World Academy of Science, Engineering and Technology
International Journal of Cognitive and Language Sciences
Vol:13, No:6, 2019

The Analysis of Deceptive and Truthful Speech: A Computational Linguistic Based Method

Seham El Kareh, Miramar Etman

***Abstract*—**Recently, detecting liars and extracting features which distinguish them from truth-tellers have been the focus of a wide range of disciplines. To the author's best knowledge, most of the work has been done on facial expressions and body gestures but only few works have been done on the language used by both liars and truth-tellers. This paper sheds light on four axes. The first axis copes with building an audio corpus for deceptive and truthful speech for Egyptian Arabic speakers. The second axis focuses on examining the human perception of lies and proving our need for computational linguistic-based methods to extract features which characterize truthful and deceptive speech. The third axis is concerned with building a linguistic analysis program that could extract from the corpus the inter- and intra-linguistic cues for deceptive and truthful speech. The program built here is based on selected categories from the Linguistic Inquiry and Word Count program. Our results demonstrated that Egyptian Arabic speakers on one hand preferred to use first-person pronouns and present tense compared to the past tense when lying and their lies lacked of second-person pronouns, and on the other hand, when telling the truth, they preferred to use the verbs related to motion and the nouns related to time. The results also showed that there is a need for bigger data to prove the significance of words related to emotions and numbers.

***Keywords*—**Egyptian Arabic corpus, computational analysis, deceptive features, forensic linguistics, human perception, truthful features.

I. Introduction

DECEPTION is a deliberate attempt to mislead others[1], and in turn, this definition will exclude: Self-deception, delusion, pathological behavior, and falsehoods due to ignorance/error. This definition focuses on the fact that humans are aware when lying to others.

Previous research and practitioner experience suggest that acoustic/prosodic and lexico/syntactic cues may signal that speakers when speaking are deceptive. While some of these cues are proposed as general, at least within a culture, there is also some evidence from diverse findings for phenomena such as pitch variation and disfluency production that there is considerable individual variation as well [1]. This study aims to examine the linguistic features that characterize deceptive speech and prove that humans perform very poorly at the task of detecting liars, and in order to examine those features or to prove the poor human perception, one of our goals is building a clean spoken corpus for deceptive and truthful speech for Arabic native speakers speaking Egyptian dialect.

Seham El Kareh is Professor of Linguistics, Department of Phonetics and Linguistics, Faculty of Arts, Alexandria University, Egypt (phone: 00201224262178, e-mail: sehamelkareh@alexu.edu.eg).

Miramar Etman is with the Department of Phonetics and Linguistics, Faculty of Arts, Alexandria University, Egypt (phone: 00201061158091, e-mail: miramar.etman@alexu.edu.eg).

II. Literature Review

A. Theory

There are many theories that were introduced by psychologists on the phenomena of lying and deception. But Paul Ekman theory [5] is the most influential one. In his theory, Ekman introduced reasoning strategies for deception such as concealment, falsification, misdirection and many other strategies. His theory is based upon that cues to deception will result from one of two flaws: the first flaw is leakage, and by this he means that part of the truth will be exposed by the liar, and the second flaw is deception cues in which he suggests there is direct indication that the person is lying as there is inconsistency in his story. In the process of developing his substantial theory, Ekman considers in detail the implications of his ideas with respect to lexical and prosodic components of speech, physical behavior, and especially, facial expressions [2].

B. Technologies for Detecting Deception

There is a wide range of technologies available to detect deception. These technologies mainly depend on biometric factors (as incensement in blood pressure, in perspiration) or depend on brain imaging strategies as Magnetic Resonance Imaging (MRI).

1. Polygraph

One of the most famous and common technology used in detecting deception is the polygraph or lie detector. Polygraph is based on measuring and recording several physiological factors such as blood pressure, pulse, respiration, and skin conductivity; all of these factors are measured while a person is answering a set of questions. Although this device is used to detect liar's reactions while lying, these physiological reactions may also occur in an individual who is suffering from stress or fear for some other reason and not lying. In addition to that, polygraph technology excludes features that are related to speech such as: Linguistic cues, prosodic and acoustic cues.

2. Magnetic Resonance Imaging (MRI)

MRI is one of the most famous technologies of brain imagining that is used to detect liars, and focuses on brain activity. Functional neuroimaging techniques (especially functional magnetic resonance imaging) have been used to study deception. In the human adult, deception and lying exhibit features consistent with their use of 'higher' or

'executive' brain systems [3].

C. Empirical Work to Detecting Deception

In this section, the researcher tries to highlight the most important empirical studies in the field of deception that focused on linguistic cues to deception.

Newman's study [4] focused on examining the linguistic manifestation of false stories by using Linguistic Inquiry and Word Count (LIWC); a linguistic analysis program that analyses either written or spoken samples on a word-by-word basis. The linguistic analysis of the false stories is done by comparing each word in the text against a file that has more than 2000 words which are divided into 72 linguistic dimensions. The Newman study was divided into five studies.

The first study included 101 undergraduate students who were videotaped while discussing both their true and false opinions on abortion. The second study included 44 undergraduates who were asked to type both their true and false opinions concerning abortion. The third study included 55 undergraduates who were also asked to write their truthful and deceptive descriptions on the issue of abortion in a counterbalanced order. The rest of the studies' targets were that the participants provide their true and false opinions.

Newman found that across the five studies, deceptive communications were characterized by fewer first-person singular pronouns and discusses that this may relate to the fact that liars attempt to disassociate themselves from the lie, fewer third-person pronouns and more negative emotions words could relate to the fact that liars may feel guilty about the lie or the topic they are lying about, and fewer exclusive words and more motions verbs suggest lower cognitive complexity. Because liars' stories are by definition fabricated, some of their cognitive resources are taken up by the effort of creating a believable story.

III. Methodology

One of the problems that researchers encounter, especially linguists in the field of detecting liars from truth tellers, is the lack of a spoken corpus that comprises both deceptive and truthful speech that could be subjected to further analysis.

This part focuses on two main points in the research. The first point is related to building the spoken corpus data and the second point is related to examining the human perception to deception.

A. Data Collection

A first step in the study was building a spoken corpus that could be subjected later to linguistic analysis.

1. Study Procedures

To ensure the validity of the data as much as is possible, the subjects should feel comfortable and unstressed in order not to affect the results later; for this reason, the subjects were told that they are to record three stories (one lie, two truth) as a part of a game which is known worldwide. They also were asked to order their stories as they wish and try to convince others who will later listen to their stories (as a part of the game) that all of their stories are true.

The choice of the topic was free to the subjects; the subjects were told to talk about any topics they feel comfortable about. Each subject was a given a sheet (see Fig. 1) to mark whether the story they said was true or lie. The sheet also included personal information about the subject such as name, gender and age.

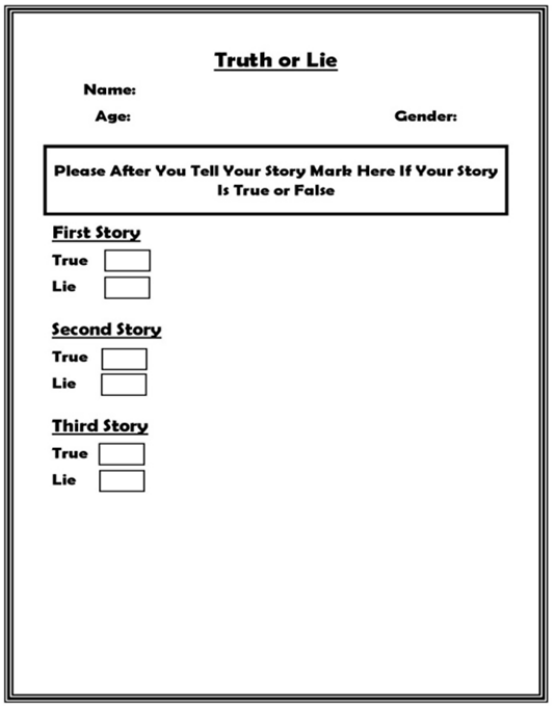

Fig. 1 The sheet given for each subject to write examples of true or false stories

The subject first enters the recording room of the speech lab with the researcher who will explain the procedures of the game and the recording process. Then, the subject was left alone to record his/her stories (both truthful and deceptive stories); however, before leaving the room, the researcher told the subject that their stories could not be heard outside in the main room and that they would only be listened to at a later time. The aim of this method is to make the subject feel relaxed and comfortable, and free to talk.

In addition, there was a convention about a group of signs between the researcher and the subject for the recording process (an example includes a sign from the subject to the researcher to start the recording session once he/she is ready and another sign when he/she finishes their story to stop the recording session), the researcher sees those signs through a glass window which separates the recording room from the main room of the speech lab.

2. Subjects

The subjects who participated in the recording process consisted of 16 (six males and 10 females) undergraduate students from Alexandria University, Faculty of Arts. However, for the final study, only seven (three males and four females) subjects out of the 16 students were selected for

further experimentation and analysis.

All the subjects are Arabic Native speakers who speak Egyptian dialect; in addition, all the speakers live in Alexandria. Their ages ranged from 20 – 23 years old (that small age gap between the subjects would help in examining the linguistic features used by this age range in deceptive and non-deceptive speech).

The subject's name was replaced by a number to protect their privacy (see Table I).

TABLE I
SUBJECTS NUMBER, GENDER AND AGE

Subject Number	Gender	Age
Subject Number 1	Male	23 Years
Subject Number 2	Male	22 Years
Subject Number 3	Male	22 Years
Subject Number 4	Female	21 Years
Subject Number 5	Female	20 Years
Subject Number 6	Female	20 Years
Subject Number 7	Female	22 Years

TABLE II
SAMPLE FROM THE PERCEPTUAL EXPERIMENT PARTICIPANTS NUMBER, GENDER AND AGE

Subject Number	Gender	Age
Participant Number 1	Female	21 Years
Participant Number 2	Female	21 Years
Participant Number 3	Female	20 Years
Participant Number 4	Female	19 Years
Participant Number 5	Female	19 Years
Participant Number 6	Female	22 Years
Participant Number 7	Female	21 Years
Participant Number 8	Female	20 Years
Participant Number 9	Female	21 Years
Participant Number 10	Female	20 Years
Participant Number 11	Female	24 Years
Participant Number 12	Female	22 Years
Participant Number 13	Female	21 Years
Participant Number 14	Female	23 Years
Participant Number 15	Female	18 Years
Participant Number 16	Female	18 Years
Participant Number 17	Female	18 Years
Participant Number 18	Female	19 Years
Participant Number 19	Female	20 Years
Participant Number 20	Female	19 Years
Participant Number 21	Female	19 Years
Participant Number 22	Female	19 Years
Participant Number 23	Female	21 Years
Participant Number 24	Female	21 Years
Participant Number 25	Female	19 Years
Participant Number 26	Male	20 Years
Participant Number 27	Male	19 Years
Participant Number 28	Male	19 Years
Participant Number 29	Male	21 Years
Participant Number 30	Male	21 Years
Participant Number 31	Male	20 Years

3. Recording Setting/Equipment

The experiment was conducted in the speech lab of the Phonetics and Linguistics Department, Faculty of Arts, Alexandria University. The CSL recording system was used in recording the data for deceptive and truthful speech.

The speech lab consists of two rooms, the main room and the recording room. In the recording room there is a chair and a desk with a microphone and pen and assessment sheets for the recording process.

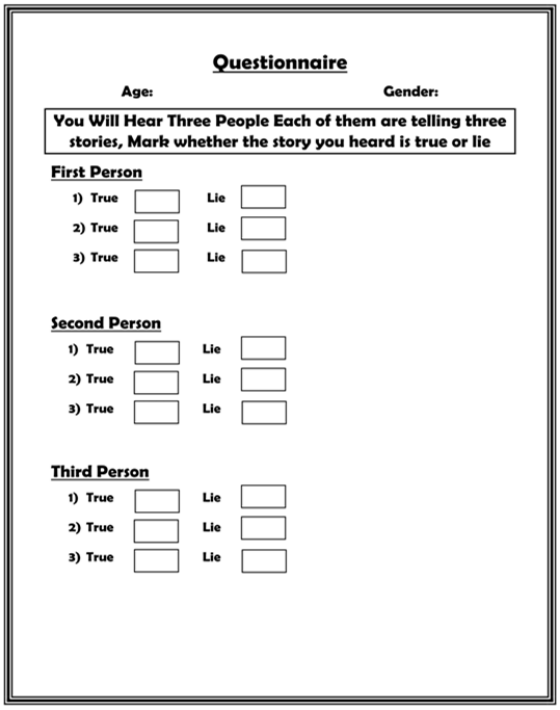

Fig. 2 The sheet given for each participant to mark whether the stories they are hearing are true or lies as a part of the perceptual experiment

B. Perceptual Experiment

Humans are notoriously poor at detecting deception [2], and as such, this experiment was done in favor of examining the human perception to detecting deception and later to compare its results with those from the computer program. This experiment is based on the previously collected data for deceptive and truthful speech.

C. Experiment Procedures

As a first step, the researcher explains to the participant(s) that they are about to hear nine audio samples from three individuals (each individual three audio samples) and that their task is to try to guess if each individual is telling the truth or lying. In addition, participants were told that the experiment is a kind of guessing game, and also the researcher did not give the participants (hearers) any information about how many stories are true and how many are deceptive.

The aim of not informing participants that there was only one deceptive story is to develop a better understanding of how humans perceive deception, because in real life, we are generally not aware of how many lies a person would tell.

The subjects were also given an assessment sheet (see Fig. 2) to mark whether what they heard was the truth or a lie; each of the participants listened to a total of nine stories (three

individuals each telling three stories).

For the listening experiment, one to five participants per session were seated in the speech lab and the researcher played the prerecorded stories; they were also advised not to communicate with each other.

After the participants (hearers) completed the listening experiment, the researcher informed them if they were able to detect the lie or not.

D. Participants

The perceptual experiment consisted of 31 participants (25 females and six males). All subjects were undergraduate students from Alexandria University, Faculty of Arts, and they ranged in age from 18 – 24 years old. The following table shows the participants number, gender and age. The name of each of the participants was replaced by a number to protect their privacy (see Table II).

E. Experiment Setting and Equipment

The experiment took place in the speech lab at Alexandria University, Faculty of Arts. The participants listened to the stories through a loud speaker while they were sitting on chairs in the lab. The participants (hearers) did not listen to each of the individual three stories consecutively, instead the participants (hearers) were given 1-2 minutes between each story and 1-2 minutes between each individual speaker to consider their thoughts and complete their evaluation of the stories on the provided sheet (see Fig. 2).

F. Experiment Results

The results, as presented in Table III, show that human perception for detecting deception is very poor. Humans try to detect deception randomly, and therefore, in most cases, they will not be able to detect liars; as well, they may even perceive the truth as if it was a lie as it is the case with subject number two which shows that 9 out of 10 participants (hearers) perceived his deceptive story as a true story. The results of this experiment show that humans cannot be accepted as accurate judges as to whether what they hear is truth or lies (see Table III).

G. Feature Selection for Analysis

A number of studies suggest that word usage provides an important cue to deception [6], [7]. The researchers in this study based their analysis on the audio corpus by selecting linguistic categories from LIWC that were significant across previous studies on deception. In addition, the researcher selected features that were evident in the collected corpus. The total number of categories examined in this study was 13 (see Table IV). The audio corpus of truthful and deceptive data was analyzed with respect to the selected categories (see Table IV) by using a linguistic analysis program that was built by the researchers.

The program calculates the percent of usage of each category and the number of words related to each category for each of the individual recorded stories.

TABLE III
THE PERCEPTUAL EXPERIMENT RESULTS

Subject Number	Stories		
	Males		
Subject Number One	First Story (True Story)	Second Story (True Story)	Third Story (Lie Story)
Number of People Perceived the Story as a Lie	6	5	5
Number of people Perceived the Story as the Truth	4	5	5
Subject Number Two	First Story (True Story)	Second Story (True Story)	Third Story (Lie Story)
Number of People Perceived the Story as a Lie	10	3	1
Number of people Perceived the Story as the Truth	0	7	9
Subject Number Three	First Story (True Story)	Second Story (True Story)	Third Story (Lie Story)
Number of People Perceived the Story as a Lie	5	3	5
Number of people Perceived the Story as the Truth	5	7	5
	Females		
Subject Number Four	First Story (True Story)	Second Story (True Story)	Third Story (Lie Story)
Number of People Perceived the Story as a Lie	6	3	5
Number of people Perceived the Story as the Truth	4	7	5
Subject Number Five	First Story (True Story)	Second Story (Lie Story)	Third Story (True Story)
Number of People Perceived the Story as a Lie	4	4	0
Number of people Perceived the Story as the Truth	6	6	10
Subject Number six	First Story (True Story)	Second Story (True Story)	Third Story (Lie Story)
Number of People Perceived the Story as a Lie	4	3	2
Number of people Perceived the Story as the Truth	6	7	8
Subject Number seven	First Story (True Story)	Second Story (Lie Story)	Third Story (True Story)
Number of People Perceived the Story as a Lie	5	2	3
Number of people Perceived the Story as the Truth	5	8	7

IV. DATA ANALYSIS AND DISCUSSION

In this section, the researcher will show and discuss the results of the calculations of the linguistic analysis program that was built using Python. As mentioned previously, most of the linguistic dimensions used here examine those linguistic features that characterize both deceptive and truthful speech. Both inter and intra linguistic features were examined. In order to find the linguistic cues characterizing deceptive and truthful speech among all the speakers, intra linguistic cues where taken as a first step. We should take in consideration the habitual way of certain speakers when telling their story, for example, while telling the deceptive story subject number 7 did not use any words related to numbers, while subject number 2 used eight words related to numbers. So, what seems to be a sign of lying for one person (as in the absence of words related to numbers with subject number 7) may not be the case for another.

From the Linguistic Analysis Program output it seems that subject 1 uses first-person pronouns more in his deceptive

story, and more importantly, it is clear he did not use any second-person pronouns in his deceptive story at all. Meanwhile, both motion and time verbs appear to characterize his truthful speech. In addition to that, emotional words, especially negative emotions, characterize his deceptive speech (see Tables V and VI).

TABLE IV
THE LINGUISTIC ANALYSIS PROGRAM CATEGORIES

Categories
Total Pronouns
First Pronouns
Second Pronouns
Third Pronouns
Total Verbs
Past Verbs
Present Verbs
Future Verbs
Imperatives
Motion Verbs
Time Words
Foreign Words
Causation Words
Emphatic Words
Sense Words
Numbers
Prepositions
Negations
Total Disfluencies
Fillers Percent
Interjections
Emotional Words
Negative Emotions
Positive Emotions

TABLE V
THE LINGUISTIC ANALYSIS PROGRAM OUTPUT FOR SUBJECT NUMBER 1 IN PERCENT

Features	True Story First Story Percent of Number of Occurrences	True Story Second Story Percent of Number of Occurrences	Lie Story Third Story Percent of Number of Occurrences
Total Pronouns	5.33 %	4.13 %	3.66 %
First Pronouns	50.0 %	73.68 %	100.0 %
Second Pronouns	50.0 %	10.52 %	0.0 %
Third Pronouns	0.0 %	15.78 %	0.0 %
Total Verbs	12.26 %	17.21 %	15.59 %
Past Verbs	47.82 %	54.43 %	41.17 %
Present Verbs	52.17 %	45.56 %	52.94 %
Future Verbs	0.0 %	0.0 %	5.88 %
Imperatives	0.0 %	0.0 %	0.0 %
Motion Verbs	36.06 %	27.84 %	11.76 %
Time Words	1.06 %	1.30 %	0.0 %
Foreign Words	1.86 %	0.21 %	0.0 %
Causation Words	0.53%	0.43 %	0.0 %
Emphatic Words	1.33 %	2.39 %	3.66 %
Sense Words	0.0 %	0.0 %	0.0 %
Numbers	1.33 %	2.17 %	1.83 %
Prepositions	6.66 %	4.57 %	6.42 %
Negations	3.2 %	2.61 %	1.83 %
Total Disfluencies	21.86 %	18.30 %	13.76 %
Fillers Percent	43.90 %	47.61 %	20.0 %
Interjections	56.09 %	52.38 %	80.0 %
Emotional Words	2.13 %	3.48 %	14.67 %
Negative Emotions	0.0 %	12.5 %	87.5 %
Positive Emotions	100.0 %	87.5 %	12.5 %

TABLE VI
THE LINGUISTIC ANALYSIS PROGRAM OUTPUT FOR SUBJECT NUMBER 1 IN NUMBERS

Features	True Story First Story Number of Occurrences	True Story Second Story Number of Occurrences	Lie Story Third Story Number of Occurrences
Total Words	375	459	109
Total Pronouns	20	19	4
First Pronouns	10	14	4
Second Pronouns	10	2	0
Third Pronouns	0	3	0
Total Verbs	46	79	17
Past Verbs	22	43	7
Present Verbs	24	36	9
Future Verbs	0	0	1
Imperatives	0	0	0
Motion Verbs	17	22	2
Time Words	4	6	0
Foreign Words	7	1	0
Causation Words	2	2	0
Emphatic Words	5	11	4
Sense Words	0	0	0
Numbers	5	10	2
Prepositions	25	21	7
Negations	12	12	2
Total Disfluencies	82	84	15
Fillers Percent	36	40	3
Interjections	46	44	12
Emotional Words	8	16	16
Negative Emotions	0	2	14
Positive Emotions	8	14	2

TABLE VII
THE LINGUISTIC ANALYSIS PROGRAM OUTPUT FOR SUBJECT NUMBER 2 IN PERCENT

Features	True Story First Story Percent of Number of Occurrences	True Story Second Story Percent of Number of Occurrences	Lie Story Third Story Percent of Number of Occurrences
Total Pronouns	4.14 %	1.96 %	5.17 %
First Pronouns	57.14 %	100.0 %	66.66 %
Second Pronouns	28.57 %	0.0 %	0.0 %
Third Pronouns	14.28 %	0.0 %	33.334 %
Total Verbs	22.48 %	12.74 %	17.24 %
Past Verbs	47.36 %	76.92 %	40.0 %
Present Verbs	39.47 %	23.07 %	50.0 %
Future Verbs	0.0 %	0.0 %	10.0 %
Imperatives	13.15 %	0.0 %	0.0 %
Motion Verbs	39.47 %	7.69 %	30.0 %
Time Words	4.14 %	9.80 %	2.87 %
Foreign Words	0.0 %	1.96 %	0.0 %
Causation Words	4.73 %	0.0 %	1.72 %
Emphatic Words	5.91 %	1.96 %	4.02 %
Sense Words	0.59 %	0.0 %	0.0 %
Numbers	4.14 %	12.74 %	4.02 %
Prepositions	2.36 %	6.86 %	4.59 %
Negations	2.95 %	0.0 %	1.72 %
Total Disfluencies	11.24 %	9.80 %	8.04 %
Fillers Percent	21.05 %	10.0 %	14.28 %
Interjections	78.94 %	90.0 %	85.71 %
Emotional Words	2.36 %	0.98 %	0.57 %
Negative Emotions	75.0 %	100.0 %	0.0 %
Positive Emotions	25.0 %	0.0 %	100.0 %

Another example for the Program output is the analysis of subject number 2 who agrees with subject number 1, as both of them preferred to use first-person pronouns in their deceptive story, and their deceptive story lacks the use of second-person pronouns. Here, also, time words are noticeable in the truthful speech. Subject number 2 data showed that present tenses appear to characterize his deceptive speech, while past tenses tend to indicate truthful speech (see Tables VII and VIII).

TABLE VIII
THE LINGUISTIC ANALYSIS PROGRAM OUTPUT FOR SUBJECT NUMBER 2 IN NUMBERS

Features	True Story First Story Number of Occurrences	True Story Second Story Number of Occurrences	Lie Story Third Story Number of Occurrences
Total Words	169	102	174
Total Pronouns	7	2	9
First Pronouns	4	2	6
Second Pronouns	2	0	0
Third Pronouns	1	0	3
Total Verbs	38	13	30
Past Verbs	18	10	12
Present Verbs	15	3	15
Future Verbs	0	0	3
Imperatives	5	0	0
Motion Verbs	15	1	9
Time Words	7	10	5
Foreign Words	0	2	0
Causation Words	8	0	3
Emphatic Words	10	2	7
Sense Words	1	0	0
Numbers	7	13	7
Prepositions	4	7	8
Negations	5	0	3
Total Disfluencies	19	10	14
Fillers Percent	4	1	2
Interjections	15	9	12
Emotional Words	4	1	1
Negative Emotions	3	1	0
Positive Emotions	1	0	1

V. CONCLUSION

The analysis of the audio corpus showed that most of the subjects preferred to use first-person pronouns in their deceptive stories, and most of these stories lacked second-person pronouns. The majority of subjects also preferred to use present tense rather than past tense when telling their deceptive stories. In addition, the motion verbs and words related to time categories were most commonly used in truthful speech. As for the words related to numbers and words that describe emotions and feelings, these also seem to be related to truthful speech.

REFERENCES

[1] DePaulo, B. M., Lindsay, J. J., Malone, B. E., et al. 2003. Cues to deception. Psychological Bulletin. 129: 74-118.
[2] Frank Enos. 2009. Detecting Deception in Speech. Ph.D. Dissertation. Columbia Univ., New York, NY, USA. Advisor(s) Julia B. Hirschberg.
[3] Spence SA, Hunter MD, Farrow TF, Green RD, Leung DH, Hughes CJ, et al. A cognitive neurobiological account of deception: evidence from functional neuroimaging. Philos Trans R Soc Lond B Biol Sci. 2004.
[4] Newman, M. L., Pennebaker, J. W., Berry, D. S. & Richards, J. M. (2003). Lying words: Predicting deception from linguistic style. Personality and Social Psych. Bull., 29,665–675.
[5] Ekman, P. (2001). Telling Lies, Clues to Deceit in the Marketplace, Politics, and Marriage (2nd Ed.). New York: W.W. Norton & Co.
[6] Newman, M. L., Pennebaker, J. W., Berry, D. S. & Richards, J. M. (2003). Lying words: Predicting deception from linguistic style. Personality and Social Psych. Bull., 29, 665–675.
[7] Zhou, L., Burgoon, J. K., Twitchell, D. P., Qin, T. & Nunamaker, J. F. (2004). A comparison of classification methods for predicting deception in computer-mediated communication. Journal of Management Information Systems, 20(4), 139–165.