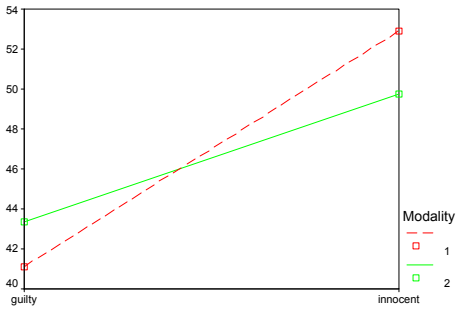
Estimated Marginal Means of average words per sentence



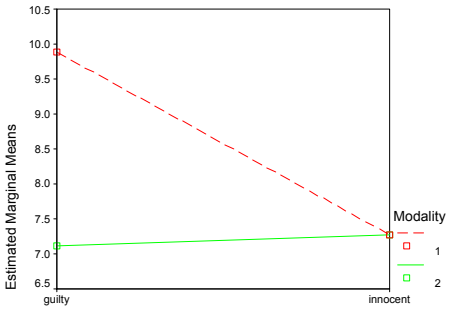
Estimated Marginal Means of Flesch-Kincaid grade level

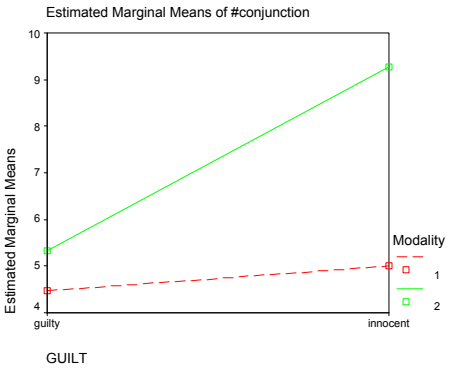


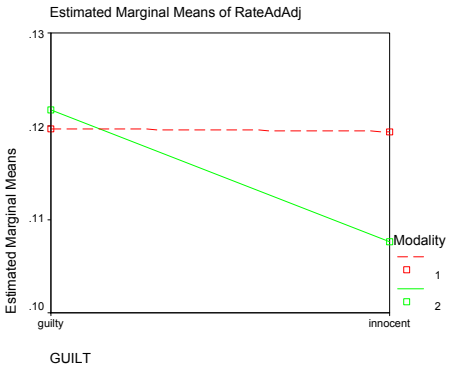
Estimated Marginal Means of sentence complexity

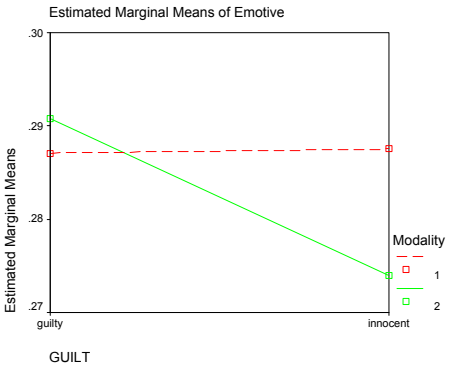


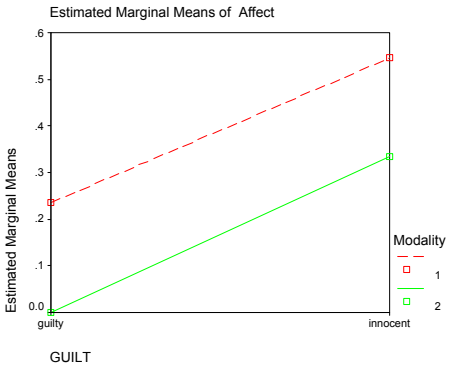
Estimated Marginal Means of vocabulary complexity











* Modality 1 = Text, Modality 2 = Audio