Dialogue & Discourse 11(2) (2020) 128-149 doi: 10.5210/dad.2020.205

Please, Please, Just Tell Me: The Linguistic Features of Humorous Deception

Stephen Skalicky STEPHEN.SKALICKY@VUW.AC.NZ
School of Linguistics and Applied Language Studies
Victoria University of Wellington

Nicholas D. Duran NDURAN4@ASU.EDU
School of Social and Behavioral Sciences
Arizona State University

Scott A. Crossley SCROSSLEY@GSU.EDU
Department of Applied Linguistics and ESL
Georgia State University

Editor: Amir Zeldes

Submitted 03/2020; Accepted 11/2020; Published online 11/2020

Abstract

Prior research undertaken for the purpose of identifying deceptive language has focused on deception as it is used for nefarious ends, such as purposeful lying. However, despite the intent to mislead, not all examples of deception are carried out for malevolent ends. In this study, we describe the linguistic features of humorous deception. Specifically, we analyzed the linguistic features of 753 news stories, 1/3 of which were truthful and 2/3 of which we categorized as examples of humorous deception. The news stories we analyzed occurred naturally as part of a segment named *Bluff the Listener* on the popular American radio quiz show *Wait, Wait... Don't Tell Me!*. Using a combination of supervised learning and predictive modeling, we identified 11 linguistic features accounting for approximately 18% of the variance between humorous deception and truthful news stories. These linguistic features suggested the deceptive news stories were more confident and descriptive but also less cohesive when compared to the truthful new stories. We suggest these findings reflect the dual communicative goal of this unique type of discourse to simultaneously deceive and be humorous.

Keywords: deception, humor, lexical semantics, applied natural language processing

1. Introduction

"You, of course, are going to play the game in which you must try to tell truth from fiction"

A person who fibs, lies, or is otherwise untruthful during a conversation possesses a decided interactional advantage: they alone are aware of their deception and can thus use that knowledge to their own benefit. This type of everyday deception ranges from the altruistic white lies used in emotionally close relationships (DePaulo and Kashy, 1998) to duplicitous speech intended to prevent interlocutors from discovering a truth (Gupta and Ortony, 2018). However, the intent to deceive need not only serve nefarious or malevolent ends. Indeed, sometimes deception can be viewed through a positive lens: as a form of creativity (Kapoor and Khan, 2017) that may evoke a sense of humor or mirth (Dynel, 2011). Satirical television news shows, humorous movie spoofs,

and radio quiz shows can all employ humorous deception to varying degrees for the purposes of entertainment. It is this type of humorous deception we address in the current study.

Because deception is typically viewed as an inherently negative conversational act, some research has worked to classify and measure features of deceptive speech so that deceptive speakers may be more easily identified. Among the different markers of deceptive conversation, quantifiable linguistic features have proven to be a promising method for the automatic detection of deception (Duran, 2009; Duran et al., 2010; Meibauer, 2018). However, deception in these studies is normally operationalized in the negative sense. This is done by asking research participants to lie or otherwise be purposefully deceptive during prompted interaction (Hancock et al., 2008; Van Swol et al., 2012), or by analyzing examples of deceptive communication (Ludwig et al., 2016). As such, the linguistic features associated with deception in these studies are based on deception for the purpose of lying in order to prevent interlocutors from discovering a truth.

The purpose of the current study is to examine the linguistic features of creative and humorous deception in a context where all parties are aware of deceptive intent. We employ a wide range of linguistic features related to lexical, syntactic, and affective features of language. By exploring which of these linguistic features (if any) can distinguish humorous deception from truth, we aim to provide a better understanding of the linguistic features of deception as it occurs in a different communicative context with different conversational goals. Moreover, though a comparison to prior studies examining malevolent deception and humor, our study can provide insight as to whether similar linguistic strategies are used during two very different types of deception. Specifically, we examine humorous deception as it occurs in a popular syndicated radio quiz show in the United States named *Wait Wait... Don't Tell Me!*.

2. The *Wait Wait... Don't Tell Me* Radio Show

Wait Wait... Don't Tell Me! (*WWDTM*) is a weekly radio program produced by a public radio station in Chicago (WBEZ) along with National Public Radio (NPR). NPR is a national, non-profit radio broadcasting company in the United States which syndicates *WWDTM* across the country. While most radio shows on NPR take a serious perspective towards reporting and commenting upon news and politics, the purpose of *WWDTM* is to provide an hour of levity. *WWDTM* does so through virtue of being a quiz show. With the aid of host Peter Sagal and quiz judge Bill Kurtis (who recently replaced long-time judge Carl Kasell), a pool of recurring panelists and radio callers compete in various trivia games and activities. Regular panelists for the show include successful writers, actors, journalists, comedians, and more. The topics of the quizzes are always related to recent news topics which occurred in the prior week. The panelists typically provide humorous banter about the various topics discussed on each episode. In its current version, *WWDTM* is taped before a live audience on Thursdays before being broadcast across the United States every Saturday or Sunday on various NPR stations.

2.1 The *Bluff the Listener* Game

Among the different recurring game segments on *WWDTM*, one game in particular includes deception as a fundamental component. This game, called *Bluff the Listener* (*BTL*), usually occurs somewhere in the middle of each *WWDTM* radio broadcast. During the *BTL* game, the weekly panelists are tasked with bluffing a radio caller who participates via telephone. Radio callers are chosen by the producers of *WWDTM* from a pool of listeners who apply in advance and agree to be

available during the taping of a particular broadcast. Each *BTL* segment begins with small talk between the host (Sagal) and the radio caller, wherein Sagal typically comments upon the geographic location of the radio caller. After this brief chat, Sagal then begins the *BTL* game by informing the participant that it is now time for them to play the "game in which you must try to tell truth from fiction." Sagal then hints at an event which has actually happened, usually with humorous flair, but purposefully leaves out specific details of that event. For example, Sagal introduced the 11 January 2020 *BTL* topic as follows:

> "As long as there has been homework, there have been excuses for not handing it in. The paleontological record shows a T-Rex once claimed his arms were too short to fill out the sheets. [LAUGHTER] This week, we heard an excuse we had never, though, heard before. It was a pretty good one. Our panelists are going to tell you about it. Pick the truthful one, you'll win our prize - the WAIT WAIT-er of your choice on your voicemail. You ready to play?" (NPR, 2020)

The three panelists then take turns reading aloud a news story that could potentially be a fuller description of the event mentioned by Sagal. The task of the radio caller is to identify which story among the three is the actual, truthful description of the event. If the radio caller correctly guesses the truthful story, they are rewarded by having a member of the show leave a message on their answering machine or voicemail service. The panelist who created the selected story also earns a point: the currency used to determine the overall winner of each *WWDTM* broadcast.

2.1.1 THE *Bluff the Listener* NEWS STORIES

The *BTL* news stories are all revised or entirely invented by the panelists. According to the *WWDTM Frequently Asked Questions* website, panelists are provided with a real news story approximately two days before taping of the episode. While one panelist is charged with presenting the real news story, the other two panelists are asked to create their own fictional news stories which are similar to the real story. The real news stories that are chosen each week are always incredible, difficult-to-believe scenarios that tend to naturally arouse suspicion. For example, the real news story associated with the homework topic cited above reported on two young Canadian snowboarders who burned their homework to keep warm after becoming lost in the wilderness due to a snowstorm.

While it is not clear exactly how much content is added by the panelist to the real news story each week, these stories tend to include punchlines at the end which highlight the incredulous and humorous nature of the story. For instance, the panelist presentation of the Canadian snowboarders' news story ended with "no word if their teachers reassigned the homework or just gave them a B for burnt." The wordplay in this final sentence ("B for burnt") provides just enough of a humorous flourish to plant a seed of doubt in the radio caller's mind as to the story's authenticity and works to reinforce the overall humorous frame of the *BTL* game.

The other two news stories are completely invented but structured to conform to the rhetorical genre of a news story. For example, to accompany the Canadian snowboarding story, one panelist wrote a fictional report describing how a graduate student studying problem-solving behavior in primates kept awakening to find his laptop missing (which contained his homework). The mystery was solved when it was later discovered the orangutan under study had devised a method for escaping its enclosure and was hiding the laptop under the orangutan's bed at night. An ironic and humorous effect is thus created through the lack of awareness of the orangutan's gifted problem-solving (by virtue of escaping and returning each night) on the part of the graduate student. The second

fictional *BTL* story included real people. Specifically, this story retold an apparent confession by famous rapper Snoop Dogg, known for his proclivity to smoke marijuana, who admitted he once accidently used his son's final biology paper to roll a comically large joint during a recording session with his colleagues. According to this fictional story, Snoop Dogg was plagued with regret and felt obliged to call the school and explain what happened, ultimately donating a large sum of scholarship money as a way of apology. The humor in this story is primarily found in the incongruity between expected parent-child relationship roles and what typically counts as an acceptable excuse for losing one's homework.

2.1.2 CLASSIFYING THE *BTL* NEWS STORIES

The invented *BTL* news stories are not a case of common, everyday deception associated with lying and untruthfulness as it occurs in regular conversation (Duran and Fusaroli, 2017; DePaulo and Kashy, 1998). Instead, the fictional *BTL* news stories are an example of specific and purposeful deception: a deception game. The radio callers are faced with a problem they must solve, and they know from the outset of the *BTL* game that two of the panelists are presenting fictional stories. Accordingly, the invented *BTL* news stories are also clearly creative products. But, just like deception, these news stories are not examples of everyday, creative language (Gerrig and Gibbs, 1988). Instead, these news stories are more appropriately defined as expert creative products carefully planned in advance. Moreover, all of the *BTL* stories violate a listener's assumptions regarding the plausiblity of the events described as well as reasonable behavior associated with the events. This violation is typically benign enough in that it can be reconciled within a hearer's understanding of the world, and thus all of the *BTL* stories, both fictional and truthful, contain some elements of humor when analysed via theoretical models of humor, such as Benign Violation Theory (Warren and McGraw, 2016). This is because the unbelievable nature of these stories creates some form of incongruity between expectations and reality, the resolution of which is thought to be an essential component of theories of humor comprehension from almost all theoretical perspectives, including the General Theory of Verbal Humor (GTVH) and other models of incongruity resolution (Attardo and Raskin, 1991; Forabosco, 1992, 2008; Ritchie, 2009). Thus, it is perhaps best to categorize the dishonest *BTL* narratives as a unique form of humorous deception.

With this distinction in mind, the fictional *BTL* stories must work within the textual and communicative contraints of the truthful *BTL* stories, which are all examples of humorous narratives because they describe unbelievable events involving both fictional and non-fictional characters for the purpose of entertainment (Chłopicki, 2017). One only need refer to internet curations of strange yet real news, such as *Yahoo!*'s Odd News section, Reddit's r/nottheonion community, or internet memes associated with Florida Man to see similar examples. Real news can be strange, and thus even the most outrageous sounding *BTL* stories could possibly be true. Therefore, the relative strangeness of an invented *BTL* news story is an essential element and one that cannot be used to reliably identify deception from truth. At the same time, it is the strength of this violation that is played with in the fictional *BTL* stories - going too far towards the absurd (or the normal) may give away the fictional nature of the deceptive *BTL* narratives. While the radio callers likely attend to this fine line, this may also affect the linguistic choices of the panelists as they strike a balance among humor and deception. Accordingly, just as prior studies of deceptive and humorous language have suggested, it may be the case that more subtle differences in linguistic features can be used to dis-

tinguish between deceptive and truthful *BTL* stories. The next section reviews related research into the linguistic features of deceptive and humorous language.

3. Linguistic Features of Humor and Deception: Prior Studies

With advances in applied Natural Language Processing (NLP) technology, a wide range of linguistic features can be modelled quantitatively. These include simple measures such as counting the number of words in a document but also include more sophisticated features, such as the sentiment associated with a text (e.g., perceptions of negativity or positivity) or the complexity of a text's syntactic structures. Among the many applications of this approach, one method is to use linguistic features to classify documents associated with different genres, registers, or communicative functions. It is through this approach that some studies have already provided insight into the linguistic features associated with humor and deception.

3.1 Linguistic Features and Deception

Several studies have used quantitative linguistic measures to classify deceptive from truthful speech. For instance, Hancock et al. (2008) recruited 35 pairs of subjects to communicate with one another using a computer messaging program. One participant in each pair was randomly asked to lie about two of the five topics discussed. Hancock et al. (2008) used an automatic text analysis program to explore differences in quantifiable linguistic features between the deceptive and non-deceptive conversational turns. The program, Linguistic Inquiry and Word Count (LIWC), reports on psychological and emotional information associated with specific words used in a text, as well as basic lexical and syntactic information (Pennebaker et al., 2001). The Hancock et al. (2008) results indicated significant differences between deceptive and non-deceptive language for several linguistic features. Specifically, deceptive language contained a greater number of words, a greater number of third-person pronouns, and more words related to senses (e.g., touch, feel) when compared to truthful language. The authors interpreted these findings to suggest deceptive language includes more detail and shifts the focus of the conversation onto the listener.

Motivated by these findings, Duran et al. (2010) reanalyzed the Hancock et al. (2008) data using a different text analysis program, Coh-Metrix (Graesser et al., 2004). Based on the findings of Hancock et al. (2008) and their own theoretical position, Duran et al. (2010) focused their analysis on six categories of linguistic features hypothesized to predict deceptive language. These categories were: total word count (measured as total number of words), immediacy of personal involvement (measured through personal pronouns and hedging), specificity of events (measured through temporal words and wh- questions), accessibility of meaning (measured through word concreteness, meaningfulness, and familiarity), complexity of language (measured through syntactic complexity and use of negation), and cohesion (measured through argument overlap and lexical similarity).

Using these categories, Duran et al. (2010) constructed a linguistic profile of deceptive language based on linguistic features which significantly differed between the deceptive and truthful language. This profile suggested that while deceptive language used relatively fewer words than non-deceptive language, (different from Hancock et al. 2008 based on how the two programs counted words), deceptive language was also more syntactically complex than truthful language. Semantically, deceptive language employed words with more accessible meanings and did not introduce new information at the same rate as truthful speech. Thus, while not all the categories described by Duran et al. (2010) proved to be important distinguishers of deceptive language, their

study replicated most of the Hancock et al. (2008) results and identified new features related to deception.

A third study employed linguistic indices from both LIWC and Coh-Metrix using a different dataset (Van Swol et al., 2012). In this study, research subjects (in pairs) participated in a game of deception wherein one participant was tasked with allocating a sum of money to the other participant. The total sum to be allocated was only known to the participant allocating the money, and the receiver could choose to reject or accept the offer presented to them by the allocator. As such, the participant allocating the money had the option of informing the receiver of the true amount of money or lying (ostensibly to be able to keep more for themselves). Van Swol et al. (2012) recorded and transcribed the conversations and then coded them for deception. The results for this study separated deceivers into two categories. For participants who simply lied about the total amount of money they were given to allocate, the findings reported a higher frequency of third-person pronouns, swearing, and use of numbers. Participants who lied via omission of truth used fewer words and causatives. As such, the Van Swol et al. (2012) study reported differences for similar linguistic features identified in the Hancock et al. and Duran et al. studies, but also highlighted how different deception strategies or functions were associated with different linguistic features.

3.2 Linguistic Features and Humor

The current body of computational and linguistic research into humor is vast when compared to similar research on deceptive language. Incongruity is a central concept to theoretical understandings of humor (Attardo and Raskin, 1991; Forabosco, 2008; Ritchie, 2009) and explains how humor is created (e.g., incongruity between linguistic form and meaning) and understood (resolution of two competing interpretations in light of context). Computational analyses of jokes, puns, and other forms of humor have worked to identify linguistic features best able to detect structural and semantic incongruity in a humorous text. For example, a computational analysis of humorous one-line jokes found that measures of unique word associations were best able to identify the correct punchlines paired with joke setups (Mihalcea et al., 2010).

Another study investigated humor as it occurred naturally in a corpus of academic essays (Skalicky et al., 2016). In this study, Skalicky et al. (2016) identified four linguistic features associated with human perceptions of humor. Together, these features suggested humorous academic essays to be more descriptive (via a greater incidence of adjectives, adverbs, and adjective predicates), more sophisticated (based on words with lower frequency), and less cohesive (based on lower sentence-to-sentence cohesion) when compared to less humorous academic essays. In this data, the lack of cohesion associated with humorous essays may have worked to signal incongruity associated with humor, differentiate the humorous paragraphs from the academic paragraphs in the essays, or a combination of both.

Finally, several studies have modelled the linguistic features of humorous satirical news texts from websites such as *The Onion* or satirical product reviews taken from Amazon.com (Burfoot and Baldwin, 2009; Mihalcea and Pulman, 2007; Reyes and Rosso, 2011; Skalicky and Crossley, 2015). Differences among these studies in terms of the texts used, linguistic features selected, and statistical models employed have led to a wide range of results with some observable trends. For instance, lexical properties of satirical texts suggest they are generally more negative (Mihalcea and Pulman, 2007; Skalicky and Crossley, 2015) and human-centered (Mihalcea and Pulman, 2007). Moreover, satirical news texts contain entities that are typically not discussed together in real news articles

(Burfoot and Baldwin, 2009), whereas satirical product reviews construct descriptive, present-tense (yet fictional) narratives (Reyes and Rosso, 2011; Skalicky and Crossley, 2015).

As a whole, these studies showcase the manner in which quantifiable linguistic features can be used to detect and describe both humorous and deceptive language. The question remains, however, as to whether similar linguistic strategies of deception or humor are employed in communicative contexts in which deceptive intention is known in advance, such as during games of deception. In these situations, the goal of deception is to convince the hearer that something false is actually true, whereas the purpose of most conversational deception is to prevent something true from being discovered (Gupta and Ortony, 2018). As such, deceivers may rely on specific linguistic strategies when cloaking fiction as truth, and these features may or may nor overlap with the findings from prior linguistic studies of deception. At the same time, the fictional *BTL* stories are designed to elicit humor and therefore may also contain linguistic features designed to present some form of humorous incongruity to the listener. This incongruity may serve two purposes: to signal humor as well as to position the fictional *BTL* story in the "unbelievable-yet-true" genre of news stories.

4. Current Study

The goal of the current study is to consider how a different conversational context may influence the linguistic features associated with deception when it is also being used for the purpose of humor. Specifically, we investigate humorous deception as it occurs in a context in which all of the interlocutors are aware that deception is present: a humorous quiz segment named *Bluff the Listener* from the popular *Wait Wait... Don't Tell Me!* radio show.

The following research questions guide this study.

1. Are there linguistic features which can distinguish between deceptive and non-deceptive language as it occurs in the *Wait Wait... Don't Tell Me!* radio show?

2. If so, how do these results compare to prior research of humorous and deceptive language?

4.1 Data Collection

A corpus of *BTL* stories was collected by manually downloading provided transcripts from NewsBank (www.newsbank.com), a curated repository of current and archived media. The corpus comprised 251 different *WWDTM* episodes broadcast over a ten-year period from 2010 to 2019 (number of episodes per year M = 25.1, SD = 5.44), for a total of 753 different *BTL* stories (502 deceptive stories and 251 truthful stories). The mean number of words for deceptive texts was 206 (SD = 49.3), and the mean number of words for truthful texts was 184 (SD = 48.7). In the corpus, each episode is accompanied with the following information: the date, the text of the three *BTL* stories presented in the episode, the panelist associated with each story, and whether the caller correctly identified the truthful story. There was a total of 50 different panelists in this data, each contributing a different number of *BTL* stories (Minimum = 1, M = 15.1, SD = 19.42, Maximum = 75). The variation in contributions is an effect of some panelists being regulars on the show, such as Paula Poundstone (n = 75) or Mo Rocca (n = 50), whereas other panelists appeared only once or twice in this data. Indeed, half of the panelists contributed two or fewer *BTL* stories (n = 25). Because the different panelists hailed from different vocations, had differing levels of experience crafting *BTL* stories, and had different numbers of truthful or deceptive stories, it is important to capture this variation in our analysis.

Year of Broadcast	Number of BTL Episodes	Caller Accuracy
2010	23	47.80%
2011	27	63.00%
2012	30	60.00%
2013	30	70.00%
2014	27	85.20%
2015	27	55.60%
2016	16	62.50%
2017	27	70.40%
2018	29	79.30%
2019	15	80.00%

Table 1: Description of dataset and caller accuracy.

The radio callers were able to identify the truthful *BTL* stories 67.33% of the time. However, the data suggests this accuracy varied by year. Table 1 displays the number of episodes and overall caller accuracy for each year in the corpus. As can be seen, there is evidence of two positive, linear trends in accuracy; one starting from 2010 and ending in 2014, and a second starting in 2015 and ending in 2019. Because the *BTL* stories were all modeled on real news stories at the time, it may be the case that differences in available topics over the years also influenced the deceptive strategies associated with different *BTL* stories which may have influenced this difference in caller accuracy. Accordingly, the variation in accuracy over time is also important to model.

4.2 Linguistic Features

Linguistic features were collected using a suite of automatic text analysis tools which consolidate a wide range of linguistic features developed for multiple applications. Table 2 summarizes these tools and the primary constructs they measure. More information can be found following the appropriate citations or by visiting www.lingusiticanalysistools.org.

4.2.1 INITIAL FEATURE SELECTION

Our analysis[1] started with the full output of linguistic features collected from the text analysis programs reported in Table 2. We then trimmed the number of indices down to avoid violating statistical assumptions. To do so, we first removed variables with variance close to zero and variables that had high zero counts. We then kept those variables that showed a significant difference between truthful and deceptive stories using simple t-tests in which the alpha value for significance was set at $p < .001$. After pruning the data, 263 indices remained. We further culled this list by removing variables which measured different variations of the same construct. For instance, from SEANCE, if two variables measured a similar construct but one version controlled for negation, we only retained the version which controlled for negation. We also opted for versions of variables which measured all of the words or all of the content words in the text (as opposed to just the function words). After removing these variables, 127 variables remained. We lastly checked these remaining 127 linguistic features for multicollinearity using Pearson correlations. For any two variables with an absolute cor-

1. Data and code for our analysis can be found on osf.io

Program Name	Construct	Reference
Tool for the Automatic Analysis of LExical Sophistication 2.0 (TAALES version 2.8.1)	Lexical sophistication	Kyle, Crossley, and Berger, 2018
Sentiment Analysis and Social Cognition Engine (SEANCE version 1.2.0)	Sentiment	Crossley, Kyle, and McNamara, 2016
Tool for the Automatic Analysis of Cohesion 2.0 (TAACO version 2.0.4)	Cohesion and coherence	Crossley, Kyle, and Dascalu, 2019
Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASC version 1.3.8)	Syntactic sophistication and complexity	Kyle, 2016
Tool for the Automatic Analysis of LExical Diversity (TAALED version 1.3.1)	Lexical diversity	In preparation, see linguisticanalysis-tools.org

Table 2: List of text analysis programs used in current study.

relation higher then .75, we removed the variable with the highest mean absolute correlation with all of the other variables. Through this process we removed an additional 49 linguistic features, bringing the total to 78 remaining linguistic features.

4.2.2 FURTHER FEATURE SELECTION: ELASTIC NET LOGISTIC REGRESSIONS

Although we drastically reduced the initial number of linguistic variables, it was still necessary to further reduce the number of remaining variables to avoid overfitting models. We did so by using supervised machine learning classification techniques. Specifically, we trained 100 versions of a logistic regression model predicting whether a *BTL* story was truthful or deceptive. For each model, we randomly split the entire *BTL* corpus into a training and test set using a 70% training and 30% test split. We trained our models in R using caret with a glmnet elastic model method. This method works to penalize overfitting by regularizing the coefficients of variables which only contribute a small amount of predictive power. Doing so mitigates the effects of overfitting associated with the large number of linguistic features entered into each model. During the training process, each training model was fit based on the results of ten-fold cross-validation with ten repeats, which was done in order to further safeguard the coefficients from overfitting. Finally, because the category membership of deceptive and truthful *BTL* stories was unbalanced (i.e., twice as many deceptive as truthful stories), we specified our models to use down sampling which equalized the number of cases in each model. After training each model, we tested the results based on the model's ability to make predictions for the remaining 30% of the data. The result of this was that each individual *BTL* story in the test data was assigned a percentage chance of how likely that story was to be truthful or deceptive based on the linguistic features included in the elastic net model. Accuracy of the predictions was then assessed by comparing the predicted category to the actual category for the truthful and deceptive stories.

As mentioned above, we repeated this entire training/test process using ten-fold cross-validation 100 times, meaning that we obtained 100 different final models from 100 different random 70/30 splits of the data. Each each of these 100 final models were themselves obtained from a ten-fold cross-validation boostrapping process. Accuracy predicting the test set for each of the 100 final models ranged from a minimum of .573 to a maximum of .716 (M = .651, SD = .029). For each model, we extracted the top 20 predictors based on the strength of the coefficients in the model (using the varImp function in caret). We counted the number of times each linguistic variable was included in the top 20 predictors for the 100 models, resulting in a feature score for each linguistic variable ranging from 0-100. We then chose features which occurred in the models at least 50% of the time (i.e., included in the top 20 variables of a model at least 50 times). This process identified 14 linguistic variables. These variables comprised the set of linguistic indices we then used in our subsequent predictive models, described below[2]. Table 3 presents an overview of these 14 linguistic features, including their definition and average values for the truthful and deceptive *BTL* stories. In order to better represent these differences, Figure 1 visually plots standardized versions of each value using z-scores. For each variable in Figure 1, the mean is set to zero, represented by the dashed line. Bars above the dashed line represent values greater than the mean, and bars below the dashed line represent values less than the mean. Figure 1 thus displays whether the truthful or deceptive *BTL* stories contain higher or lower amounts of each particular linguistic feature. These features can be grouped into several larger categories, described below.

4.2.3 SENTIMENT AND SEMANTIC GROUPINGS

Sentiment measures affective perceptions associated with words, such as valence (positive/negative emotional associations with words). Semantic groupings are clusters of words related to some similar semantic category, such as words related to motivation, cognition, and so on. Five variables from Table 3 are of this category and include Abstract Words, Strength Adjectives, Dominance Ratings: Nouns, Time and Space Words, and Vader Polarity: Adjectives. The Abstract Words, Strength Adjectives, and Time and Space Words all belong to lists of semantic categories originally compiled as part of the General Inquirer database (Stone et al., 1966). Abstract Words are content words which represent abstract concepts, such as *duty* and *truth*. Strength adjectives are adjectives which imply strength, such as *alert* or *muscular*. The Time and Space Words category includes words related to temporal and spatial meaning, including locations (*somewhere*) and locative prepositions (*above*), measurement words (*diameter*), and words of distance and time (*inch*, *soon*). Dominance represents affective perceptions of dominance originally collected as part of the Affective Norms for English Words (ANEW) database (Bradley and Lang, 1999). Perceptions of dominance measure whether a word is associated with something that is in control versus something that is being controlled (*leader* is more dominant than *ache*). VADER is a sentiment analysis framework specifically designed for social media which takes into account variations in punctuation, emoticon use, and other features of shorter texts to provide a state-of-the-art measure of valence (Hutto and Gilbert, 2014). The specific measure in this list is the average VADER sentiment ratings for the adjectives in each *BTL* story.

2. To clarify, the purpose of this portion of the analysis was to explore which set of linguistic features were consistently chosen as predictive of whether a *BTL* story was deceptive or truthful in the current data. Therefore, this part of the analysis was not intended to be confirmatory or related to testing our research questions.

Variable Name	Variable Description	Truth *M*(*SD*)	Deception *M*(*SD*)
Abstract Words	Words representing abstract concepts (higher = more of these words)	0.032 (0.016)	0.036 (0.016)
CW Repetition: Sentence	Average number of a times any content word repeats in the next two sentences (higher = more repetition)	0.121 (0.055)	0.111 (0.051)
VAC Strength (SD)	Standard deviation of average strength between constructions and verbs (higher = stronger strength)	0.090 (0.064)	0.106 (0.076)
Dominance Nouns	Average dominance values for all nouns (higher = more dominant)	5.300 (0.498)	5.420 (0.355)
Sentence Similarity (LDA)	Average semantic similarity between adjacent sentences using LDA (higher = more similarity)	0.943 (0.042)	0.934 (0.050)
Word Similarity (LSA)	Average word similarity for all words in a text using LSA (higher = more word similarity)	0.166 (0.021)	0.170 (0.020)
Mean Length CW TTR (MTLD)	Average span length for content words with an average TTR of .720 (higher = lower average lexical diversity)	87.200 (24.200)	97.600 (23.800)
Noun Complexity (SD)	Standard deviation of number of dependents per noun subject (higher = more variation in number of dependents)	0.992 (0.304)	1.080 (0.344)
Positive Causal Conn.	Causal connectives with positive sentiment (higher = more of these words)	0.017 (0.011)	0.014 (0.009)
CW Repetition: Text	Average ratio of any one content word to all words in a text (higher = more content word repetition)	0.174 (0.048)	0.181 (0.047)
Type-Token Ratio	Average variety of lexical items (higher = more variety)	0.624 (0.060)	0.612 (0.058)
Strength Adjectives	Words representing strength (higher = more of these words)	0.242 (0.133)	0.271 (0.128)
Time and Space Words	Words with spatial or temporal meaning (higher = more of these words)	0.070 (0.025)	0.075 (0.023)
Vader Polarity: Adjectives	Average valence of adjectives (higher = more positive adjectives)	0.194 (0.564)	0.382 (0.533)

CW = Content Words, VAC = Verb Argument Construction, SD = Standard Deviation, LDA = Latent Dirichlet Allocation, LSA = Latent Semantic Analysis, TTR = Type-Token Ratio, MTLD = Measure of Textual Lexical Diversity. Variable names as reported by the text analysis tools are included in Appendix A

Table 3: Fourteen linguistic variables included in the top 20 predictors by at least 50% of the elastic net logistic regression models.

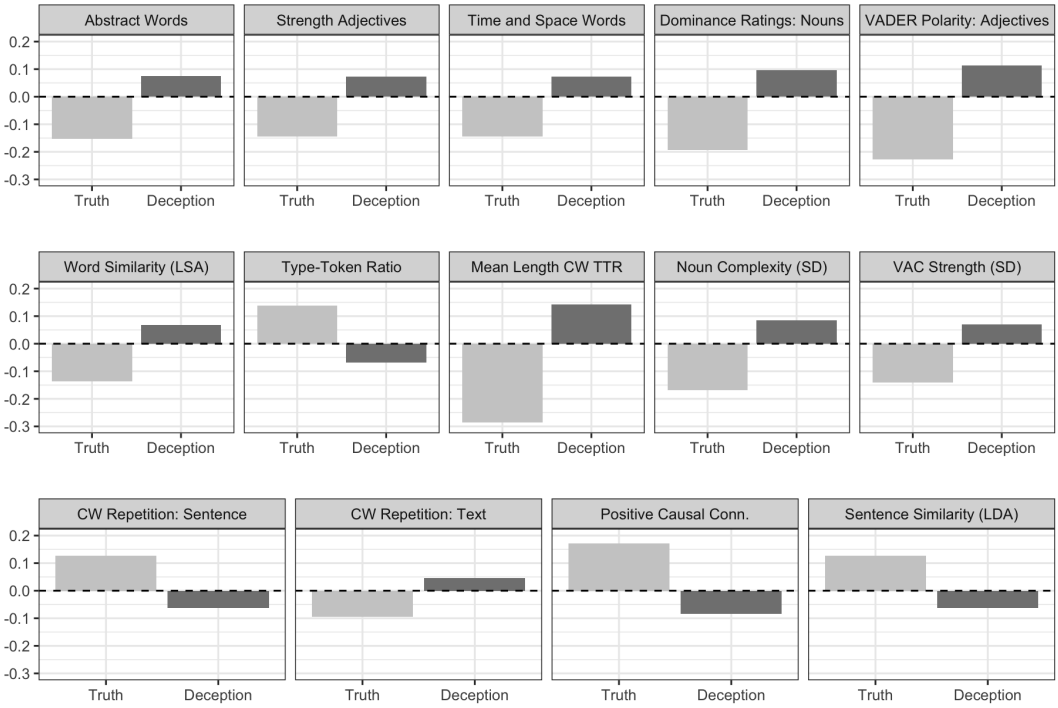

Figure 1: Standardized average values for the 14 linguistic variables which appeared in at least 50 of the 100 elastic net logistic regression models predicting truthful and deceptive *Bluff the Listener* stories. Row 1 = Sentiment variables, Row 2 = Complexity variables, Row 3 = Cohesion and coherence variables. Dashed lines set at zero represent mean value for all texts; shaded bars represent change for each individual variable in truthful or deceptive texts. CW = Content Words, VAC = Verb Argument Construction, SD = Standard Deviation, LDA = Latent Dirichlet Allocation, LSA = Latent Semantic Analysis.

4.2.4 LEXICAL AND SYNTACTIC COMPLEXITY

Lexical and syntactic complexity measure the overall sophistication of the words and grammatical constructions employed in a text. Five variables from Table 3 were related to these constructs: Word Similarity via Latent Semantic Analysis (LSA), Type-Token Ratio (TTR), Mean Length of Content Word Type-Token Ratio (MTLD), Noun Complexity via Standard Deviation (SD), and Verb Argument Construction (VAC) Strength via Standard Deviation (SD). The Word Similarity (LSA) feature measures the distributional similarity of words across texts. The LSA values used here are pretained values from the Touchstone Applied Science Associates Inc. (TASA) corpus. The TASA corpus contains over 37,000 texts representing a variety of different genres (Günther et al., 2015). Thus, our Word Similarity (LSA) feature measures the distributional similarity of words in the *BTL* stories when compared to the general language corpus (i.e., TASA). In this manner, this feature represents the lexical distinctiveness of vocabulary used in a text because texts with greater distributional similarity will contain words that can be used in a greater number of contexts. Higher average LSA values thus reflect lower lexical distinctiveness and less sophisticated vocabulary. The next two variables are a measure of how many different words types are used in a text. Type-Token Ratio (TTR) is the simplest version and is the number of unique word types divided by the

total number of word tokens in a text. A higher TTR means a greater amount of different word types exist. The Mean Length Content Word TTR (MTLD) is a more precise measure of TTR which calculates the average span length of Content Words (CW) which maintain a TTR above .720 (McCarthy and Jarvis, 2010). In this manner, TTR captures global word use in a text, whereas the MTLD value captures consistency of TTR throughout a text. Noun Complexity (SD) is the standard deviation of the average number of dependents attached to noun phrases in a text (e.g., each direct or indirect object attached to a noun phrase would count as a dependent). Because this measure is the standard deviation, it captures the rate of variation for this measure. Finally, VAC Strength (SD) is a measure of how strongly verbs and their syntactic dependents (e.g., adverbs or nouns attached to a main verb) are associated based on their frequency of occurrence in a regular English usage. A higher value would suggest verbs are being used with typical or more commonly used syntactic dependents. In the same manner to Noun Complexity (SD), VAC Strength (SD) measures the standard deviation and thus the rate of variation for this measure.

4.2.5 COHESION AND COHERENCE

Cohesion and coherence measure the similarity of words across sentences and paragraphs in a text as well as how well ideas in a text are connected. Four variables from Table 3 were related to these constructs: Content Word (CW) Repetition: Sentence, Content Word (CW) Repetition: Text, Positive Causal Connectives, and Sentence Similarity via Latent Dirichelet Allocation (LDA). The first two measure the number of repeated content words (i.e., nouns, verbs, adjectives) between adjacent sentences as well as in the overall text. Positive Causal Connectives measures the frequency of occurrence for words which create positive causal links between sentences, such as *because* and *moreover*. The Sentence Similarity (LDA) feature uses Latent Dirichlet Allocation to calculate the probability occurrence of latent topics within and across texts. In this case, LDA is used to calculate the similarity of sentence topics in a particular text.

4.3 Predictive Model: Generalized Linear Mixed Effects Model

The distribution of the 14 features as they appear in Figure 1 suggests descriptive differences between the truthful and deceptive *BTL* stories. However, it is important to model these potential differences in light of the variation that may be associated with differences in style associated with particular panelists as well as the available pool of news topics each year. To do so, we fit a generalized linear mixed effects model with text category as the outcome variable (truth versus deception with truth as baseline), the 14 linguistic features in Table 3 as fixed effects, and panelist and year as random effects[3]. We used an automatic backfitting algorithm to assess the significance of each fixed effect via model comparisons using relative log-likelihood comparisons with the Akaike's Information Criterion (AIC) as a criterion of model fit (Tremblay and Ransijn, 2015).

The resulting model parameters are reported in Table 4. Three of the 14 features in Table 3 were removed during the model backfitting procedure (Abstract Words, Type-Token Ratio, and VAC Strength SD), suggesting that they did not contribute a significant amount of additional explanatory power in light of the remaining variables. Although the random effects structure captured the predicted variance associated with different panelists, the random effect of year explained close to zero variance in the model (resulting in a singular fit) and was thus removed. The marginal R^2 (fixed effects only) was .182 and the conditional R^2 (fixed and random effects combined) was .223 (using

3. glmer(truth or deception ~ linguistic_feature1 + ... + linguistic_feature14 + (panelist) + (year))

Random Effects				**Model Effect Sizes**			
	Variance	SD		Marginal R^2		Conditional R^2	
Panelist	0.239	0.489		.182		.223	
Fixed Effects							
	Estimate	SE	z	p	OR	5%	95%
(Intercept)	0.894	0.139	6.452	$< .001$	2.445	1.947	3.071
Sentiment and Semantic Groupings							
Strength Adjectives	0.206	0.089	2.314	0.021	1.229	1.062	1.424
Time and Space Words	0.250	0.090	2.768	0.006	1.284	1.107	1.490
Dominance: Nouns	0.293	0.095	3.090	0.002	1.341	1.147	1.567
VADER: Adjectives	0.349	0.088	3.975	$< .001$	1.417	1.227	1.637
Cohesion and Coherence							
CW Repetition: Sentence	-0.302	0.118	-2.564	0.010	1.352	0.609	0.897
CW Repetition: Text	0.393	0.117	3.360	0.001	1.482	1.222	1.796
Positive Causal Conn.	-0.269	0.089	-3.022	0.003	1.309	0.660	0.885
Sentence Similarity (LDA)	-0.247	0.095	-2.590	0.010	1.280	0.668	0.914
Lexical and Syntactic Complexity							
Word Similarity (LSA)	0.230	0.094	2.458	0.014	1.259	1.079	1.469
Mean Length TTR (CW)	0.476	0.108	4.389	$< .001$	1.609	1.347	1.924
Noun Complexity (SD)	0.251	0.093	2.701	0.007	1.285	1.103	1.497
DV baseline = truth. OR = odds ratio. For ease of interpretation, the OR for terms with negative estimates were transformed to positive odds ratios using 1/OR. Refer to Table 3 for variable descriptions.							

Table 4: GLMER results predicting truthful and deceptive *Bluff the Listener* stories

the delta method), which indicates the linguistic variables in this model were able to account for approximately 18% of the variation in text type.

5. Discussion

The goal of this study was to investigate the linguistic features of creative and humorous deception as it occurs in a unique conversational context wherein all interlocutors are aware that deception is present. To do so, we collected a corpus of truthful and fictional news stories used during a recurring segment of a radio quiz show named *Bluff the Listener* (*BTL*), part of the popular American radio show *Wait Wait... Don't Tell Me!*. We gathered quantitative measures for a variety of linguistic features related to lexical and syntactic sophistication, sentiment, cohesion, and coherence for the deceptive and truthful *BTL* stories using a suite of automatic text analysis tools. We first identified linguistic features consistently chosen as predictive of text category (deceptive versus truthful) using 100 penalized logistic regressions with cross-validation and down sampling. This process identified 14 linguistic variables as significant predictors of text type, which we then fit into a generalized linear mixed effects model which also took into account variance associated with different authors of the *BTL* news stories as well as the time the story was produced.

Our first research question asked whether linguistic features could be used to distinguish between deceptive and non-deceptive language as it occurs in the *BTL* segment of the *WWDTM* radio show. Our initial results identified 11 linguistic variables as significant predictors of deceptive texts which accounted for approximately 18% of the variance in our data set. These 11 features represent three general categories of linguistic properties: sentiment and semantic groupings, cohesion and coherence, and lexical and syntactic complexity. Our second research question asked how the results obtained in answering our first research question compare to similar prior studies of the linguistic features of humorous and deceptive language. Below, we provide a discussion of our results in light of these two research questions for each of the three broader categories of linguistic properties described above.

5.1 Sentiment and Semantic Groupings

The results for the four significant linguistic variables in this category suggest specific lexical differences between the deceptive and truthful *BTL* stories. Texts containing a higher number of adjectives related to perceptions of strength as well as a higher number of time and space words were more likely to be deceptive rather than truthful. Adjectives that belong to the strength semantic grouping include words related to physical qualities (*athletic*, *large*, *healthy*), certainty (*undeniable*, *last*, *most*), behavior (*nonchalant*, *steady*), performance (*perfect*, *proficient*), and other related terms. Words that belong to the time and space semantic grouping represent temporal and spatial relations. These include prepositions, adjectives, and verbs related to physical location (*around*, *southern*, *surround*) as well as nouns related to distance and time (*kilometer*, *era*). The other two linguistic features in this category were related to sentiment or affect. Deceptive texts were associated with language that included higher average dominance ratings for nouns (*crown* is more dominant than *hostage*) as well as higher VADER scores for adjectives (meaning more positive adjectives).

As a whole, these features coalesce to suggest the vocabulary of deceptive *BTL* texts is marked by confident and positive descriptions of entities or actions during specific times and/or in specific locations. This may suggest that deceptive *BTL* stories are therefore more specific or detailed in some regards when compared to the truthful *BTL* news stories. Specificity of language was a feature previously investigated by both Hancock et al. (2008) as well as Duran et al. (2010) and was operationalized as a measure of temporal and spatial words and the number of wh-questions produced. However, specificity was predicted to be lower for deceptive language as a strategy to obscure falsified information with little to no veracity, consistent with prior findings suggesting that liars tend to use language which includes fewer details (DePaulo et al., 2003). The reverse trend is seen here in the current data and may be reflective of the very different communicative context associated with the *BTL* quiz game. Indeed, because the authors of the deceptive *BTL* news stories are tasked with making the fictional seem plausible, it may be the case that including more specific information and vivid detail lends an air of authenticity to the fictional stories. In this manner, these confident, accurate description mirror linguistic features of humor identified in academic essays as well as satirical product reviews (Skalicky and Crossley, 2015; Skalicky et al., 2016), both of which were found to employ a higher degree of description and certainty. These features may therefore align closer with the humorous than deceptive aspect of the *BTL* stories.

5.2 Lexical and Syntactic Complexity

Complexity was also investigated in prior research of both humor and deception. In terms of deception, Duran et al. (2010) included two measures of syntactic complexity: the use of negative connectors and the mean number of words which appear before a verb in each clause. Deceptive language contained significantly more of the second of these features, which Duran et al. (2010) interpreted to represent the need for deceptive speakers to spend more time formulating their lies and deception on-the-fly (a stalling strategy). Much like the previous category, the findings of the current study are different.

The three linguistic features in the current study related to lexical and syntactic complexity were the Word Similarity via Latent Semantic Analysis (LSA), mean length of high type-token ratio for content words (MTLD), and standard deviation of noun complexity measures. Deceptive *BTL* stories were associated with greater amounts of all three of these features. In terms of word similarity via LSA, the deceptive stories were marked by words which were more strongly related to other words. As such, the vocabulary of the deceptive texts was less sophisticated and less contextually diverse when compared to the non-deceptive texts. Additionally, the sentences in the deceptive stories used a greater variety of words for longer spans before repeating words (supported by the lower incidence of content word overlap discussed below), which aligns with the descriptive findings demonstrating the deceptive texts were on average longer than the truthful texts. As such, this form of complexity may in turn reflect a level of exaggerated or fabricated complexity on the part of the deceptive *BTL* news stories. Finally, the measure of noun complexity was the standard deviation of the average number of dependents attached to a noun phrase. This means that the noun phrases (and, by extension, embedded phrases, clauses, and whole sentences) in the deceptive *BTL* stories had much greater variation in structure, length, and word types when compared to the truthful *BTL* stories. This does not suggest that the deceptive *BTL* stories were necessarily more or less syntactically complex than the truthful stories, but rather that that deceptive stories were less consistent in their syntactic choices. All together, these features suggest that the deceptive *BTL* stories are on the whole more descriptive and varied in their sentence complexity, which likely reflects a combined strategy of deception (exaggerated detail) and humor (incongruity).

5.3 Cohesion and Coherence

The direction of the coefficients for each of the four individual variables in this category align to suggest two key differences between the deceptive and truthful *BTL* stories for cohesion and coherence. First, the deceptive texts were associated with lower sentence-to-sentence cohesion when compared to the truthful texts. For instance, the measures of lexical overlap and semantic cohesion at the sentence level both suggest the truthful texts more consistently used the same words (Content Word Repetition: Sentence) and repeated similar ideas (Sentence Similarity via Latent Dirichelet Allocation) across adjacent sentences. The truthful texts were also associated with a greater number of positive causal connectives, which link sentences and ideas using words like *arise* and *moreover* to signpost additional positive information linked to any particular idea (as opposed to negative causation words such as *however*). This suggests truthful *BTL* texts were marked by greater overall coherence of ideas because the sentences were more cohesive and more explicitly connected.

The second key difference between deceptive and truthful *BTL* texts for cohesion and coherence was demonstrated by the measures of lexical repetition. Namely, deceptive texts were marked by

more repetition of content words in a text overall (Content Word Repetition: Text) as contrasted with repetition in adjacent sentences. Thus, the deceptive texts were more cohesive at a general, lexical level, but had lower overall coherence among ideas when compared to the truthful texts. This may reflect the manner in which topics and entities are discussed in the two different texts types. For the truthful *BTL* stories, ideas may be typically presented in a progressive manner with more explicit and coherent links among ideas, as befitting a news report. The fictional narratives that comprise deceptive *BTL* stories, however, may lack this level of coherence because they rely more heavily on invented situations.

Cohesion and coherence have been investigated in prior similar investigations of deceptive texts. For instance, in their results, Duran et al. (2010) found deceptive language to be more redundant than truthful language, and therefore more cohesive, which is opposite the findings reported in the current data. Duran et al. (2010) suggested higher redundancy associated with deception in their study may have been reflective of a strategy to focus on a small number of ideas as to avoid introducing information which may give away the deceptive intent. At the same time, they argued that higher cohesion may have also reflected the relative lack of links to memorized, prior experiences, and naturally the deceptive conversational turns were forced to rely on the smaller number of ideas conjured in the moment.

Although not realized in their results, Duran et al. (2010) had also predicted that the same lack of memory and experience associated with fictional and deceptive events may lead to less cohesion among ideas in deceptive speech. This hypothesis may explain the results for the *BTL* stories, where the truthful *BTL* stories were more cohesive than the deceptive stories. The truthful *BTL* stories are all adapted from news reports based on actual events with real entities and locations but are also within the genre of strange or unbelievable news. Because the deceptive *BTL* stories also must tread the line between the merely absurd and the unbelievable, it may be the case that the introduction of less cohesive ideas and words served to enhance the appropriateness of the deceptive *BTL* news stories for this particular genre (i.e., to create a sense of strange-yet-real news). In other words, the introduction of lower cohesion may actually serve to better align the deceptive *BTL* stories with the unique genre and communicative context of the *BTL* quiz show, but still lacks the cohesion a story based on genuine events can contain. At the same time, this lack of coherence may reflect incongruity associated with humor in the fictional *BTL* texts.

5.4 Summary and Implications

A bird's eye view of the results paints the deceptive *BTL* stories as fictional narratives which are overly descriptive but also lack higher-level cohesion and coherence. In this manner, the deceptive *BTL* stories are a messy mirror of the truthful *BTL* stories and may represent the individual chaos injected into these stories by each of the different panelist authors for the dual purpose of humor and deception. The panelists are simultaneously competing against each other to earn points while also entertaining the *WWDTM* audience. Thus, there may exist a tension between being purely deceptive and the desire to present a fictional *BTL* story which is humorous and entertaining, and this likely translates into a specific instantiation of deceptive humor seen only in this and similar genres.

As mentioned in section 2.1.2, both the fictional and the truthful *BTL* stories engage in some form of incongruity or violation of expectations which can result in humor (Warren and McGraw, 2016; Attardo and Raskin, 1991). The strength of this violation must be carefully attended to by the authors of the deceptive *BTL* stories. We suggest that the increased lexical descriptiveness and lack

of higher-level cohesion and coherence in the fictional *BTL* stories may be a linguistic manifestation of the forced, fictional incongruity required in order to meet the demands of this unique genre. In order to create some level of incongruity or violation expected in any *BTL* news story, an invented situation must be constructed and thus the tendency for overly descriptive language in the fictional *BTL* narratives might help to construct this reality for both the audience as well as the panelist. And, as mentioned above, the lack of coherence and cohesion may have been an unconscious linguistic decision to inject a certain level of incongruity that the truthful *BTL* stories naturally possessed. However, our results suggest that this invented, forced incongruity cannot fully replicate what is found in the truthful *BTL* stories, and perhaps successful radio callers are able to attend to these subtle linguistic differences. As the saying goes, truth is stranger than fiction.

6. Conclusion and Future Directions

In this study we explored which linguistic features were representative of deceptive and truthful news stories in a specific genre and communicative context: a radio quiz game show where radio callers must pick a truthful story from among three possibilities. We opted for this approach because it was difficult to make theoretical predictions from prior research into deceptive language due to the stark differences in genre and context. Our findings support this notion, in that the features predictive of deceptive language in this study did not reflect findings from prior research attempting to catalogue the linguistic features of deception operationalized as lying and/or truth avoidance. However, we believe this is not a contradiction but rather a reflection of how deceptive language adapts to local communicative situation (as does any language use). In other words, because the deceptive *BTL* stories are a very different type of deception, it is not surprising that linguistic features were used differently, likely reflecting the overt humor associated with this form of deception. There are big differences between spontaneous and prepared lies. The underlying cognitive constraints involved will differ, and thus so will their manifestations in language. As such there may be no generic suite of linguistic features related to deception in general, and we believe our current approach provides a roadmap for how linguistic methods can be applied to different examples of deception detection in future studies of this nature.

A natural next step for an analysis of this nature would be to model caller accuracy as a variable to measure which of the deceptive stories were more likely to be chosen (incorrectly) as truthful. While our data did contain caller accuracy, we were unable to verify which callers were accurate based on their knowledge of the news cycle (e.g., knowing the truthful story because they had encountered it in the news already or using the internet during the call to check the stories) and which were truly naïve participants. One potential solution to this issue would be to qualitatively analyze the responses by the radio caller for hints as to the strategies used to determine the truthful stories. A further possibility would be to present these stories in a laboratory setting to research participants.

Another approach for future work in this area would be to consider additional types of humor which rely on varying levels of deception to create or amplify a humorous effect. One potential source of data for such an analysis may be found in the speech of stand-up comedians and other explicit comedic contexts. For example, the late Mitch Hedberg routinely employed a strategy based on linguistic subversion, wherein expectations with specific phrases and collocations were subverted (e.g., “I used to do drugs. I still do, but I used to, too.”). Other examples can be found in jokes which initially retreat from their punchline using phrases such as “just kidding” before

then doubling down on the initial joke (see Skalicky et al. 2015 for a more detailed discussion of these jokes). It could be argued that in both of these cases the audience is partially deceived, and this deception is crucial for incongruity resolution and ultimately humor. Because the incongruity in these forms of humor is tied more strongly to violation of expectations at the linguistic level, it would be fruitful to investigate whether these differences are realized in quantifiable linguistic features and how they may differ from the features identified in the current study.

A final limitation of our study was the relatively small amount of variance accounted for by the linguistic features used here (approximately 18%). This suggests that there are likely other factors characterizing the deceptive stories which may include linguistic and discourse features not captured in the current analysis as well as features of the individual panelists. Our random effects structure suggested the panelists explained an additional 4% of the variance in text category, but future studies might model other aspects of the panelists, such as their vocation, experience on the show, and more. In tandem with the linguistic feature selection process we have described in the current study, this type of additional information offers improvements for any future analyses attempting to distinguish truth from fiction.

References

Salvatore Attardo and Victor Raskin. Script theory revis(it)ed: Joke similarity and joke representation model. *Humor-International Journal of Humor Research*, 4(3–4):293–348, 1991.

Margaret M. Bradley and Peter J. Lang. Affective norms for English words (ANEW): Instruction manual and affective ratings. *The Center for Research in Psychophysiology, University of Florida*, 1999.

Clint Burfoot and Timothy Baldwin. Automatic satire detection: Are you having a laugh? In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, page 161–164. Association for Computational Linguistics, 2009.

Władysław Chłopicki. Humor and narrative. In Salvatore Attardo, editor, *The Routledge Handbook of Language and Humor*, pages 143–157. Routledge, 2017.

Bella M. DePaulo and Deborah A. Kashy. Everyday lies in close and casual relationships. *Journal of Personality and Social Psychology*, 74(1):63–79, 1998.

Bella M. DePaulo, James J. Lindsay, Brian E. Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. Cues to deception. *Psychological Bulletin*, 129(1):74–118, 2003. doi: 10.1037/0033-2909.129.1.74.

Nicholas D. Duran. Expanding a catalogue of deceptive linguistic features with NLP technologies. In *Proceedings of the 22nd AAAI Conference on Artificial Intelligence*, pages 243–248, Sanibel, FL, USA, 2009. Association for the Advancement of Artificial Intelligence.

Nicholas D. Duran and Riccardo Fusaroli. Conversing with a devil's advocate: Interpersonal coordination in deception and disagreement. *PLOS ONE*, 12(6):1–25, 2017. doi: 10.1371/journal.pone.0178140.

Nicholas D. Duran, Charles Hall, Philip M. Mccarthy, and Danielle S. Mcnamara. The linguistic correlates of conversational deception: Comparing natural language processing technologies. *Applied Psycholinguistics*, 31(3):439–462, 2010. doi: 10.1017/S0142716410000068.

Marta Dynel. A Web of Deceit: A Neo-Gricean view on types of verbal deception. *International Review of Pragmatics*, 3(2):139–167, 2011. doi: 10.1163/187731011X597497.

Giovannantonio Forabosco. Cognitive aspects of the humor process: The concept of incongruity. *Humor*, 5(1/2):45–68, 1992.

Giovannantonio Forabosco. Is the concept of incongruity still a useful construct for the advancement of humor research? *Lodz Papers in Pragmatics*, 4(1):45–62, 2008. doi: 10.2478/v10016-008-0003-5.

Richard J. Gerrig and Raymond W. Gibbs. Beyond the lexicon: Creativity in language production. *Metaphor and Symbol*, 3(3):1–19, 1988.

Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2):193–202, 2004.

Swati Gupta and Andrew Ortony. Lying and Deception. In Jörg Meibauer, editor, *The Oxford Handbook of Lying*, pages 148–169. Oxford University Press, 2018. doi: 10.1093/oxfordhb/9780198736578.013.11.

Fritz Günther, Carolin Dudschig, and Barbara Kaup. LSAfun - An R package for computations based on Latent Semantic Analysis. *Behavior Research Methods*, 47(4):930–944, 2015. doi: 10.3758/s13428-014-0529-0.

Jeffrey T. Hancock, Lauren E. Curry, Saurabh Goorha, and Michael Woodworth. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, 45(1):1–23, 2008. doi: 10.1080/01638530701739181.

C. J. Hutto and Eric Gilbert. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*, page 216–225, 2014.

Hansika Kapoor and Azizuddin Khan. Deceptively yours: Valence-based creativity and deception. *Thinking Skills and Creativity*, 23:199–206, 2017. doi: 10.1016/j.tsc.2016.12.006.

Stephan Ludwig, Tom van Laer, Ko de Ruyter, and Mike Friedman. Untangling a web of lies: Exploring automated detection of deception in computer-mediated communication. *Journal of Management Information Systems*, 33(2):511–541, 2016. doi: 10.1080/07421222.2016.1205927.

Philip M. McCarthy and Scott Jarvis. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392, 2010. doi: 10.3758/BRM.42.2.381.

Jörg Meibauer. The linguistics of lying. *Annual Review of Linguistics*, 4(1):357–375, 2018. doi: 10.1146/annurev-linguistics-011817-045634.

Rada Mihalcea and Stephen Pulman. Characterizing humour: An exploration of features in humorous texts. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, page 337–347. Springer, 2007.

Rada Mihalcea, Carlo Strapparava, and Stephen Pulman. Computational models for incongruity detection in humour. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, page 364–374. Springer, 2010.

NPR. Bluff The Listener, January 2020. URL `https://www.npr.org/2020/01/11/795534428/bluff-the-listener`.

James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: LIWC 2001, 2001.

Antonio Reyes and Paolo Rosso. Mining subjective knowledge from customer reviews: A specific case of irony detection. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, page 118–124. Association for Computational Linguistics, 2011.

Graeme Ritchie. Variants of incongruity resolution. *Journal of Literary Theory*, 3(2):313–332, 2009. doi: 10.1515/JLT.2009.017.

Stephen Skalicky and S. A. Crossley. A statistical analysis of satirical Amazon.com product reviews. *The European Journal of Humour Research*, 2(3):66–85, 2015.

Stephen Skalicky, Cynthia M. Berger, and Nancy Bell. The functions of "just kidding" in American English. *Journal of Pragmatics*, 85:18–31, 2015.

Stephen Skalicky, C. M. Berger, S. A. Crossley, and Danielle S. McNamara. Linguistic features of humor in academic writing. *Advances in Language and Literary Studies*, 7(3):248–259, 2016.

Philip J. Stone, Dexter C. Dunphy, and Marshall S. Smith. *The General Inquirer: A computer approach to content analysis.* MIT Press, 1966.

Antoine Tremblay and Johannes Ransijn. LMERConvenienceFunctions: Model Selection and Post-hoc Analysis for (G)LMER models. *R package version 2.10*, 2015. URL `https://CRAN.R-project.org/package=LMERConvenienceFunctions`.

Lyn M. Van Swol, Michael T. Braun, and Deepak Malhotra. Evidence for the Pinocchio Effect: Linguistic differences between lies, deception by omissions, and truths. *Discourse Processes*, 49 (2):79–106, 2012. doi: 10.1080/0163853X.2011.633331.

C. Warren and A. P. McGraw. Differentiating what is humorous from what is not. *Journal of Personality and Social Psychology*, 110(3):407–430, 2016.

Appendix A. Complete Names of Linguistic Features

Output name from program	Name used in manuscript	Program
Abs_GI_neg_3	Abstract Words	SEANCE
adjacent_overlap_2_cw_sent	CW Repetition: Adjacent Sentences	TAACO
all_av_delta_p_const_cue_stdev	VAC Strength (SD)	TAASC
Dominance_nouns_neg_3	Dominance Nouns	SEANCE
lda_1_all_sent	Sentence Similarity (LDA)	TAACO
lsa_average_all_cosine	Word Similarity (LSA)	TAALES
mtld_ma_wrap_cw	Mean Length TTR (CW)	TAALED
nsubj_NN_stdev	Noun Complexity (SD)	TAASC
positive_causal	Positive Causal Conn.	TAACO
repeated_content_lemmas	CW Repetition: Text	TAACO
simple_ttr_aw	Type-Token Ratio	TAASC
Strong_GI_adjectives_neg_3	Strength Adjectives	SEANCE
Timespc_Lasswell_neg_3	Time and Space Words	SEANCE
vader_compound_adjectives	Vader Polarity: Adjectives	SEANCE