A Corpus-based Multidimensional Analysis of Linguistic Features of Truth and Deception

Zheng Fanghua
School of Foreign Languages, Peking University
zhengfh2016@pku.edu.cn

Abstract

This study sets out to examine the linguistic difference between truthful and deceptive texts. In order to take more linguistic features into consideration, this research applied multidimensional analysis, which can reduce many linguistic features into several factors. This study used a self-built corpus containing 100 truthful texts and 100 deceptive texts. TextMind was employed to annotate these Chinese texts automatically. SPSS version 20 was utilized for t-tests and multidimensional analysis. The discussion of the data was divided into two parts: word count and word per sentence, and multidimensional analysis. This research reveals that word count and word per sentence of deceptive discourse are significantly smaller than those of truthful discourse. The results of multidimensional analysis suggest that deceptive discourse displays a weaker performance on dimensions of narration, interpersonal relationship, and perception.

Key words: Truth, Deception, Corpus, Multidimensional Analysis

1 Introduction

Lying is considered as an important and frequent part of everyday social interactions of our age and some studies suggest that on average people tell one or two lies everyday (DePaulo, Kashy, Kirkendol, Wyer, & Epstein, 1996; Hancock, Thom-Santelli, & Ritchie, 2004). In Ekman's (1992) definition, the word "deception" and "lie" can be used interchangeably, which means that a person intentionally or deliberately misleads another person without mentioning his purpose in advance. Ekman (1992) further explains that there are two ways to lie: to conceal and to falsify. To conceal means that the liar withholds some information that misleads the listener. To falsify means to fabricate information, to distort facts or to convey an opposite attitude.

Since deception has become prevalent in everyday life, especially with the development of technology, how to detect deception becomes a key issue. People are poor at detecting lies (Vrij, Edward, Roberts & Bull, 2000). Vrij et al. mentions that when detecting lies through non-verbal cues, such as face expression, pitch of voice or speech rate, the accuracy rates usually ranges from 40% to 60%, sometimes even lower than a chance of guess (50%). In the past two decades, a growing number of studies have been done to find linguistic cues to improve the deception detection accuracy (DePaulo, Lindsay, Malone, Muhlenbruck, Charlton, & Cooper, 2003; Newman, Pennybaker, Berry & Richards, 2003; Bond & Lee, 2005). Many linguistic cues have been analyzed, including pronouns, negative emotions, conjunctions, prepositions and so on. The number of these linguistic features is often calculated through Linguistic Inquiry and Word Count (LIWC), which is a computerized text analysis program (Pennebaker, Francis, & Booth, 2012). Inspired by LIWC, Computational Cyber-Psychology Lab developed TextMind, which can be used to analyze Chinese texts.

These studies, however, only select a limit number of linguistic features to analyze. In order to obtain a more comprehensive view, this study aims to include more linguistic features and take the multidimensional approach to explore the differ-

Copyright 2018 by the author

ence between truthful and deceptive discourse based on two self-built comparable corpora.

Since Biber proposed multidimensional analysis in 1984, it has been extensively used to study the linguistic features of different registers. Multidimensional analysis has the advantage of examining hundreds of linguistic features at a time and classifying them into several dimensions according to the frequency of co-occurrence among these features. The self-built corpus in this study contains 100 truthful texts and 100 deceptive texts in Chinese. These texts will be firstly analyzed through TextMind to get quantitative data. Then tens of linguistic features will be selected for multidimensional analysis. The factor scores of truthful texts and deceptive texts will be calculated to see whether there are significant differences between truth and deception on these dimensions.

2 Literature Review

This part firstly reviews the methods for data collection in relevant studies. Advantages and disadvantages of these methods are also mentioned. Then, it summarizes the major findings of previous research on verbal cues of deception. Finally, it concludes two major limitations of these studies, which can be improved in the current study.

2.1 Methods for Data Collection

In the past two decades, studies on verbal cues in deception have employed a number of methods for collecting truthful and deceptive statements. According to the different nature of the deceptive statements, data can be classified into two types: data collected from natural condition and data collected from controlled experiments. Natural data of deception are often collected from high stakes lies including criminal statements, police interrogations and legal testimony (Fitzpatrick & Bachenko, 2012). Such kind of data is often analyzed to identify theft, fraud or scams for the purpose of solving criminal case.

However, the most common way to collect lies employed by the researchers is to collect them through controlled experiments. A classic example is Newman et al.'s experiment in 2003, in which they instruct participants to give true and deceptive opinions on abortion and feelings about friends. Another way to collect lies is to ask participants to conduct a mock crime. Newman et al. (2003) recruit 60 students in a money-stealing mock crime. Half of the students are asked to steal the money in the book and half is not. When an experimenter questions them, those who steal the money need to deny taking the money by telling lies, while those who do not take the money deny taking the money by telling truth.

Some researchers point out a crucial disadvantage of collecting data through experiments. These lies cannot replicate the true situation of lying (Fitzpatrick & Bachenko, 2012). In other words, participants know that they will not be punished for lying in such studies. Hence the motivation to lie is weak. Therefore, the language they use in such lies might not be the same as lies in real life situation. However, the advantage of this method is also very clear, that is, researchers have more control over the number and the kind of data they want to collect. They can collect true and deceptive statements at the same time through different modes, such as emails, recordings, face-to-face communication, or handwritten texts. Thus, this study employs controlled experiment to collect data.

2.2 Findings of Relevant Studies

In the past two decades, a growing number of studies have been done to find linguistic cues to facilitate the detection of deception. Many studies find that liars tend to use fewer self-reference words and refer to others frequently (Bond & Lee, 2005; Newman et al., 2003). Also, in deceptive descriptions, there are more negative emotions and motion words (Newman et al., 2003; DePaulo et al., 2003).

However, the results of some research are inconsistent with the findings mentioned above (Hauch, Blando ́n-Gitlin, Masip & Sporer, 2012). In some corpus, the number of first person pronouns in deceptive texts and non-deceptive texts is not significantly different. Also, in some language like Italian, the first personal pronoun is often omitted. Besides, the percentage of negative emotions may have something to do with the topic of the lies. Therefore, it is worth mentioning that linguistic cues, such as the number of first number pronouns and negative emotions may be sensitive to the topic and context of the lies.

Different from the linguistic cues mentioned above, Burgoon et al. (2016) recently studies deceptive utterances from the following five aspects:

utterance length and specificity, complexity, hedging and uncertainty; comprehensibility; affect. This research finds that deceptive utterances are laden with more information and details and have more uncertainty language. One of the most important implications of Burgoon et al.'s research is that they find deceptive utterances "differed systemically" from non-deceptive ones. In prepared lies, the language is more complex and is less comprehensible.

A review of recent studies on verbal cues of deception demonstrates that it may be feasible to detect lies through linguistics cues, such as pronouns, emotions, word count, specificity, hedging, affect and so on. However, there are still some limitations in previous studies. First, the number of linguistic features included in deception research is very limited, which lacks a comprehensive view. Second, compared with the increasing number of deception studies on western languages, few studies have explored the linguistic cues of deception in Chinese. The statistical performance of linguistic features may vary when lies are in different languages.

3 Experiment Design and Data Collection

The current study included 100 participants who were asked to give both truthful and deceptive speeches in Chinese concerning the topic "my favorite teacher". In the first speech, the participants were instructed to describe their favorite teachers according to their true attitudes. In the second speech, they were instructed to describe a teacher they actually dislike in a way as if they like this teacher. Participants were asked to describe each teacher for one minute (no more than one and a half minutes), during which their speeches were tape-recorded. In the truthful speeches, participants were encouraged to provide true evidence or to tell honestly how they feel. In the deceptive speeches, participants were encouraged to convey a convincing false impression. They were allowed to invent stories or to use false evidence to help them convey deceptive information.

Altogether, 200 recordings were collected, with 100 truthful recordings and 100 deceptive recordings. Each verbal sample was transcribed, and truthful and deceptive texts entered into different text files. Paralinguistic cues (e.g., smile, tone, intonation) were removed, but the pauses were remained and annotated manually for the purpose of basic statistic analysis. A corpus of the transcribed texts was set up, with a total word count of 30657 Chinese characters, with truthful discourse counting 16447 characters and deceptive discourse 14210 characters.

3.1 Automatic Annotation by TextMind

Inspired by LIWC 2007, Computational Cyber-Psychology Lab developed TextMind, which is a Chinese language psychological analysis system. TextMind utilizes Language Technology Platform (LTP) for parsing and annotation of simplified Chinese. Each of the 200 transcribed texts was analyzed individually by TextMind.

Although TextMind can analyze texts along more than 100 linguistic features, this study only selects 38 linguistic features for multidimensional analysis. The rest of linguistic features were excluded from analysis for two reasons. First, the linguistic features that were used at extremely low rates were excluded. Second, linguistic features that were less likely to be related to the topic or were unlikely to reflect the difference between truth and lies, such as money, death, ingest and so on, were excluded. Third, the linguistic features for which Chinese does not have overt markings were excluded, such as tense and articles. Usually, such features also have low frequency, thus they would be excluded from the analysis anyway.

3.2 Multidimensional Analysis by SPSS 20

The results of TextMind provided us with the raw frequency of linguistic features. Before multidimensional analysis, we standardized these data by calculating their frequency per hundred words. Then, we applied SPSS version 20 for the multidimensional analysis. We reduced the dimensions of these 38 linguistic features by applying factor analysis in SPSS. The results displayed the major factors (dimensions), the linguistic features under each factor, the loading of each feature, etc. Finally, we calculated factor scores of truth and deception in each factor. Then, T-test for individual samples were applied to test whether there was a significant difference between the factor scores. If there is a significant difference in a certain dimension, it indicates that the linguistic features under this dimension might be representative linguistic features of deception.

Copyright 2018 by the author

4 Results and Discussion

This part will firstly present the results of t-test of basic statistical data including word count and word per sentence. Then it will illustrate the results of multidimensional analysis and interpret the major dimensions in details

4.1 Word Count and Word Per Sentence

The results of t-test for individual samples show that the word count of deceptive discourse is smaller than that of truthful discourse and the difference is significant. There is also a significant difference of word per sentence between deceptive discourse and truthful discourse. The results are displayed in Table 4-1.

	Type	Mean	S.D.	Sig. (two-tail)
Word Count	Truth	164.47	48.42	.003
	Deception	142.10	54.82	< .05
Word Per Sentence	True	28.68	13.12	0.046
	Deceptive	25.26	10.88	< .05

Table 4-1. T-test of word count and word per sentence

The results above are consistent with the previous studies. Truthful speeches contain more information than deceptive speeches. Utterance length is a feature that is commonly used to distinguish deceptive speech from truthful speech (Burgoon & Qin, 2006). When liars do not have the opportunity to prepare, the deceptive utterances are usually shorter than truthful ones. The reasons may be twofold. First, lying requires more cognitive energy for thinking and reasoning. Therefore, when the duration of discourse is the same, the utterance length is shorter. Second, some researchers regard this as a deception strategy, because talking less can help them to avoid being detected.

4.2 Results of Multidimensional Analysis

SPSS 20 was applied for multidimensional analysis. The result of KMO and Bartlett's test is 0.779>0.7, meaning that these 38 linguistic features are suitable for the multidimensional analysis. After the analysis, 9 major factors have been extracted. The extraction sum of squared loadings of these 9 factors is up to 73.74%, meaning that these 9 factors can explain 73.74% of the 38 linguistic features selected in this study, which is satisfactory. Due to the limited space of this paper, Table 4-2 only displays the first 7 factors and the linguistic features they contain. This table also lists factor loadings that are greater than 0.4. Negative loadings are excluded.

F1	**Loadings**	**F2**	**Loadings**
space	.778	conj	0.871
number	.697	exclusive	0.821
relative	.695	adverb	0.711
quantative	.685	time	0.636
motion	.650	cogmech	0.627
work	.639	funct	0.582
preps	.551	interjunc	0.537
F3	**Loadings**	**F4**	**Loadings**
ppron	0.861	we	0.773
pronoun	0.778	inclusive	0.115
i	0.754	**F5**	**Loadings**
shehe	0.706	affect	0.890
social	0.567	posemo	0.721
cause	0.514	negemo	0.622
F6	**Loadings**	**F7**	**Loadings**
percept	0.848	certain	0.731
hear	0.721	psycho	0.461
see	0.626	auxverb	0.445

Table 4-2 Seven factors and factor loadings

According to Biber (1995), multidimensional analysis can identity the groupings of linguistic features that have strong co-occurrence associations. These linguistic features with frequent co-occurrence are classified into one dimension or factor. Each factor has a specific function and interpretation to the text. The functions and interpretations of the seven factors in this paper will be discussed later, the explanation of which will be based on Biber's theory.

Then, we may ask whether these factors can distinguish truthful and deceptive texts or not. We can calculate the factor scores and use t-test to see if there is a significant difference between truthful and deceptive discourse in each dimension. The results of t-test are displayed in Table 4-3.

	Type	Mean	Standard Deviation	Sig. (two-tail)
Factor 1	Truth	4.88	3.42	.284 > .05
	Deception	4.31	4.02	
Factor 2	Truth	18.86	7.20	*.001< .05
	Deception	15.52	7.43	
Factor 3	Truth	9.34	6.00	*.002< .05
	Deception	7.06	4.25	
Factor 4	Truth	7.58	4.12	.932 > .05
	Deception	7.53	4.44	
Factor 5	Truth	1.50	2.59	.959 > .05
	Deception	1.48	3.19	
Factor 6	Truth	4.68	4.17	*.004< .05
	Deception	3.08	3.60	
Factor 7	Truth	4.37	3.47	.159 > .05
	Deception	3.70	3.18	

Table 4-3. T-test of factor scores

As illustrated in Table 4-3, on Factor 2, 3, 6, there are significant differences of factor scores between truthful discourse and deceptive discourse, while on Factor 1, 4, 5, 7 the differences between these two types of discourses are not significant. The following part will discuss the data from three aspects: 1) what is the interpretation of the Factor; 2) what does the result of t-test imply; 3) the underlying reasons or possible explanations for the result.

Factor 1 includes linguistic features such as space, number, quantity, relative, preposition, etc. The grouping of these linguistic features may be associated with objective description. Space, number and quantity are objective data that reflect temporal and physical information pertaining to the topic. Besides, according to Biber (1995, pp.136), relativity words "function to specify or elaborate the identities of referents" and prepositions function to provide more information. In addition, interpersonal linguistic features, such as personal pronouns, first-person pronouns and social, have negative loadings on Factor 1, which suggests that Factor 1 depicts impersonal and objective information.

The mean scores of deceptive and truthful texts on Factor 1 are 4.31 and 4.88 respectively. That is to say, generally, truthful texts contain more objective descriptions than deceptive texts. However, the difference is not significant. In other words, there is little difference between deceptive and truthful texts when participants depicted the basic information of the teachers they like or dislike. This is understandable because in this experiment, when participants were lying, they still talked about people that they really know. Therefore, the language they use for objective description about the teachers they dislike is still true. If the topic of this experiment is changed to "lying about a hotel that you have never lived in", then we may predict that the performance of truthful discourse and deceptive discourse on Factor 1 should be different.

Factor 2 consists of linguistic features including conjunction, adverb, time, cognitive mechanism, functional words, verbs, etc. These features have a strong association with narration, because in narratives, people tend to use more verbs, conjunctions and time to keep the flow of the story. On factor 2, the mean factor score of deceptive texts (15.52) is smaller than that of truthful texts (18.86) and the difference is significant (p=.001). This indicates that in deceptive texts, fewer narrative devices are used. The narrative feature of deceptive texts is significantly weaker than truthful speeches.

Factor 3 is obviously connected with interpersonal relationship. It includes linguistic features such as pronouns, first-person pronouns, third-person pronouns, social, cause and so on. Biber (1995) maintains that the use of first or second pronouns reflect the direct participation of the speaker. On this factor, truthful texts have a higher mean score than deceptive texts and the difference is significant, meaning that in truthful texts, participants told more information about interpersonal relationships and include more social interactions, while in deceptive texts, participants conveyed less information on interpersonal relationship. This is reasonable because when participants were talking about teachers they do not like.

Besides, it is worth mentioning that second-person pronoun “you” is not included in Factor 3 probably because the experiment of current study is not a face-to-face conversation. When participants recalled their favorite teachers, they were talking about someone who was not there, thus third-person pronouns were more frequently used than second-person pronouns.

In addition, the first-person plural pronoun “we” is included in Factor 4 rather than Factor 3. Factor 4 is a dimension pertaining to inclusiveness. The concept of inclusiveness is to some extent different from interpersonal relationships depicted by Factor 3. According to Biber (1995), conversations and letters are marked on the dimension of inclusiveness. In this study, the experiment is not interactive in nature, thus involving few inclusive features. Also, the result of t-test shows that there is no significant linguistic difference between truth and deception on this dimension. Therefore, this factor is not very useful to detect deceptive discourse in this study.

Factor 5 includes affect words, such as positive and negative emotions. This factor is associated with people’s attitudes and emotions. The mean score of truthful discourse on this factor is 1.50, which is higher than that of deceptive discourse (1.48). However, the difference is not significant. That is to say the Factor of emotion is not effective enough to distinguish truth and deception in this study. Before the study, we hypothesize that when participants lie about their favorite teachers, they might use fewer emotional words. However, the result is not in accordance with our predication. In this study, truthful and deceptive texts have no significant difference on this dimension. Previous studies also have contradictory results on this dimension. A possible guess would be that the dimension of emotion is sensitive to the topic.

Factor 6 involves linguistic features pertaining to perception, such as see, hear and feel. The t-test result indicates that the linguistic features of perception are more salient in truthful speeches than in deceptive ones. This may suggest that when people are telling truth, they tend to depict their feelings in more details, because they have really experienced that process. However, in deceptive speeches, since they have not experienced the process, they are unable to provide too many direct feelings. Therefore, dimension of perception can be utilized to distinguish truth and deception in this study.

Factor 7 consists of features, such as possibility modals, which are expressions of certainty or uncertainty. The result illustrates that the mean value of truthful discourse on this factor is higher than that of deceptive ones, but the difference is not significant. This may suggest that the dimension of modals is not very idealistic for distinguishing truthful and deceptive discourse in this study.

5 Conclusion

This study applies multidimensional analysis to see whether linguistic dimensions are capable of distinguishing truthful and deceptive discourse. After the multidimensional analysis, nine major factors were extracted. Seven of them were analyzed in details. The result shows that three factors are salient in differentiating deceptive texts from truthful texts. They are: 1) Factor 2: Linguistic features about narration, such as conjunctions, time, verbs and etc. 2) Factor 3: Interpersonal features including first-person pronouns, third-person pronouns and so on. 3) Factor 6: Perceptive features, including seeing, hearing and feeling. On these three factors, truthful texts have higher factor scores than deceptive texts and the differences are significant. Therefore, these three factors are capable of distinguishing lies from truth in this study.

This study also provides several implications. First, it is effective to study linguistic cues of deception through multidimensional analysis based on corpus. Multidimensional analysis can classify linguistic features into dimensions systematically and scientifically. Second, some linguistic features or dimensions are sensitive to the topic. Factors that cannot distinguish deceptions and truth in this study may be capable of doing that when the topic is changed. Therefore, if we want to detect deception of a certain text type, we can use hundreds of samples of truthful and deceptive texts to do a multidimensional analysis. After finding out the patterns, we can predict whether a new text is deceptive or not by calculating its factor scores on each dimension.

Copyright 2018 by the author

References

Biber, D. (1995). *Dimensions of Register Variation: A Cross-linguistic Comparison*. Cambridge: Cambridge University Press.

Bond, G. D., & Lee, A. Y. (2005). Language of lies in prison: Linguistic classification of prisoners' truthful and deceptive natural language. *Applied Cognitive Psychology, 19,* 313-329.

Burgoon, J. K., & Qin, T. (2006). The dynamic nature of deceptive verbal communication. *Journal of Language and Social Psychology, 25*(1), 76-96.

Burgoon, J., Mayew, W. J., Giboney, J. S., Elkins, A. C., Moffitt, K., Dorn, B., Byrd, M., & Spitzley, L. (2016). Which Spoken Language Markers Identify Deception in High-Stakes Settings? Evidence From Earnings Conference Calls. *Journal of Language and Social Psychology, 35*(2) 123–157.

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), 979-995.

DePaulo, B.M., Lindsay, J.J., Malone, B.E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to Deception. Psychological Bulletin, 129(1), 74-118.

Ekman, P. (1992). *Telling lies, Clues to Deceit in the Marketplace, Politics, and Marriage*. New York: W.W. Norton & Company.

Fitzpatrick, E., & Bachenko, J. (2012). Building a data collection for deception research. *Proceedings of the EACL 2012 Workshop on Computational Approaches to Deception Detection*. 31-38.

Hancock, J. T., Thom-Santelli, J., & Ritchie, T. (2004, April). Deception and design: The impact of communication technology on lying behavior. *In Proceedings of the SIGCHI conference on Human factors in computing systems*. 129-134.

Hauch, V., Blando ́n-Gitlin, I., Masip, J. & Sporer, S. L. (2012). Linguistic Cues to Deception Assessed by Computer Programs: A Meta-Analysis. *Proceedings of the EACL 2012 Workshop on Computational Approaches to Deception Detection*. 1-4.

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,* 665–675.

Pennebaker, J.W., Francis M.E. & Booth, R.J. (2001). *Linguistic Inquiry and Word Count* (LIWC). Erlbaum Publishers.

Vrij, A., Edward, K., Roberts, K. P., & Bull, R. (2000). Detecting deceit via analysis of verbal and nonverbal behavior. *Journal of Nonverbal behavior*, *24*(4), 239-263.

.

Copyright 2018 by the author