Marketing Science Institute Working Paper Series 2017
Report No. 17-124

Detecting Fictitious Consumer Reviews: A Theory-Driven Approach Combining Automated Text Analysis and Experimental Design

Ann Kronrod, Jeffrey K. Lee, and Ivan Gordeliy

Report Summary

Fraudulent user-generated content is harmful for both consumers and marketers and increases uncertainty about consumption experiences and offerings. To improve consumer experience online and increase consumer trust, marketers need a robust method to identify potentially fictitious product reviews.

Here, Ann Kronrod, Jeffrey Lee, and Ivan Gordeliy address this need via a novel method leveraging linguistic theory, experiment-driven data sampling, and automated text analysis on the language used in reviews.

Relying on literature about linguistics of experienced and imagined events, they develop the following conceptualization: reviews with fraudulent user-generated content should exhibit (1) *fewer verbs in the past tense* (given the lack of memory of sequences of events), (2) *fewer unique words* (because the teller relies on general knowledge rather than on unique experience), and (3) *more abstract language* (given the lack of concrete memories to share).

The authors tested these predictions by using automatic text analysis tools on authentic and fictitious reviews written by volunteer participants for the purpose of this work. As expected, they found that writers of authentic reviews used significantly more *past tense verbs*, *unique words*, and *concrete nouns*, compared with writers of fictitious reviews. Importantly, they found that these features of authentic reviews are difficult to falsify. Even when writers of fictitious reviews received clues about these aspects of authentic reviews, they were unable to replicate the frequencies of these aspects used by authentic review writers.

Implementing an experimental design, the authors also investigated people's ability to detect fictitious reviews. Replicating previous findings in the literature, participants were unable to distinguish fictitious and authentic reviews at a better level than chance (49%-52% successful detection). Interestingly, some participants were informed about the linguistic aspects that distinguish authentic from fictitious reviews. These participants became more suspicious, labelling more reviews as fictitious, but overall their detection rates did not improve. These findings suggest that a computerized detection approach offers advantages relative to an approach reliant on human judgement of review authenticity.

These findings offer insights to consumers as well as managers of digital platforms that depend on consumer trust and on an abundance of authentic user-generated content. The study contributes to theory regarding the linguistic features of a lie, and educates consumers on how to avoid naïve reading of product reviews. The results also demonstrate the advantages of using automatic tools to detect potentially fraudulent online content, and provides the basis to develop practical methods for detecting deception in consumer reviews.

Ann Kronrod is Assistant Professor of Marketing, Department of Marketing, Entrepreneurship and Innovation, Robert J. Manning School of Business, University of Massachusetts Lowell.
Jeffrey K. Lee is Visiting Assistant Professor of Marketing, New York University, Shanghai.
Ivan Gordeliy is a Postdoctoral Researcher, Group for Neural Theory, LNC, DEC, ENS, École Normale Supérieure, Paris.

Acknowledgments
The authors would like to thank Marketing Science Institute for their funding support for this project. The authors are thankful to Seshadri Tirunillai for his help in developing the code and computational approach in the earlier stages of this work. We thank Ravi Kiran and Wang Wan for their assistance in this project. We would also like to acknowledge the valuable input from Alireza Alemi, and to thank Liudmyla Kushnir, Pantelis Leptourgos, Vasily Pestun, Vasil Khalidov and Alexey Arbuzov for fruitful discussions which promoted this work. Part of this project was conducted while the first author was Assistant Professor of Advertising, The College of Communication, Michigan State University.

Introduction

As user-generated content continues to proliferate online, there is a rise in fraudulent information, which is harmful for both consumers and marketers. For example, Anderson and Simester (2014) found that approximately five percent of product reviews on a large retailer's website were posted without record of the author ever purchasing the product. Restaurants with uncertain reputations may attempt to "improve" their reputation by faking positive reviews, especially in a highly competitive market (Anderson and Magruder 2012; Mayzlin, Dover and Chevalier 2012; Luca and Zervas 2013). Fraudulent reviews challenge consumers, marketers and market researchers (Anderson and Simester 2014; Malbon 2013; Streitfeld 2012) and increase uncertainty about consumption experiences and offerings (Zhao et al. 2013)

The problem of fraudulent reviews is compounded by the fact that people tend to be naïve about the authenticity of the reviews they read. User-generated content is often perceived by consumers as an objective sharing of unbiased opinions about consumption experiences that actually occurred (C. Schellekens, J. Verlegh, and Smidts 2010; Chen and Xie 2008; Kronrod and Danziger 2013; Moore 2012). Some commercial websites such as Yelp or Epinion have developed algorithms to de-select reviews that are likely to be fraudulent (Li et al. 2011; Luca and Zervas 2016). Existing literature has identified quantitative factors that may increase the likelihood that a review is fictitious, such as the number of reviews written by a "first time reviewer" (Wu et al. 2010), whether the reviewer has authored many reviews in a single day with identical high or low ratings (Lim et al. 2010) and other behavioral footprints of the reviewer (Mukherjee et al. 2013).

In addition, over the years there has been interest in identifying the sincerity of a text by its *linguistic* cues (Hauch et al. 2015; Newman et al. 2003; Vrij 2008). The notion underlying

this approach is that it is possible to infer a person's thoughts, feelings, attitudes and motivations from the language they use. This literature has found that people who lie tend to use fewer first person pronouns, perhaps to dissociate themselves from the lie (Knapp, Hart, and Dennis 1974; Wiener and Mehrabian 1968). Appendix A provides a summary of studies on the language of insincere texts. These studies have collectively laid an important foundation for thinking about the way truthful texts differ linguistically from insincere texts.

A unifying theory is an important contribution to current approaches to insincerity detection, because it offers a framework to predict consumer behavior, based on universal psychological concepts tying thought and language. The purpose of this work, therefore, is to propose such a theory and to test its ability to generate several deep linguistic features that can distinguish between authentic and fictitious reviews.

Our fundamental question is: how does the language describing an experience differs if the teller has (versus has not) been through an experience? We propose that if the teller has not been through an experience, his or her language would exhibit (1) fewer verbs in the past tense, (2) fewer unique words, and (3) more abstract language. We explain the importance of these linguistic features of insincerity via the psycholinguistic literature, which distinguishes between the cognitive processes associated with experience-driven versus fictitious descriptions.

As a test of the predictive limitation of our propositions, we explore whether liars could strategically leverage these linguistic features to improve the perceived authenticity of their reviews. Just as liars strategically employ more first-person pronouns (Berzack 2011; Ott et al. 2011) in order to come across as more truthful (Vrij, Edward, and Bull 2001), it may be possible for liars to successfully fake reviews using our proposed features. At the same time, if liars are unable to use our features to fake reviews, it can be said that our theory is especially useful in

detecting fictitious reviews, by offering language features that are difficult to manipulate. We predict that while liars may be able to disguise their reviews through increased usage of the past tense, they will be unable to use more unique words or increase the concreteness of their reviews, because these linguistic features require the teller to have actually had the experience. In other words, our theory suggests linguistic features that are both predictive of deception and difficult to leverage by those who wish to deceive.

Finally, we explore whether human readers benefit through awareness of our proposed linguistic features of insincerity. Similar to our predictions on liars' strategic usage of these language features, we predict that readers would benefit from knowing the relationship between lowered past tense usage and fraudulent descriptions. At the same time, we predict that readers will not benefit from knowing the relationship between reduced use of unique words and lowered concreteness and fraudulent descriptions. The latter prediction is supported by the deep psycho-linguistic relationship between these two features and truthfulness, which cannot be easily detected by naïve readers even when being aware of them, because of the limited processing capacity of human cognition. Such a finding would also provide support for the usefulness of automated text analysis tools to detect fraudulent reviews based on these linguistics features, as they are not easily identified by humans.

We test our predictions by analyzing authentic and fictitious reviews written by participants for the purpose of this work. We implement an experimental design in our studies, by assigning our participants into conditions where they receive clues about the different features of fictitious reviews, to assess whether they are able to adjust their written language with these clues in mind. We then use automatic text analysis tools and algorithms that help us calculate differences in the use of language features such as the past tense, the use of unique words and

review concreteness. Finally, we investigate participants' ability to detect fictitiousness after providing them with awareness of the relationship between these features and fictitious writing. We also examine participants' lay beliefs about the features of authentic and fictitious reviews.

The contribution of this work is both theoretical and practical. Specifically, we provide a coherent theoretical framework that (1) predicts consumer behavior in both the reading and writing of product reviews, (2) contributes to research connecting consumer experiences and the linguistic features used when describing these experiences, and (3) contributes to the development of a psycho-linguistic theory of lies. Additionally, we inform marketers' understanding of the process of writing authentic and fictitious reviews, and suggest feasible ways to detect insincerity in user-generated content via automated tools, which may be especially helpful for digital platforms that depend on consumer trust and an abundance of authentic user-generated content. Finally, this work combines experimental methods (involving both consumer production and evaluation of user-generated content) with automated text analysis, highlighting potential advantages of a multi-method approach in the detection of fictitious reviews.

THEORETICAL DEVELOPMENT

Linguistic Characteristics of Lies

Literature examining the psycho-linguistic features of in-person and online communication suggests a wealth of insights regarding the language people employ when being deceptive (see Appendix A for a summary). This literature provides consistent evidence that lies are characterized by reduced usage of first person pronouns fraudulent (Hancock et al. 2007; Li

et al. 2011; Newman et al. 2003; Toma and Hancock 2010), because liars seek to avoid personal and specific referencing (Villar, Arciuli, and Barsaglini 2010). In the domain of product reviews, a recurrent finding is that insincere reviews tend to be more extreme – i.e. very positive or highly negative (Li, Huang, Yang and Zhu 2011; Luca and Zervas 2013).

While a few of the extant findings in the literature are theory-driven (Newman et al. 2003), among many of these results, little is known about why these linguistic features reflect deception. For example, while some research finds that deceptive reviews tend to be more extreme (Li et al. 2011; Mukherjee et al. 2013), this result is hard to reconcile with the general finding that reviews also tend to generally distribute in a bipolar way, with a high frequency of one and five star reviews, compared with less extreme reviews (Li and Hitt 2008). As a result, it is hard to leverage the extremity finding to generate a reliable indicator of deception.

Indeed, some documented markers of insincerity also contradict each other. For example, Li, Huang, Yang and Zhu (2011) find that fraudulent reviews are shorter that authentic ones, but Hancock, Curry, Goorha and Woodworth (2008) find that people who lied produced more words. Similarly, Van Swol and colleagues (2012a, 2012b) find that the bigger the lie, the bigger the number of words used by the liar. A second example of contradictory findings lies in negative emotion words. Some researchers find that liars express fewer negative emotion words (Newman et al. 2003). However, Anderson and Simester (2014) and Newman, Pennebaker, Berry and Richards (2003) find more negative emotion words in fraudulent reviews. Consistent with the latter findings, Hancock et al. (2008) and Toma and Hancock (2010) suggest that, compared to truthful statements, lies generally include more negations. Finally, while some literature demonstrates reduced usage of first person pronouns in fraudulent writing (Hancock et

al. 2007; Li et al. 2011; Toma and Hancock 2010), Berzack (2011) shows increased usage of first person pronoun in fraudulent texts.

In sum, findings regarding the linguistic characteristics of insincere texts are sometimes contradicting, and oftentimes lacking with respect to knowledge of the underlying mechanism or explanation for the finding. Many of the findings presented in Appendix A have not been replicated, and these findings are rarely supported by theory. In the next section, we develop a theoretical conceptualization of the processes of fictitious review writing, and we propose a set of linguistic features that arise from our theory and can be used to detect insincerity.

The Characteristics of Authentic versus Fictitious[1] Descriptions of Experiences

Imagine how you would write about your last flight to the moon. How would the language of your tale about your flight to the moon be different from your description of your last flight abroad? In other words, what are the specific linguistic features used by people who haven't actually had the experience?

A fictitious review can be defined as one that contains descriptions of events or experiences that have not actually happened to the author. We suggest that authors of fictitious reviews must activate capacities other than the mere recall of the experience. Specifically, the

[1] Note on Definitions and Terminology: Various terms are used in literature to describe insincere reviews, such as fictitious, insincere, fake, fraudulent, non-authentic, deceptive, and fabricated. The terms used for true reviews are: true, real, authentic, sincere, and genuine. The phenomenon we are investigating is of people writing a consumer review for a product or service they have not experienced. This behavior is characterized by writing up an opinion for and/or reporting an experience that did not take place. Distinctions such as whether the reviewer was paid (shill reviews) or unpaid, or whether they were motivated in any other way or not are not relevant to the question whether the reviewer has been through the experience before telling about it. Further, this paper is concerned only with cases of intentionally (consciously) made-up reviews. Thus, a consumer who tells about an experience that actually happened, but non-consciously does not describe the product experience properly, does not fall into our scope of interest (although analyzing the language of mistaken or inaccurate reviews bears its own value). In the current work, therefore, we chose to use the term "authentic" for reviews of actual consumer experiences, to reflect the behavior of actually *re-viewing* an incident that has happened before. We refer to reviews of experiences that did not occur in reality as "fictitious" reviews, to emphasize the nature of composing these reviews by making them up.

deceptive writer is engaged in the invention and construction of a description that is *not* based on actual, previous experiences. Instead, during this invention process, the teller employs general previous knowledge about the domain (e.g. travelling) relating to the description. In other words, the liar is likely to employ semantic memory (i.e. conceptual or categorical knowledge) as opposed to the truth teller, who would be more likely to employ episodic memory (i.e. recall and recognition of the details of particular experiences) (Mantonakis, Whittlesea and Yoon 2008; Tulving 1985a, 1985b).

Research on autobiographical memory suggests that episodic memory is based on re-experiencing the remembered event. By contrast, semantic memory is based on familiarity with terms associated with similar events (Burt 1999; Tendolkar 2008). While in episodic memory the mind searches for links with experiences retained in memory, in semantic memory the links lead to categorized knowledge, obtained via automatic abstraction of generalized terms and meanings (Kanwisher 1987). Thus, the recall of an event from episodic memory leverages aspects of the event (time, space, reenactment of sensations, thoughts and emotions etc.). By contrast, when a person makes up a fictitious story, the information for the story comes not from an actual event, but rather an organization of semantic meanings of words which constitute knowledge.

Based on this distinction between semantic and episodic memory, we suggest that when reviewing an actual experience that they have gone through, people rely on episodic memory. However, when writing a fictitious review, people do not rely on episodic memory for a specific event, but rather on the semantic memory of ideas that are frequent in the domain of their tale. Furthermore, given this distinction in memory processes, we suggest that the language of authentic reviews will be different from that of fictitious reviews. A wealth of literature has documented the link between cognition and language (Pinker 2013). The mounting evidence

indicates that cognitive processes are reflected in an individual's language, and there is also evidence that memory processes are exhibited in language use. For example, episodic memory can influence the linguistic clarity and detail of autobiographical descriptions (Irish et al. 2016). Similarly, semantic memory can influence lexicon structure and in turn word choice in communication (Takashima et al. 2017).

In sum, we suggest that reviews based on episodic memory are linguistically different from reviews that are based on semantic memory. We suggest three specific linguistic differences between authentic and fictitious reviews:

Use of the Past Tense. When describing an experience that occurred in the past, people usually employ the past tense. Indeed, previous research showed that descriptions of an event that did not occur entailed a reduction in the usage of the past tense (Dulaney 1982).We suggest that since describing a fictitious event does not rely on episodic memory, there is no temporal link between the fictitious events being described and any actual events that occurred in the past. Thus, the liar does not sufficiently associate the events in his or her story with proceedings that occurred in the liar's past. This prediction is consistent with research suggesting that lying witnesses use nouns and adjectives over verbs and adverbs (Filipović 2007; Sokolowski 1977), because they lack memories of prior actions (which would be best captured by verbs and adverbs). In sum, we suggest that authors of fictitious reviews are less likely to employ the past tense in their writing.

Unique Words. When describing an experience that actually happened, people tend to use words and expressions that reflect their unique experience, based on their episodic memory of that experience. By contrast, deceptive writers are forced to use their imagination, basing descriptions on semantic memories that are relevant to the topic and on general knowledge about

the domain (Abraham and Bubic 2015). Previous research has also shown that some semantic memories are created through frequency and repetition (as opposed to novelty and uniqueness). For instance, the meaning of words in semantic memory come from the most frequent occurrences of the word and their associated contexts (Kelly et al. 2001). In sum, we suggest that authors of fictitious reviews tend to use words that are less unique and to repeat words that they previously used (Kelly et al. 2001). At the same time, truth tellers are more likely to use unique words and expressions, which are available to them through the recollection of the elements associated with their unique experience.

Concrete Language. As mentioned earlier, fabricating a story requires the use of words and ideas that have undergone automatic abstraction through the process of knowledge acquisition (Kanwisher 1987). Further, the making up of an experience involves drawing facts from imagination. Imagination involves the abstraction of content from other experiences (Allport and Postman 1947; Hansen and Wänke 2010; Plous 1993; Tversky 1982), and abstract descriptions attend to overall information and more holistic attributes (Aggarwal and Law 2005).Given that liars are reliant on their imagination to produce fictitious reviews, we suggest that generating fictitious reviews will involve more abstract language, and less concrete language. By contrast, truth-tellers are likely to leverage concrete language as they recall specific incidents and events within their lived experience.

Formally, we provide several hypotheses relating to the linguistic markers of fictitious reviews:

H1a: Authors of fictitious reviews will use *less past tense*, compared with authors of authentic reviews.

H1b: Authors of fictitious reviews will use *fewer unique words*, compared with authors of authentic reviews.

H1c: Authors of fictitious reviews will use *less concrete (more abstract) language*, compared with authors of authentic reviews.

The Predictive Power of the Proposed Theory

One of the difficulties in distinguishing authentic from fictitious texts is that some liars may strategically disguise their insincerity, by incorporating known or intuitively obvious features of truthful text into their lies. For instance, review spammers (writers who deceptively mass-produce reviews on specific products to artificially raise or lower the ratings of these products) disguise themselves as genuine reviewers by avoiding known signals of deception (such as wordiness), and embracing known signals of truthfulness, such as increased use of first person pronouns and conjunctions (Berzack 2011). Similarly, sufficiently motivated liars avoid overly positive review writing - another known signal of deception - by using more negation in their language (Ott et al. 2011; Ott, Cardie, and Hancock 2012). Finally, liars can also increase perceptions of truthfulness by substantiating their lies with factual support about their behavior or the environment (Liebes 2001).

Given the potential ability of liars to increase the perceived truthfulness of their reviews using strategic textual modifications, the task of identifying deception can be very difficult. However, we assert that not all linguistic features of deception can be easily disguised, even if writers are (or made) aware of them. Specifically, with regards to our theoretically-derived features of fictitious reviews (reduced past tense, increased use of unique words, and increased

language concreteness), we suggest that liars will find it difficult to employ more past tense, unique words or concrete descriptions, relative to truth tellers, *even if made aware of these potential features of truthful reviews*. This prediction follows from liars' lack of mental links to actual experiences stored in their episodic memory (Mantonakis, Whittlesea and Yoon 2008), which reduces the availability of these descriptions to them.

We therefore predict that:

H2: Writers of fictitious reviews will not be able to use as much past tense, unique words and concrete language, as authentic review authors, even after being informed about those features of authentic reviews.

Consuming Authentic and Fictitious Reviews

Overall, humans are not very successful at detecting lies (Malbon 2013). On average, humans are about 53% accurate when attempting to detect whether communication contains deception or not, which is nearly identical to guessing at chance (Anderson and Magruder 2012; Bond Jr and DePaulo 2006; Malbon 2013). Moreover, awareness of the possibility that a text is fictitious does not improve detection (Van Swol, Braun, and Kolb 2015). It is also difficult to train humans to improve their accuracy at deception detection (e.g. by training humans to recognize nonverbal, paraverbal, and verbal cues associated with lying or telling the truth). In a meta-analysis of 30 studies, Hauch et al. (2016) find only a small to moderate training effect on the accuracy of deception detection by lab participants. More surprisingly, Aamodt and Custer (2006) find that local and federal law enforcement agents - experts with years of training - may still be no better at detecting deception in potential crime situations than college students.

Nonetheless, some previous literature has shown that people can intuitively identify certain features of insincere language on their own (Vrij et al. 2001). Additionally, people can become aware of certain structural features of lies, such as reduced use of first person pronoun, and this knowledge can help them identify insincerity (Baskett and Freedle 1974; Vrij et al. 2001). At the same time, the aforementioned literature reports the relative weakness of human detection of insincerity, and suggest that training may only have a modest effect. Given this mixed literature, we pose the following Research Question:

RQ1: If readers are made aware of the relationship between deception and the three linguistic features proposed by our theory (that is, reduced use of past tense, unique words and concreteness), would they be able to effectively incorporate this knowledge into their detection of fictitious reviews?

METHOD

Overview of Methodological Approach

To test our predictions we first obtained a set of authentic and fictitious reviews for a hotel stay (Study 1). In this study, participants were randomly assigned to one of six review-writing conditions. In one condition, participants wrote a review about an actual hotel stay they experienced (e.g. an authentic review), while in another, participants created a review about a hotel stay that they did not actually experience (e.g. a fictitious review). Among the remaining four conditions, participants were asked to write fictitious reviews as well. However, we provided these participants with a clue about one of our predicted linguistic features of fictitious

reviews (e.g. reduced use of past tense, concreteness, and unique words), to see whether this would affect their review writing. We used the resulting sets of texts to assess our H1a-c and H2, and we test our hypotheses via automatic text analysis methods.

In a second study, we examine whether human readers can identify fictitious versus authentic reviews. We assigned participants to one of five conditions, and asked them to label a subset of the reviews from Study 1 as either authentic or fictitious. In some of the conditions, participants received clues about one of our predicted linguistic features of fictitious reviews. Through these clues, we assessed whether participants could be given knowledge that would improve their ability to identify fictitious reviews.

Study 1 – Obtaining the Database, Performing Text Analysis to Distinguish Authentic from Fictitious Reviews, and Testing the Usefulness of Clues to Disguise Fictitious Reviews

Procedure

Following previous research that employed Amazon Mechanical Turk workers to generate fictitious reviews (Gokhman et al. 2012), we recruited 1,261 MTurkers for this study. 71 participants were eliminated from our data analyses due to incomplete responses, missing reviews, or reviews that included less than one full sentence. This resulted in a final dataset of 1,190 respondents (593 women; mean age = 33, Min. age 18, Max. age 77).

All participants were instructed to write a review for a hotel or motel stay. Some participants were asked to write about a stay that they truly experienced. The instructions for this authentic review writing task read as follows:

"We would like you to write a review for a hotel/motel in which you actually stayed for at least 2 consecutive nights. Please recall your experience at the hotel and share it in detail in the space below. Please make your review about 10 sentences long. (If you have not stayed in a hotel/motel in the past 12 months, please state that in the space)."

Other participants were asked to write a fictitious review, about a hotel stay that they did not actually experience. These participants read the following instructions:

"We would like you to write a review for a hotel/motel in which you have not actually stayed. Please write your review as if you stayed in the hotel for at least 2 consecutive nights. Please share your experience at the hotel in detail in the space below. Please make your review about 10 sentences long."

Conditions

Participants were randomly assigned to one of six conditions. Two conditions involved writing either (1) an authentic review or (2) a fictitious review, with the instructions as described in the previous section. Three conditions also involved the writing of a fictitious review, but participants in these conditions were also given one "clue" about a linguistic characteristic of fictitious reviews, as a test of whether this would affect participants' writing. A sixth condition also involved the writing of a fictitious review with a clue, but the clue given to these participants suggested that fictitious reviews exhibit fewer first person pronouns (a finding from previous research). By including this sixth condition, we can test whether the use of first person pronoun is one that writers can easily fake, as proposed in previous literature (Berzack 2011). More generally, we are also able to explore whether fictitious writers indeed use fewer first person pronouns in our dataset. If so, this might indicate that the reviews we collected for this study do not significantly deviate from previously datasets, given that the first person pronoun finding is well-documented.

Participants in the four clue conditions received one of the four descriptions below, depending on condition:

Past tense: “Please note: Scientists have discovered some characteristics of insincere (fake) reviews. One such characteristic is less use of the past tense. A fake review is more likely to use present or future tense. When composing your fake review, we encourage you to use this clue to improve your text.”

Unique words: “Please note: Scientists have discovered some characteristics of insincere (fake) reviews. One such characteristic is less use of special words. A fake review is more likely to use frequent words that are common for that product (for example the word "keyboard" or "typing" for a computer). When composing your fake review, we encourage you to use this clue to improve your text.”

Abstract language: “Please note: Scientists have discovered some characteristics of insincere (fake) reviews. One such characteristic is less use of concrete language. A fake review is more likely to use abstract descriptions (for example, "comfortable" for a couch). When composing your fake review, we encourage you to use this clue to improve your text.”

Personal Pronouns: “Please note: Scientists have discovered some characteristics of insincere (fake) reviews. One such characteristic is less use of first person narration ("I" or "we"). A fake review is more likely to use other pronouns ("you", "they" etc.)”

Subsequently, all participants wrote their review in a text box below the instructions and clues (if applicable). Then participants indicated their age, gender, whether English was their native language, and their level of education.

Results and Analyses

Table 1 provides descriptive statistics for each condition.

Table 1: Study 1 - Descriptive Statistics for Conditions

	Authentic	Fictitious	Past Clue	Unique Words Clue	Concreteness Clue	First Person Pronoun Clue
N reviews	174	202	189	217	205	202
Total Word Count per condition	22941	24980	24297	26980	26533	25705
Average Word Count per review	132	123	129	124	129	127
Minimum words per review	35	30	30	13	19	20
Maximum words per review	394	294	379	295	580	320
SD word count per review	51	42	49	46	57	45

Text Analysis Method and Results

In this section, we describe our text analysis methodology and present results from our study. A full and detailed description of the specific programming code and chosen functions can be found in the Tech Appendix for this article. We began our approach by cleaning our review data. To do so, we parsed the set of reviews to a SQL database table using Python code. Next, we removed all characters that were not letters or punctuation from the texts, and corrected spelling mistakes via Python's Enchant library.

Subsequently, we developed methods to detect each of our predicted linguistic features (e.g. past tense, unique words, concreteness and first person pronouns). In the following sections, we briefly describe these methods. Then, we leverage our method to test hypotheses H1a-c and H2, by quantifying and comparing differences across our study conditions.

1. Past Tense (H1a) – Definition and Analysis

We used the Treebank Project list[2] as well as the Brown Parts of Speech Tagger in Python's NLTK package to identify the tenses of verbs in our review dataset[3]. We then identified the number of verbs in each condition, and the number of verbs in the past simple tense among them. We used the proportion of past tense verbs out of all the verbs in our analyses. Table 2 reports descriptive statistics from our dataset.

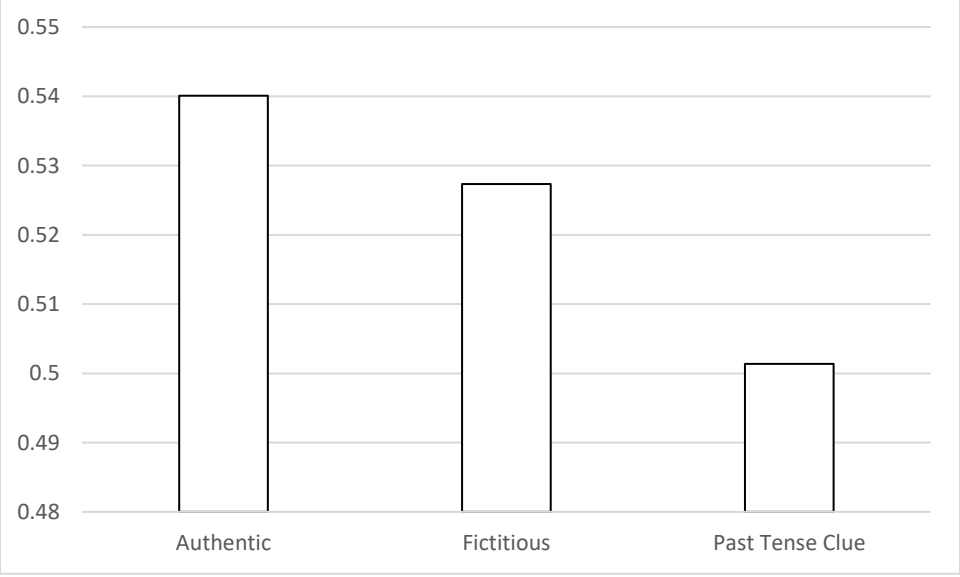
Table 2: Descriptive Statistics for Past Tense Analysis

	Authentic Reviews	Fictitious Reviews	Past Tense Clue	Total in Authentic, Fictitious and Past Tense Clue Conditions	Total in Entire Corpus
Past Simple Verbs	2162	2335	2216	6713	17038
Total Number of Verbs	4003	4428	4420	12834	32599
Proportion of Past Tense Verbs	.540	.527	.501	.52	.52

[2] https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
[3] http://www.nltk.org/book/ch05.html

As expected, writers of fictitious reviews exploited less verbs in the past tense, compared with authentic review writers, but this difference was only directional (Z = 1.17, p = .121). Participants used significantly fewer past tense verbs when they received a clue (0.595), compared with when they wrote fictitious reviews but did not receive a clue (Z = 2.44, p = .007), or compared to when they wrote authentic reviews (Z = 3.55, p < .001). Figure 1 presents the differences across conditions in the proportion of past tense verbs. All in all, it can be seen in Table 2 that the authentic reviews had a greater proportion of past tense usage, compared with all other conditions, and compared with the full corpus, which was predominantly made of fictitious reviews.

Figure 1: differences in the proportion of past tense use between Authentic, Fictitious and Clue conditions



Discussion

Our results on past tense usage directionally support the idea that writers of fictitious reviews avoid using the past tense in their language, relative to writers of authentic reviews. The difference may have been insignificant because of the potential ability of writers to make up for their insincerity by adding reference to past tense (Berzack 2011). However, our findings in the past tense clue condition suggest that these participants avoided using past tense even more. This result suggests that the clue did not help participants increase past tense usage in order to appear more authentic. It is possible that the clue imposed an additional cognitive load on our participants, which impeded their ability to efficiently write fictitious reviews by increasing their use of the past tense.

2. Unique Words (H1b) - Operationalization, Definition and Calculation

H1b predicts that writers of fictitious reviews would employ fewer unique words in their review, compared with writers of authentic reviews. Instead, fictitious review writers would use words that are relatively common to the topic they are writing about (e.g. hotels). To quantify the uniqueness of words, we examined the frequency distribution of each word form[4], within each condition.

[4] A word form is any sequence of letters with a meaning. For example, 'empty' is one word form, and 'emptiness' is another word form

The overall corpus in our dataset contained 181,144 words. The distribution of word count and word forms in the three focal conditions was:

Authentic	22,941 words	2,724 word forms
Fictitious	24,890 words	2,473 word forms
Unique-Words Clue	26,980 words	2,591 word forms
Total	84,820 words	6,820 word forms

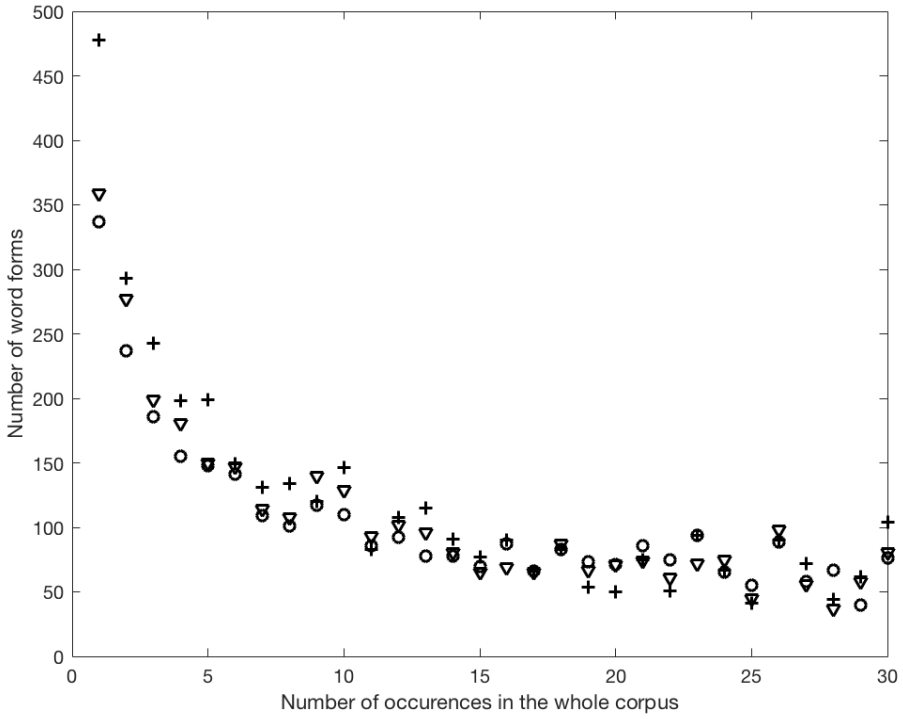
Words in our database distributed according to a power law. See full description of the distributions and calculations in the Tech Appendix. Figure 2 portrays the distributions of the frequencies (number of occurrences) of the different word forms, by condition. In all conditions a large proportion of the words (around 40%) were used only once in the whole condition. This distribution is similar to findings in other papers on language use (Dunning 1993).

To measure uniqueness, we began by looking at the frequency of occurrence of different word forms in the full corpus. We examined each frequency level (one occurrence, two occurrences, etc.) and counted the number of words forms that occur at that particular level. We then counted how many of the word forms in each frequency level are present in each of the three focal conditions: authentic, fictitious and unique-words clue condition (see Table 3). Next, we defined ’rare’ words as words that occurred in the full corpus no more than 27 times (see Technical Appendix for an explanation of this definition)[5]. The number of rare words in

[5] The highest frequency of word form occurrence was more than 13,000 times in the whole text, indicating the significance of defining rare occurrence at only 27 times.

authentic, fictitious and clue condition reviews was 3326, 2889, and 3219, respectively. We then calculated the ratio of rare word forms to total word count in each of the conditions, and compared the difference in proportions across conditions. We found that within authentic reviews, there was a significantly higher proportion of rare words (P = 0.145) compared with fictitious reviews (P = 0.116, Z = 9.39, $p < .001$) and clue condition reviews (P = 0.119, Z = 8.47, $p < .001$). Among the two fictitious conditions (no clue versus clue), there were no significant differences in the proportion of low-frequency word forms (Z = 1.14, $p = .127$).

Figure 2: Distribution of Number of Occurrence of the Different Word Form Frequencies in Authentic, Fictitious and Unique-Words Clue Conditions



+ Authentic ○ Fictitious Δ Clue

Table 3: Word Form Full Corpus Frequencies (first 20 levels) Occurrence by Condition

Number of occurrences per condition/Frequency of occurrence in the whole corpus	Authentic	Fictitious	Unique-Words Clue
1	478	337	359
2	293	237	277
3	243	186	199
4	198	155	181
5	199	148	150
6	149	141	147
7	131	109	114
8	134	101	108
9	120	117	140
10	146	110	129
11	83	86	93
12	108	92	102
13	115	78	96
14	91	78	81
15	77	70	65
16	90	87	69
17	66	66	65
18	83	83	87
19	54	73	67
20	50	71	71

An additional observation: we noticed that generally, the word count in the authentic reviews condition was smaller (22941) than in the fictitious reviews condition (24890) and the clue condition reviews (26980), while the vocabulary size (number of *different* word forms) was greater for the authentic reviews condition (2,724), compared with the fictitious reviews (2,473) and the clue condition reviews (2,591). This means that the language used in the authentic reviews was more diverse, whereas in the fictitious reviews and in the clue condition reviews there were more repetitions of the same terms.

Discussion

Results of our analyses support H1b, indicating that participants who wrote authentic reviews used significantly more unique words, relative to those who wrote fictitious reviews (with or without a clue). Providing support for H2, our analyses also show that participants who received a clue were not able to significantly increase their usage of unique words in their reviews. Our additional observation regarding the number of different words out of the overall number of words per condition provide further support for hypothesis H1b.

3. Concreteness (H1c) - Operationalization, Definition and Calculation

H1c predicts that writers will use less concrete (and more abstract) language when writing fictitious (relative to authentic) reviews. This prediction is derived from literature linking the imagination of experiences and conceptual generalization (or abstractness) in the mind. We predicted that fictitious reviews would be more likely to employ more general language because abstract descriptions tend to include more holistic attributes (Aggarwal and Law 2005). Previous

research has defined the concreteness (or abstractness) of a term as a function of the term's ambiguity and the number of associations that the term evokes[6] (Davidson and Laroche 2014; Dickson 1982; Krishnan, Biswas, and Netemeyer 2006; Lambert 1955; Paivio 1963; Rossiter and Percy 1978; Sadoski, Goetz, and Fritz 1993). We rely on this definition and operationalize concreteness following previous literature (Changizi 2008; Iliev and Axelrod 2017; Nelson 2017) that used the WordNet taxonomy (Princeton 2010) to establish a measure of concreteness, and specifically we focus on the hierarchy of nouns suggested by this taxonomy[7].

WordNet's taxonomy of nouns is organized by hierarchies of terms with more and less general meaning. For a given noun (such as *hotel*), words higher in the hierarchy are more abstract, and are termed "hypernyms" (e.g. *building*, *structure*). By contrast, words lower in the hierarchy are more concrete, and are termed "hyponyms" (e.g. *motel*) [8]. The hierarchy of nouns in WordNet is organized so that one super-hypernym (the noun *entity*) is at its top, and the rest of the nouns are organized on descending levels of the hierarchy. As an example, the noun *hotel* is located six levels below the hypernym *entity*. See figure 3 for a full representation of the hierarchy for the noun *hotel*. Note also that as in an ontology of words, WordNet hypernyms can have more than one descendant hyponym each. As a result, one can draw a tree of hypernyms and hyponyms, with lower branches of the tree reflecting deeper levels of noun concreteness.

[6] We are aware of the seminal and influential work by Semin & Fiedler (1988) who have defined the Linguistic Category Model (LCM) to describe levels of abstractness of a text. However, since this model is governed by the degree of interpretation of the event, we find it less relevant to this work and prefer other definitions that suit better with our theory and method.

[7] See Tech Appendix for an explanation of our choice to calculate concreteness using nouns only (and not other parts of speech).

[8] Note: WordNet has separate categories for fixed noun phrases (e.g. natural language processing). We did not include fixed noun phrases in our analysis.

Figure 3: Full representation for the WordNet hierarchy of the word hotel

- S: (n) **hotel** (a building where travelers can pay for lodging and meals and other services)
  - *direct hyponym* / *full hyponym*
  - *part meronym*
  - *direct hypernym* / ***inherited hypernym*** / *sister term*
    - S: (n) building, edifice (a structure that has a roof and walls and stands more or less permanently in one place) *"there was a three-story building on the corner"; "it was an imposing edifice"*
      - S: (n) structure, construction (a thing constructed; a complex entity constructed of many parts) *"the structure consisted of a series of arches"; "she wore her hair in an amazing construction of whirls and ribbons"*
        - S: (n) artifact, artefact (a man-made object taken as a whole)
          - S: (n) whole, unit (an assemblage of parts that is regarded as a single entity) *"how big is that part compared to the whole?"; "the team is a unit"*
            - S: (n) object, physical object (a tangible and visible entity; an entity that can cast a shadow) *"it was full of rackets, balls and other objects"*
              - S: (n) physical entity (an entity that has physical existence)
                - S: (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

Using the WordNet 3.1 taxonomy, we measured the relative "depth" of each noun by counting the number of steps down the hierarchy from the word *entity* to that noun (also called the length of the hypernym path). The greater the number of steps, the deeper the noun (Nelson 2017). Further, we assumed that the depth value assigned to each noun also represents the number of more abstract nouns that can be used instead of this noun.

Once all nouns in a given review have been assigned a depth value (equal to the number of steps down from the word *entity*), we were able to quantify the overall concreteness (or depth) of the review. We derive this overall concreteness score by calculating, via a combinatorial formula, the number of alternative texts that could have been generated for each review, using hypernyms instead of the actual noun used by the writer (see Tech Appendix, formula 1). To avoid dealing with the very large numbers representing the results of this calculation, we used the logarithm of this value as the concreteness score for each review. A full description of the method to calculate concreteness can be found in the Tech Appendix.

Results

Because the overall concreteness scores did not follow a normal distribution (see table 4 and figure 4). We used the log transformed the concreteness scores in our statistical analyses. An omnibus ANOVA test comparing the concreteness scores among authentic, fictitious and abstractness clue conditions revealed significant differences among these conditions ($F(2,578) = 3.49$, $p = .031$). Contrasts tests indicated that the language in authentic reviews was significantly more concrete than the language in the fictitious and abstractness clue conditions. However, there were no significant differences between the latter two conditions in abstractness ($t(1, 578) = 77.62$, $p < .001$). Figure 5 presents these results.

Table 4: Descriptive Statistics for Distributions of Concreteness Scores for Reviews

	N	Mean concreteness	Minimum	Maximum	SD	Skewness	Kurtosis
Overall	581	52.3	10.2	220.3	21.99	1.966	8.63
Authentic	174	56.02	17.94	184.35	24.08	1.663	4.804
Fictitious	202	50.39	10.27	136.66	18.78	1.242	2.724
Concreteness-Clue	205	51.10	10.20	220.31	22.78	2.580	14.877

Figure 4: Distributions of the Three Conditions of Concreteness – Scaled Comparison

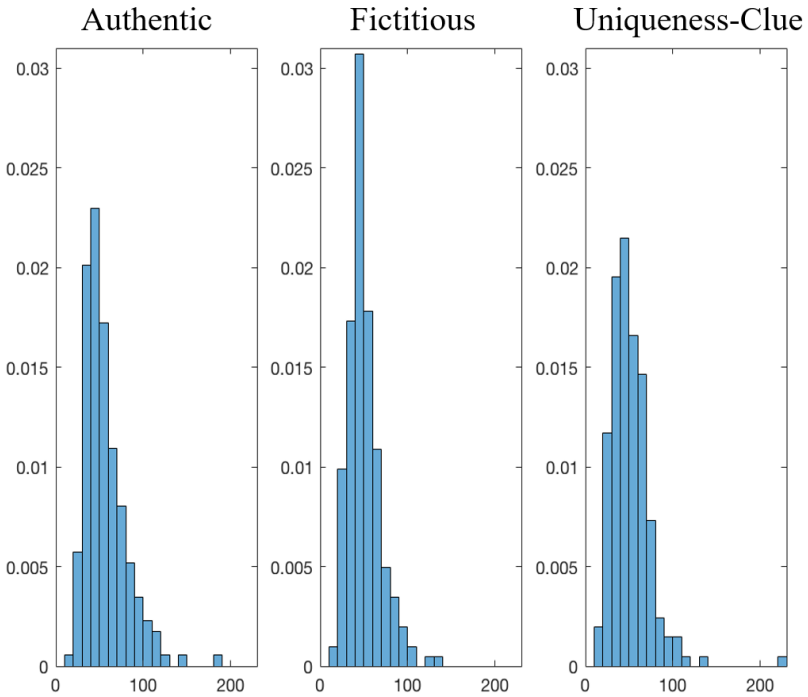
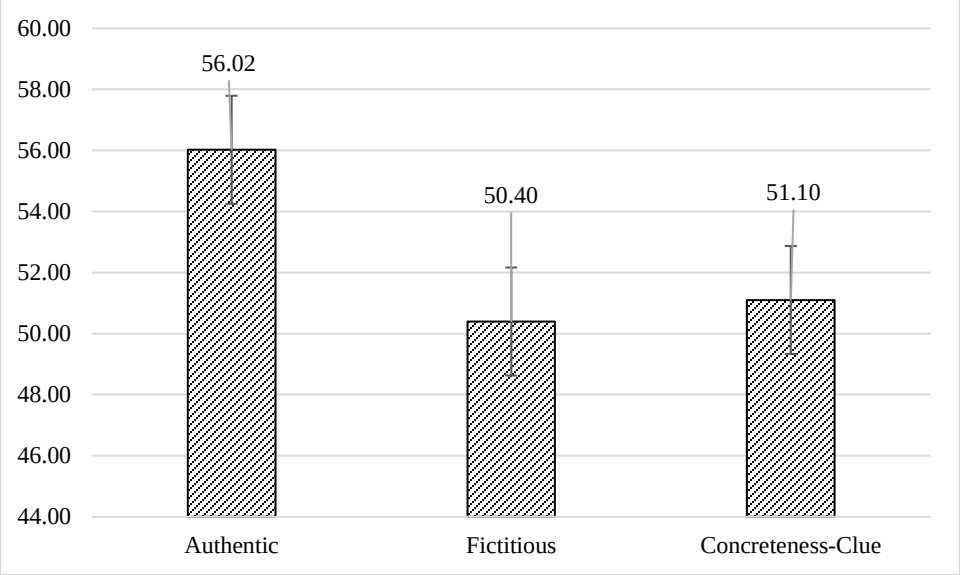


Figure 5: Differences in Average Concreteness between Authentic, Fictitious and Concreteness-Clue Conditions



Discussion

Results on language concreteness support H1c, suggesting that authentic reviews were more concrete in language than fictitious review (with a clue or without). Further, H2 was partially supported by our results as well, as participants were unable to generate more concrete reviews after receiving the clue.

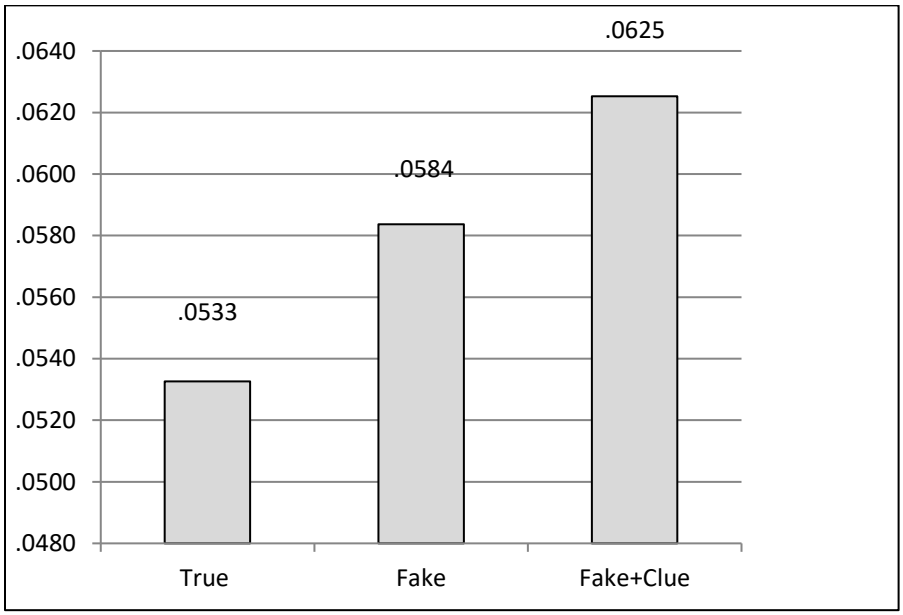
Use of First Person Pronoun (FPP)

To check the typicality of our corpus, relative to those analyzed in previous research (Berzack 2011; Hancock, Landrigan, and Silver 2007; Li et al. 2011), we compared the frequency of first person pronoun terms (e.g. I, we, me, us, my, our, mine, ours) in the three conditions. Descriptive statistics for first person pronoun usage in the three conditions are in Table 5. Z scores (comparing the proportion of first person pronouns within reviews) suggest that in the authentic reviews condition, participants used the first person pronoun significantly less than in the fictitious review condition, as well as the fictitious review with a clue condition ($Z = 3.25$, $p < .001$; $Z = 5.19$, $p < .001$, respectively). This result is consistent with Berzack (2011), who found no lower use of first person pronouns in lying texts. Figure 6 presents these results.

Table 5: Descriptive statistics for First person Pronoun and total Personal Pronouns in the Authentic, Fictitious and FPP-Clue Conditions

	True	**Fictitious**	**FPP-Clue**	**Total**
Word Count	22,941	24,890	24,297	72,128
"I"	603	757	752	5371
"we"	264	294	446	2499
"my"	156	234	246	1685
"our"	110	104	172	1000
"me"	47	64	65	495
"us"	44	49	68	453
"ours"	1	0	1	6
"mine"	2	1	0	5
Total	1227	1503	1750	11514

Figure 6: Proportion of First person Pronoun out of Total Words in Authentic, Fictitious and Fictitious-Clue Condition



Discussion

The result for first person pronouns suggests that some linguistic features of insincere text are intuitive and can be easily disguised. Further, participants who received a clue about the use of first person pronouns were able to use it even more than participants who wrote fictitious reviews without a clue.

Study 1 Conclusion and Discussion

Our findings in study 1 draw an interesting picture: first, we were able to support H1b and H1c, finding that writers of fictitious reviews tend to use less unique words and more abstract language. We also observed lesser use of the past tense in fictitious texts, but contrary to our prediction participants were not able to increase their use of past tense after receiving a clue about it. We did not observe lesser use of first person pronouns in fictitious reviews, and participants who received a clue increased their use even more. These results suggest that some linguistic features may be more intuitive and writers can manipulate them in less predictable ways. These results echo previous works suggesting that some authors are being successful in manipulating their reviews to disguise their insincerity.

How successful are consumers of product reviews at identifying insincerity? Study 2 investigates consumption of reviews and tests the predictability of our theory on linguistic features of authentic and fictitious reviews through experimenting with readers of the reviews that were composed in study 1.

Study 2: Detecting Authentic and Fictitious Reviews

The purpose of Study 2 was to address our research question, which asks whether readers will be successful at distinguishing between authentic and fictitious reviews when made aware of the linguistic features of fictitious reviews developed in this work. To investigate this question we ask participants to determine for 60 reviews whether they are authentic of fictitious, using the same set of clues employed in study 1.

Participants and Procedure

355 MTurk workers (mean age = 32.8, 143 women) took part in this study. Participants were instructed to read 60 reviews and to label each review as being either "true" or "fake". The instructions read as follows: "In this assignment you will read 60 reviews for a hotel. For each review, you will make a decision whether it is a **true review** (the reviewer actually stayed at the hotel and wrote the review after that) or a **fake review** (the reviewer has not actually stayed at the hotel and made up the review). Please make sure to leave time to read and rate all reviews in one sitting. Most of the reviews do not exceed 5 lines. Please note, your reading is timed, as you are expected to read the reviews before you rate them. To begin reading and rating the reviews please press the "next" button."

The 60 reviews were comprised of a randomly selected set of (1) 30 reviews from the set of authentic hotel reviews in Study 1, and (2) 30 reviews from the set of fictitious hotel reviews in Study 1 (where participants did not have a clue). Reviews were presented in a random order, and participants read and rated one review at a time. As a measure of effort, we timed the speed of ratings submission for each review. At the end of the ratings task, participants indicated their frequency of staying in hotels (1=less than once a year; 7=once in a few days), their age, gender, education level and whether English was their native language. Subsequently, participants were thanked and paid.

All participants were randomly assigned to one of five conditions. In one condition, participants merely received the above instructions and evaluated the reviews. In the other four conditions, participants read one of the four clues provided to participants in Study 1.

Results

Attrition, reading time and participant inclusion. 27 respondents skipped at least half of the reviews; therefore we omitted their responses from the analyses, leaving us with a dataset of 328 participants. On average, participants spent 24 seconds (SD = 21 sec) reading a review and rating it as authentic or fictitious. Previous research has suggested a variance in reading speed of about 30%-40% among readers (Miyata et al. 2012; Skinner et al. 2009). With this finding in mind, we took the average time spent reading reviews ($24_{seconds}$*$60_{reviews}$ = 1440 seconds), and multiply it by 0.6, which would represent 40% faster reading time (1440*0.6 = 864 seconds) to generate a threshold for unusually fast reading speeds in our study. We identified 99 participants who met this criteria, and we compared the results of analyses with and without this group. We found similar results, and therefore these participants were not removed from the data and results are reported for the full database. Interestingly it took participants on average less time to read a fictitious review than an authentic one, though this difference was not significant[9] ($F(1,327) = 0.22$, $p = 0.642$).

Successful classification of authentic and fictitious reviews. We used the number of correctly identified authentic reviews, correctly identified fictitious reviews, and total number of correctly classified reviews, in our analyses. Frequency of staying in hotels did not significantly differ across conditions ($p = .315$); thus, it was not included as a variable in the analyses.

[9] Note that Sphericity assumption was violated, because the Mauchly's test was significant. Corrections for sphericity violation (e.g. Greenhouse) all showed that this mean difference was still not significant.

Successful classification across clue conditions. A 2-way mixed repeated measures analysis with number of correct authentic versus fictitious review detections as a within subjects factor and the experimental condition as a between subjects factor revealed a marginally significant main effect for condition ($F(5,322) = 2.13$, $p = .062$).

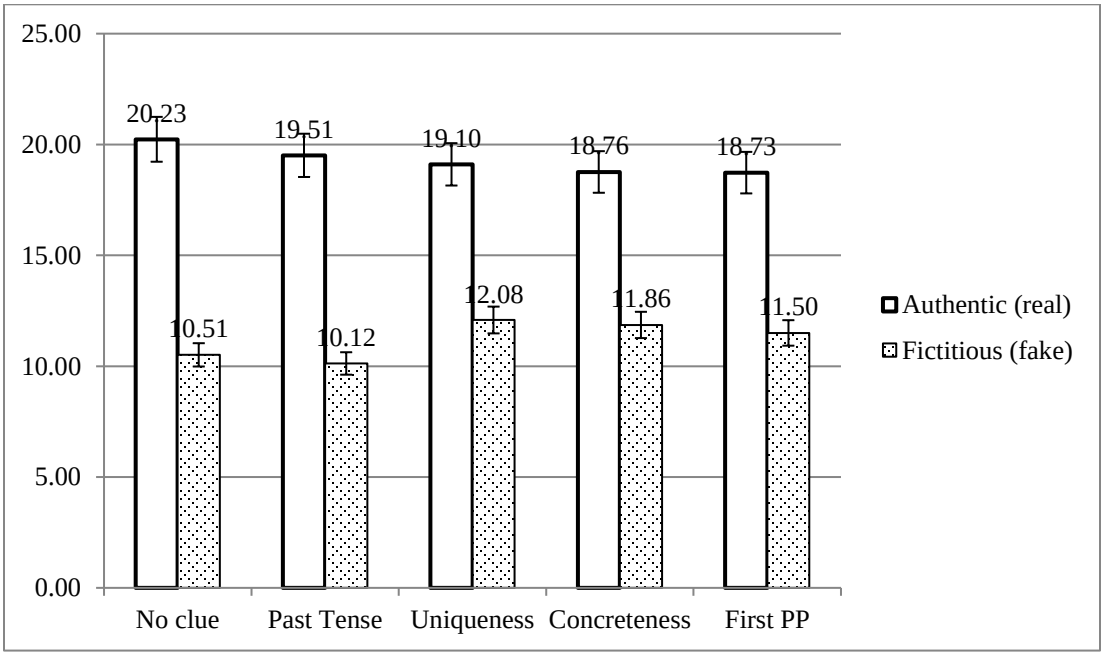
We found a significant interaction of clue-condition and review type (authentic/fictitious) ($F(5,322) = 2.31$, $p = .044$). A contrast test for the differences between the number of correct detections of authentic versus fictitious reviews within each condition (see figure 7) suggests that within the authentic reviews, our participants were marginally more successful at detecting authentic reviews when they had no clue ($M = 20.23$), compared with people who received a clue about First Person Pronouns ($M = 18.73$, $p = .052$), and about abstractness ($M = 18.76$, $p = .055$). There were no differences among the clue conditions in correct detection of the authentic reviews.

For detection of fictitious reviews, we found participants had marginally fewer correct detections when given no clue relative to a unique-words clue ($M = 12.08$, $p = .058$). Participants who did worse in the detection of fictitious reviews were those who were given a past tense clue ($M = 10.12$); those in this condition were significantly less successful than participants in the unique-words clue ($p = .017$), and concreteness clue ($p = .033$) conditions.

Aggregate impact of clues on accuracy. To explore the overall impact of clues on judgment accuracy, we collapsed all the clue conditions (First Person Pronoun, Past, Unique Words, Concrete) into an overall "clues" experimental cell. We then compared this cell with the "no clues" control cell. We also separate our analysis through authentic versus fictitious reviews categorization. A 2-way mixed repeated measures analysis revealed a significant interaction

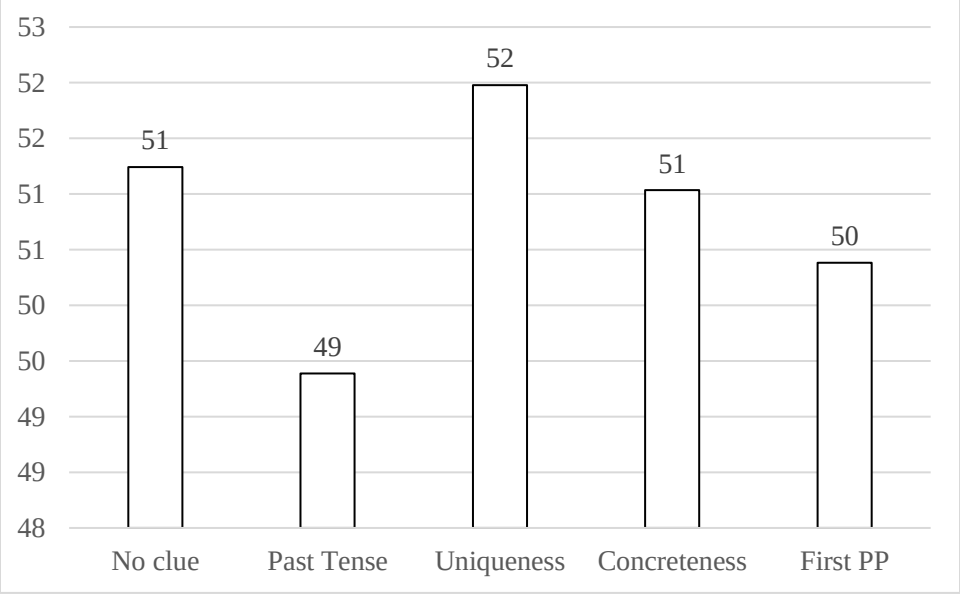
($F(1, 322) = 5.21$, $p = 0.023$). We can decompose this interaction by comparing judgements on authentic versus fictitious reviews. When participants judged fictitious reviews, providing them with a clue increased the number of fictitious reviews they accurately categorized as fictitious ($M = 11.77$, $F(1,326) = 3.81$, $p = 0.052$) relative to participants who did not receive clues ($M = 10.51$). At the same time, when participants judged authentic reviews, providing them with a clue decreased the number of authentic reviews they accurately categorized ($M = 19.02$, $F(1, 326) = 4.21$, $p = 0.041$) relative to those who did not receive clues ($M = 20.23$). Overall, clues helped participants to detect fictitious reviews, but also hurt them when they judged actual reviews. This pattern of results suggests that clues drove participants to increased suspicion, but did not increase judgment accuracy, as participants merely overall deemed more reviews as fictitious, but were not more successful at distinguishing between authentic and fictitious reviews.

Figure 7: Average Number of Correct Detections of Authentic and Fictitious Reviews, by Clue Condition



Aggregate analysis of correct classification. Figure 8 present results on correct classification of reviews, aggregated among both authentic and fictitious reviews. There was a marginally significant difference among the clue conditions in success of detection ($F(5,322) = 2.13$, $p = .062$). All classification rates were no better than guessing at chance, replicating previously findings suggesting a 50% level of success from human judges (Anderson and Magruder 2012; Bond Jr and DePaulo 2006; Malbon 2013). Although participants appeared to be most successful at detection when they were clued-in about the more complex features of language, such as unique and abstract language, these participants still did not outperform participants who received no clues to assist them in their evaluation.

Figure 8. Percent of Successful Detection of Authentic and Fictitious Reviews across Conditions



Exploratory analysis of open ended reflections. To better understand the features participants may have focused on to detect authentic versus fictitious review writing, we examined their open-ended responses on how they decided which reviews were fictitious and which were authentic. We submitted their responses to a simple word count algorithm, to determine the most commonly occurring features and themes among participants. The most recurring features used by participants to detect fictitious review writing were:

1. Avoidance of detail and specific descriptions or names
2. Vagueness and omission of information
3. Overly short or long reviews
4. Extreme valence: overly negative/positive/good/bad
5. Lack of descriptions of personal experience
6. The presence of grammar/spelling errors

This exploratory analysis suggest that participants were somewhat able to think of some of our predicted linguistic features of deception, such as lack of concreteness (as evidenced by Theme 1) and lack of unique words relating to personal experience (as evidenced by Theme 5). However, as the previous results in Study 2 demonstrated, reminding participants of these features of deceptive writing did not help them to correctly classify fictitious and authentic reviews.

Study 2: Conclusion and Discussion

Results of the current study suggest that humans tend to be highly inaccurate when judging the authenticity of product reviews. Providing participants with clues to detect fictitious reviews did not improve their overall judgement accuracy. Even when given clues to detect features of fictitious reviews, participants were still overly trusting. This finding echoes previous literature suggesting that people tend to assume that reviews are mostly authentic and truthful (Chen and Xie 2008; Kronrod and Danziger 2013).

Furthermore, clues had a positive effect on accurate detection of fictitious reviews, likely because clues increased suspicion, thus helping participants to label a fictitious review appropriately when they discerned features that reflected fictitious writing. Yet on authentic reviews, clues *reduced* participant accuracy, possibly because they became overly suspicious when evaluating these reviews. Participants were twice as inaccurate when classifying fictitious reviews relative to authentic reviews. This result is consistent with previous work suggesting that participants are poor at discerning fictitious versus authentic reviews (Anderson and Magruder 2012; Bond Jr and DePaulo 2006; Malbon 2013).

This pattern of results shows the limitations of interventions that provide participants with information on what constitutes a fictitious review. Clues increase suspicion, which helps in accurately detecting fictitious reviews, but also hinders the accurate judgement of authentic reviews. Our results suggest that providing humans with information to detect fictitious writing can backfire, because suspicion also leads to inaccurate judgment and perception of authentic writing. For these reasons, an automated (machine-driven) approach may be far more efficient at classify authentic versus fictitious reviews, relative to human judges.

GENERAL DISCUSSION

"I suggest that to make progress we do not need fully artificial intelligent text analysis; rather, a mixture of computationally-driven and user-guided analysis may open the door to exciting new results."

(Marti A. Hearst in the paper "*Untangling Text Data Mining*", 1999).

Hearst's statement mirrors the approach adopted in this work: the combination of automated text analysis and theory-driven linguistic hypotheses. Our investigation was driven by a theory regarding the language of reviews when an author has not actually experienced the product or service being discussed. Our theorizing yielded three predicted linguistic features of fictitious reviews, which we were able to empirically confirm. Fictitious reviews are characterized by reduced usage of the past tense, a smaller proportion of unique words, and reduced language concreteness. Furthermore, in two experiments with both writers (Study 1) and readers (Study 2) of hotel reviews, we found that two of our theory-driven features cannot be imitated, even when participants are made aware that these features exemplify fictitious writing. We also find that awareness of our predicted linguistic features cannot help readers of the reviews to correctly classify fictitious and authentic reviews. In fact, none of the clues we offered our participants were helpful in improving their accuracy rates. Also, in the open-ended responses in Study 2, many participants seemed somewhat aware of our theoretically predicted linguistic features of fictitious texts. However, this awareness did not translate into an accuracy rate that was better than at-chance guessing. An automated approach to the detection of fictitious review text, driven by linguistic theory, can help identify this type of deceptive text.

Theoretical Implications, Limitations and Future Research

We were careful in distinguishing among various language features used by review writers. However, there may be some correlations among our predicted language features. For example, it is possible that abstract language also involves common (relative to unique) words, because when describing abstract thoughts, people may rely less on unique and unusual experiences or features. Thus, an observed decrease among some of our conditions in the use of unique words may actually be due to an increased emphasis on abstractness. Indeed, there is a relationship between "buzz words" (which tend to be popular and common) and the abstractness of text (Heath and Heath 2007),supporting the idea that abstract thinking and common language may be correlated. In sum, some of our results may be driven by the relatedness among our predicted language features.

At the same time, future research could examine other language features, such as the linguistic complexity in a given text. Linguistic complexity is defined as the use of infrequent words, a higher level of syntactic structure (for example the use of clauses in clauses), implementing connectors and conclusion markers, and longer words, expressions, clauses and sentences (Gordon and Stuecher 1992; Juola 1998; Saslow et al. 2014; Whissell 1999). This concept is believed to represent cognitive complexity, given that the complexity of one's language is said to reflect the complexity of one's thoughts (Bard et al. 2007; McKimmie et al. 2013; Newman et al. 2003). Previous research has also found that truth-tellers are capable of higher cognitive complexity in their communication, compared with liars, because making up a story consumes cognitive resources more than merely recalling an experience that has actually occurred (Newman et al. 2003; Vrij et al. 2011). It is possible that the activation of semantic (relative to episodic) memory processes could place a high cognitive load on fictitious review

writers. If so, linguistic complexity could be another reliable marker of truth-telling, that is both difficult to fake and difficult for humans to detect (relative to a machine).

Finally, future research could explore our linguistic features in other consumer contexts. For instance, it would be theoretically important to know how sensitive our findings are to product or service category differences. Would findings be similar in categories where consumption experiences are inherently more abstract and subjective (e.g. artwork), versus more concrete in nature (e.g. filling up at a gas station)? Similarly, Schweidel and Moe (2014) found that different social media venues may impact the way consumers express brand sentiment. It is worth exploring whether our results would vary under different review generation and consumption circumstances. Future research could also quantify the benefits of an automated text analysis method for the social media venues themselves. Such a method could result in greater helpfulness review ratings for user-generated content, given the relationship between authenticity and helpfulness (Li et al. 2011), or increased positivity and engagement in the social media platform.

Potential Moderators for our Effects

Several moderating variables could influence the relationship between certain linguistic features (e.g. abstractness, common language usage) and deceptive writing. First, language fluency may play a key role. Previous research has suggested that people are more likely to believe that a non-native speaker is lying, compared with a native speaker (Evans and Michael 2014). This finding may be due to reduced usage of linguistic markers of truth-telling by non-natives, or it may be a more general bias against language disfluencies. Given increased global connectivity and the access of social platforms to an international, multi-lingual user base, the impact of language proficiency merits additional examination. Furthermore, if consumers are

indeed biased against language disfluencies, an automated approach to detecting fictitious writing may be truly superior to a human approach that is less forgiving of language disfluencies, and therefore more vulnerable to bias.

A second factor at play could be financial incentives. All of our participants received the same payment for their participation, and indeed, some did not complete our requested task correctly. Indeed, increased financial incentives could increase participant success in "faking" certain language features when given clues to write with these features. Nonetheless, previous research has shown that many consumers post fictitious reviews without any financial incentive to do so (Anderson and Simester 2014), suggesting that financial incentives are not necessary to spur deceptive behavior. In our view, even the most motivated liars will find it difficult to overcome their lack of episodic memories when writing about a fictitious experience. Future research may explore the role of incentives in fictitious review writing, as well as the types of clues provided to review writers who aim to deceive.

Finally, in our data analysis, we did not include additional aspects of the review or reviewer, such as expertise or gender. These differences have been found to affect the inferences readers of reviews make about the content of the reviews, and to affect the persuasiveness of reviews in general. Accordingly, gender and perceived expertise of the review writer (as well as other factors) could influence the perceived truthfulness of the writing. Furthermore, the relationship between such factors and our linguistic features could be complex. For example, in research on the persuasiveness of language complexity, McKimmie et al. (2013) found that when evaluating an expert's persuasiveness, complex language was associated with male experts, whereas simple language was associated with female expert persuasiveness. Extended to our content, it could be that male review writers come across as honest and truth when using

linguistic complexity (described in more detail below), while female writers do not gain the same benefits when using such language. Interestingly, Van Swol, Braun and Kolb (2013) find that people are more successful at detecting a lie in computer mediated communication, rather than face to face. This may be due to various styles of lie that are possible or impossible in these different communication situations.

Other works combined analysis of linguistic and non-linguistic factors to investigate the role and influence of consumer communication online on product success. Schweidel and Moe (2014) modelled brand sentiment in social media posts integrating the effect of different venues in different industries. The authors find that different social media venues may have a different effect on the way consumers express brand sentiment and the conclusions that marketers can derive about their brands. It is worth exploring whether our results will be different under different review generation and consumption circumstances. Further, experimentation with downstream effects such as engagement or attitudes towards the product or firm as a result of using an automatic detector of linguistic features of authenticity may reveal additional important outcomes of employing such a tool on commercial and noncommercial websites.

Practical Implications

Some previous work has focused on the automatic detection of fake product reviews. For example, the Pheme project, a collaboration among five European universities and four companies, aims to construct an automatic lie detector for social media[10]. Websites such as Yelp have an automatic selection system to identify fictitious reviews, according to features in the content. But psycho-linguistic, theory-driven research on the language in user-generated content

[10] EU project to build lie detector for social media. Source: http://www.sheffield.ac.uk/news/nr/lie-detector-social-media-sheffield-twitter-facebook-1.354715. Accessed on Feb 26 2014.

has yielded only a handful of discoveries (C. Schellekens et al. 2010; Kronrod and Danziger 2013). Moreover, automated detection of fake reviews currently relies heavily on numerical features, such as the number of reviews written by first-time reviewers on www.TripAdvisor.com (Wu et al. 2010) or the frequency of first person pronouns in the review (Ott et al. 2011, 2012).

The methodology in this paper adds to the aforementioned tools, by introducing predictive linguistic features that allow practitioners to reliably analyze the language of reviews relating to their businesses. These tools help practitioners to more accurately learn about attitudes and psychological aspects of their customers, by allowing them to filter out content that is likely to be fictitious and unrepresentative. Social network platforms can also enhance the consumer experience online via our tools, by ensuring that online content is more truthful, which can generate positive feedback loops for the platform (via increased consumer trust in the platform). Finally, knowledge of our predicted linguistic features can help marketers to participate more effectively in social media, employing language that might come across as more authentic and appropriate (Kozinets et al. 2010).

Human reviewers do have some intuition of when someone is lying. For instance, implicit measures of lie detection were found to be more accurate than explicit measures of lie detection (ten Brinke and Carney 2014; ten Brinke, Stimson, and Carney 2014). Nonetheless, most research reports a relatively low success rate on human detection of fictitious writing (Anderson and Magruder 2012; Malbon 2013). Based on statistics reported in Bond and DePaulo(2006) and Hauch et al. (2016)'s meta-analyses, we estimated that training can improve human deception detection up to 55.3% for the top 5% of human judges, well short of the

benchmarks for automated textual analysis[11]. By contrast, computerized analyses have traditionally yielded an accuracy rate of 67% (Newman et al. 2003). Our automatic detection method adds to the extant stream of work on computerized methods for detection of fictitious user-generated content. Our method contributes to the ways in which marketers can analyze and validate online communication about their products, without having to rely on explicit measures of lie detection collected from human judges.

Conclusion

The important role of social media in the success or failure of products is well known and widely researched, but except for a few exceptions (Jurafsky et al. 2014; Ordenes et al. 2014) the potential effect of the *language* that consumers use when communicating online about their experiences has been largely overlooked. Through a combination of automated text methods, linguistic theory, and experimentation, we aimed to develop a better understanding of the linguistic features of fictitious reviews. Our findings revealed the difference between humans and machines in detecting these features, as well as the inability of humans to replicate these features in writing. Given these results, our work adds to theoretical reasoning on the connection between language and memory, while also providing an approach that can be implemented to automatically detect fictitious writing on social media platforms. Ultimately, we hope that this work will increase interest in the language of product reviews and the implications of studying this rich (and oftentimes, revealing) type of data.

[11] . We arrived at the estimate of 55.3% accuracy (for the top 5% of trained human judges) using statistics from previous meta-studies. First, we noted Bond and DePaulo (2006) mean and SD (53.46%, 4.52%) accuracy rates for non-trained participants. Second, we noted Hauch et al (2016)'s general effect size of training, which was 0.331, as well as the confidence interval for the effect size (0.262 to 0.400). With these figures, we can compure the accuracy of the average participant, post-training, which should be (4.52 * 0.331 + 53.46) = 55% with training. For the elite participant (i.e. the top 5% participant), we can also compare their post-training accuracy, which should be (4.52 * 0.400 + 53.46) = 55.3% accuracy. .

REFERENCES

Abraham, Anna and Andreja Bubic (2015), "Semantic Memory as the Root of Imagination," *Frontiers in psychology*, 6, article No. 325 pp. 1-5.

Aggarwal, Pankaj and Sharmistha Law (2005), "Role of Relationship Norms in Processing Brand Information," *Journal of consumer research*, 32(3), 453–464.

Allport, Gordon W. and Leo Postman (1947), The Psychology of Rumor, Oxford, England: Henry Holt.

Anderson, Eric T. and Duncan I. Simester (2014), "Reviews without a Purchase: Low Ratings, Loyal Customers, and Deception," *Journal of Marketing Research*, 51(3), 249–269.

Anderson, Michael and Jeremy Magruder (2012), "Learning from the Crowd: Regression Discontinuity Estimates of the Effects of an Online Review Database," *The Economic Journal*, 122(563), 957–989.

Bard, Ellen Gurman, Anne H. Anderson, Yiya Chen, Hannele BM Nicholson, Catriona Havard, and Sara Dalzel-Job (2007), "Let's You Do That: Sharing the Cognitive Burdens of Dialogue," *Journal of Memory and Language*, 57(4), 616–641.

Baskett, Glen D. and Roy O. Freedle (1974), "Aspects of Language Pragmatics and the Social Perception of Lying," *Journal of Psycholinguistic Research*, 3(2), 117–131.

Berzack, Antony (2011), "Language Use of Successful Liars," Master of Science Thesis at Cornell University (https://core.ac.uk/download/pdf/6103650.pdf Accessed on June 29 2017).

Bond Jr, Charles F. and Bella M. DePaulo (2006), "Accuracy of Deception Judgments," *Personality and social psychology Review*, 10(3), 214–234.

ten Brinke, Leanne and Dana R. Carney (2014), "Wanted Direct Comparisons of Unconscious and Conscious Lie Detection," *Psychological science*, 25(10), 1962–1963.

ten Brinke, Leanne, Dayna Stimson, and Dana R. Carney (2014), "Some Evidence for Unconscious Lie Detection," *Psychological Science*, 25(5), 1098–1105.

Burt, Christopher DB (1999), "Categorisation of Action Speed and Estimated Event Duration," *Memory*, 7(3), 345–355.

C. Schellekens, Gaby A., Peeter W. J. Verlegh, and Ale Smidts (2010), "Language Abstraction in Word of Mouth," *Journal of Consumer Research*, 37(2), 207–23.

Changizi, Mark A. (2008), "Economically Organized Hierarchies in WordNet and the Oxford English Dictionary," *Cognitive Systems Research*, 9(3), 214–228.

Chen, Yubo and Jinhong Xie (2008), "Online Consumer Review: Word-of-Mouth as a New Element of Marketing Communication Mix," *Management Science*, 54(3), 477–91.

Davidson, Alexander and Michel Laroche (2014), "Consumer Patternicity: Investigating the Influence of Abstract Mindsets on Personal Need For Structure," *ACR North American Advances*, http://www.acrwebsite.org/volumes/v42/acr_v42_17007.pdf.

Dickson, Peter R. (1982), "The Impact of Enriching Case and Statistical Information on Consumer Judgments," *Journal of Consumer Research*, 8(4), 398–406.

Dulaney, Earl F. (1982), "Changes in Language Behavior as a Function of Veracity," *Human Communication Research*, 9(1), 75–82.

Dunning, Ted (1993), "Accurate Methods for the Statistics of Surprise and Coincidence," *Computational Linguistics*, 19(1), 61–74.

Evans, Jacqueline R. and Stephen W. Michael (2014), "Detecting Deception in Non-Native English Speakers," *Applied Cognitive Psychology*, 28(2), 226–237.

Filipović, Luna (2007), "Language as a Witness: Insights from Cognitive Linguistics," *International Journal of Speech, Language & the Law*, 14(2), 245-267.

Gokhman, Stephanie, Jeff Hancock, Poornima Prabhu, Myle Ott, and Claire Cardie (2012), "In Search of a Gold Standard in Studies of Deception," in *Proceedings of the Workshop on Computational Approaches to Deception Detection*, Association for Computational Linguistics, 23–30, http://dl.acm.org/citation.cfm?id=2388620.

Gordon, Randall A. and Uwe Stuecher (1992), "The Effect of Anonymity and Increased Accountability on the Linguistic Complexity of Teaching Evaluations," *The Journal of Psychology*, 126(6), 639–649.

Hancock, Jeffrey T., Christopher Landrigan, and Courtney Silver (2007), "Expressing Emotion in Text-Based Communication," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 929–932, http://dl.acm.org/citation.cfm?id=1240764.

Hansen, Jochim and Michaela Wänke (2010), "Truth from Language and Truth from Fit: The Impact of Linguistic Concreteness and Level of Construal on Subjective Truth," *Personality and Social Psychology Bulletin*, 36(11), 1576–1588.

Hauch, Valerie, Iris Blandón-Gitlin, Jaume Masip, and Siegfried L. Sporer (2015), "Are Computers Effective Lie Detectors? A Meta-Analysis of Linguistic Cues to Deception," *Personality and Social Psychology Review*, 19(4), 307–342.

Heath, Chip and Dan Heath (2007), *Made to Stick: Why Some Ideas Survive and Others Die*, New York: Random House,

Iliev, Rumen and Robert Axelrod (2017), "The Paradox of Abstraction: Precision Versus Concreteness," *Journal of Psycholinguistic Research*, 46(3), 715–729.

Irish, Muireann, Jody Kamminga, Donna Rose Addis, Stephen Crain, Rosalind Thornton, John R. Hodges, and Olivier Piguet (2016), "'Language of the past'–Exploring Past Tense Disruption during Autobiographical Narration in Neurodegenerative Disorders," *Journal of Neuropsychology*, 10(2), 295–316.

Juola, Patrick (1998), "Measuring Linguistic Complexity: The Morphological Tier," *Journal of Quantitative Linguistics*, 5(3), 206–213.

Jurafsky, Dan, Victor Chahuneau, Bryan R. Routledge, and Noah A. Smith (2014), "Narrative Framing of Consumer Sentiment in Online Restaurant Reviews," *First Monday*, 19(4), http://uncommonculture.org/ojs/index.php/fm/article/view/4944.

Kanwisher, Nancy G. (1987), "Repetition Blindness: Type Recognition without Token Individuation," *Cognition*, 27(2), 117–143.

Kelly, Stephen W., A. Mike Burton, Takashi Kato, and Shigeru Akamatsu (2001), "Incidental Learning of Real-World Regularities," *Psychological Science*, 12(1), 86–89.

Knapp, Mark L., Roderick P. Hart, and Harry S. Dennis (1974), "An Exploration of Deception as a Communication Construct," *Human Communication Research*, 1(1), 15–29.

Kozinets, Robert V, Kristine de Valck, Andrea C Wojnicki, and Sarah J.S Wilner (2010), "Networked Narratives: Understanding Word-of-Mouth Marketing in Online Communities," *Journal of Marketing*, 74(2), 71–89.

Krishnan, Balaji C., Abhijit Biswas, and Richard G. Netemeyer (2006), "Semantic Cues in Reference Price Advertisements: The Moderating Role of Cue Concreteness," *Journal of Retailing*, 82(2), 95–104.

Kronrod, Ann and Shai Danziger (2013), "'Wii Will Rock You!' The Use and Effect of Figurative Language in Consumer Reviews of Hedonic and Utilitarian Consumption," *Journal of Consumer Research*, 40(4), 726–39.

Lambert, Wallace E. (1955), "Associational Fluency as a Function of Stimulus Abstractness.," *Canadian Journal of Psychology (Revue Canadienne de Psychologie)*, 9(2), 103.

Li, Fangtao, Minlie Huang, Yi Yang, and Xiaoyan Zhu (2011), "Learning to Identify Review Spam," in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, 2488, http://www.aaai.org/ocs/index.php/IJCAI/IJCAI11/paper/download/3097/3721.

Li, Xinxin and Lorin M. Hitt (2008), "Self-Selection and Information Role of Online Product Reviews," *Information Systems Research*, 19(4), 456–474.

Liebes, Tamar (2001), "'Look Me Straight in the Eye' the Political Discourse of Authenticity, Spontaneity, and Sincerity," *The Communication Review*, 4(4), 499–510.

Lim, Ee-Peng, Viet-An Nguyen, Nitin Jindal, Bing Liu, and Hady Wirawan Lauw (2010), "Detecting Product Review Spammers Using Rating Behaviors," in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, ACM, 939–948, http://dl.acm.org/citation.cfm?id=1871557.

Luca, Michael and Georgios Zervas (2016), "Fake It till You Make It: Reputation, Competition, and Yelp Review Fraud," *Management Science*, 62(12), 3412–3427.

Malbon, Justin (2013), "Taking Fake Online Consumer Reviews Seriously," *Journal of Consumer Policy*, 36(2), 139–157.

Mantonakis, Antonia, Bruce WA Whittlesea, and Carolyn Yoon (2008), "Consumer, Memory, Fluency and Familiarity," *Handbook of Consumer Psychology*, 77–102.

McKimmie, Blake M., Sara A. Newton, Regina A. Schuller, and Deborah J. Terry (2013), "It's Not What She Says, It's How She Says It: The Influence of Language Complexity and Cognitive Load on the Persuasiveness of Expert Testimony," *Psychiatry, Psychology and Law*, 20(4), 578–589.

Miyata, Hiromitsu, Yasuyo Minagawa-Kawai, Shigeru Watanabe, Toyofumi Sasaki, and Kazuhiro Ueda (2012), "Reading Speed, Comprehension and Eye Movements While Reading Japanese Novels: Evidence from Untrained Readers and Cases of Speed-Reading Trainees," *PloS One*, 7(5), e36091.

Moore, Sarah G. (2012), "Some Things Are Better Left Unsaid: How Word of Mouth Influences the Storyteller," *Journal of Consumer Research*, 38(6), 1140–1154.

Mukherjee, Arjun, Abhinav Kumar, Bing Liu, Junhui Wang, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh (2013), "Spotting Opinion Spammers Using Behavioral Footprints," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 632–640, http://dl.acm.org/citation.cfm?id=2487580.

Nelson, Laura (2017), "Computational Grounded Theory: A Methodological Framework." *Working Paper under review.*

Newman, Matthew L., James W. Pennebaker, Diane S. Berry, and Jane M. Richards (2003), "Lying Words: Predicting Deception from Linguistic Styles," *Personality and Social Psychology Bulletin*, 29(5), 665–675.

Ordenes, Francisco Villarroel, Babis Theodoulidis, Jamie Burton, Thorsten Gruber, and Mohamed Zaki (2014), "Analyzing Customer Experience Feedback Using Text Mining: A Linguistics-Based Approach," *Journal of Service Research*, 17(3), 278–295.

Ott, Myle, Claire Cardie, and Jeff Hancock (2012), "Estimating the Prevalence of Deception in Online Review Communities," in *Proceedings of the 21st International Conference on World Wide Web*, ACM, 201–210, http://dl.acm.org/citation.cfm?id=2187864.

Ott, Myle, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock (2011), "Finding Deceptive Opinion Spam by Any Stretch of the Imagination," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, Association for Computational Linguistics, 309–319, http://dl.acm.org/citation.cfm?id=2002512.

Paivio, Allan (1963), "Learning of Adjective-Noun Paired Associates as a Function of Adjective-Noun Word Order and Noun Abstractness," *Canadian Journal of Psychology (Revue Canadienne de Psychologie)*, 17(4), 370.

Pinker, Steven (2013), *Language, Cognition, and Human Nature: Selected Articles*, Oxford, England: Oxford University Press.

Plous, Scott (1993), *The Psychology of Judgment and Decision Making*, Mcgraw-Hill Book Company, http://psycnet.apa.org/psycinfo/1993-97429-000.

Rossiter, John R. and Larry Percy (1978), "Visual Imaging Ability As a Mediator of Advertising Response", in *NA - Advances in Consumer Research* Volume 05, ed. Kent Hunt, Ann Abor, MI: Association for Consumer Research, Pages: 621-629.

Sadoski, Mark, Ernest T. Goetz, and Joyce B. Fritz (1993), "Impact of Concreteness on Comprehensibility, Interest, and Memory for Text: Implications for Dual Coding Theory and Text Design," *Journal of Educational Psychology*, 85(2), 291.

Saslow, Laura R., Shannon McCoy, Ilmo Löwe, Brandon Cosley, Arbi Vartan, Christopher Oveis, Dacher Keltner, Judith T. Moskowitz, and Elissa S. Epel (2014), "Speaking under Pressure: Low Linguistic Complexity Is Linked to High Physiological and Emotional Stress Reactivity," *Psychophysiology*, 51(3), 257–266.

Skinner, Christopher H., Jacqueline L. Williams, Jennifer Ann Morrow, Andre D. Hale, Christine E. Neddenriep, and Renee O. Hawkins (2009), "The Validity of Reading Comprehension Rate: Reading Speed, Comprehension, and Comprehension Rates," *Psychology in the Schools*, 46(10), 1036–1047.

Sokolowski, Wojciech L. (1977), "Structure and Characteristics of the Individually Used Language and Adequacy of the Reality Description in Statements of Witnesses," *Przegląd Psychologiczny*, 20(3), 503-521.

Streitfeld, David (2012), "Fake Reviews, Real Problem," *New York Times.*(http://query.nytimes.com/gst/fullpage.html?res=9903E6DA1E3CF933A2575AC0A9649D8B63, Accessed on June 29 2017).

Takashima, Atsuko, Iske Bakker, Janet G. Van Hell, Gabriele Janzen, and James M. McQueen (2017), "Interaction between Episodic and Semantic Memory Networks in the Acquisition and Consolidation of Novel Spoken Words," *Brain and Language*, 167, 44–60.

Tendolkar, Indira (2008), "How Semantic and Episodic Memory Contribute to Autobiographical Memory. Commentary on Burt," *Language Learning*, 58(s1), 143–147.

Tulving, Endel (1985a), "Elements of Episodic Memory," https://philpapers.org/rec/TULEOE.

Tulving, Endel (1985b), "How Many Memory Systems Are There?" *American Psychologist*, 40(4), 385.

Tversky, Amos (1982), *Judgments of and by Representativeness, 84-98. D. Kahneman, P. Slovic, and A. Tversky (Eds).* Judgment under Uncertainty: Heuristics and Biases, New York: Cambridge University Press.

Van Swol, Lyn M., Michael T. Braun, and Miranda R. Kolb (2015), "Deception, Detection, Demeanor, and Truth Bias in Face-to-Face and Computer-Mediated Communication," *Communication Research*, 42(8), 1116–1142.

Villar, Gina, Joanne Arciuli, and Alessio Barsaglini (2010), "Can Reduced Use of Pronouns during Deceptive versus Truthful Speech Be Observed in a Language Other than English?," in *13th Australasian International Conference on Speech Science and Technology*, http://assta.org/sst/SST-10/SST2010/PDF/AUTHOR/ST100003.PDF.

Vrij, Aldert (2008), "Nonverbal Dominance versus Verbal Accuracy in Lie Detection: A Plea to Change Police Practice," *Criminal Justice and Behavior*, 35(10), 1323–1336.

Vrij, Aldert, Katherine Edward, and Ray Bull (2001), "Stereotypical Verbal and Nonverbal Responses While Deceiving Others," *Personality and Social Psychology Bulletin*, 27(7), 899–909.

Vrij, Aldert, Pär Anders Granhag, Samantha Mann, and Sharon Leal (2011), "Outsmarting the Liars: Toward a Cognitive Lie Detection Approach," *Current Directions in Psychological Science*, 20(1), 28–32.

Whissell, Cynthia (1999), "Linguistic Complexity of Abstracts and Titles in Highly Cited Journals," *Perceptual and Motor Skills*, 88(1), 76–86.

Wiener, Morton and Albert Mehrabian (1968), *Language within Language: Immediacy, a Channel in Verbal Communication*, New York: Appleton-Century-Crofts, Century Psychology Series.

Wu, Guangyu, Derek Greene, Barry Smyth, and Pádraig Cunningham (2010), "Distortion as a Validation Criterion in the Identification of Suspicious Reviews," in *Proceedings of the First Workshop on Social Media Analytics*, ACM, 10–13. http://dl.acm.org/citation.cfm?id=1964860.

Zhao, Yi, Sha Yang, Vishal Narayan, and Ying Zhao (2013), "Modeling Consumer Learning from Online Product Reviews," *Marketing Science*, 32(1), 153–169.

APPENDIX A: LINGUISTIC FEATURES OF INSINCERE TEXT

Feature	References
Less pronouns	Villar, Arciuli, and Barsaglini (2010)
Less use of first person pronouns (“I”/”we”)	Hancock et al. (2008); Li, Huang, Yang and Zhu (2011); Newman, Pennebaker, Berry and Richards (2003); Toma and Hancock (2010)
More first person pronouns (“I”/”we”)	Berzack 2011
More words	Berzack (2011); Hancock, Curry, Goorha and Woodworth (2008); Van Swol and colleagues (2012a, 2012b)
Less words	Li, Huang, Yang and Zhu (2011)
More negations	Hancock et al. (2008) and Toma and Hancock (2010)
More negative emotion words	Anderson and Simester (2014); Newman, Pennebaker, Berry and Richards (2003)
More extreme	Li, Huang, Yang and Zhu (2011); Luca and Zervas (2013)
More sense words (see/touch)	Hancock et al. (2008) and Toma and Hancock (2010)
Less causal terms	Hancock et al. (2008) and Toma and Hancock (2010)
More facts	Liebes (2001)
Reduced adjectives	Villar, Arciuli, and Paterson (2013)
More conjunctions	Berzack 2011
Less conjunctions	Newman, Pennebaker, Berry and Richards (2003)
Similarity to language in other reviews of the same reviewer	Li, Huang, Yang and Zhu (2011)
More “um” in speech	Arciuli, Mallard, and Villar (2010)
Too slow or too fast response in speech	Baskett and Freedle 1974

TECHNICAL APPENDIX

The purpose of this technical appendix is to provide readers with specific tools and analysis approaches that were used to arrive at the conclusions presented in the paper. A secondary goal of this appendix is to elaborate on certain ideas regarding our approach to text analysis.

1. Technical Details for Text Pre-processing

As the first step, we created a "Reviews" table in an SQLite database, where we parsed all the data from the original reviews dataset. We then pre-processed the original reviews, by (1) removing all characters other than letters or punctuation and (2) correcting spelling mistakes.

Cleaning and Spellchecking

To clean up the text we used regular expressions ('re' library in Python), which allowed us to remove all characters other than letters and punctuation. Then, we corrected spelling mistakes with the Python PyEnchant library. The spelling algorithm goes word by word through the text and checks if each word is not in the library, i.e. it is misspelled. If the word is misspelled, the algorithm then uses a function built into the PyEnchant library to generate a list of alternative words from the dictionary, ordered from most likely as a viable replacement to least likely (http://pythonhosted.org/pyenchant/tutorial.html). The first word in this list is then selected as replacement for the misspelled word, and inserted into the text instead of the misspelled word. This algorithm is simple and efficient in terms of speed, and has been shown to yield an accuracy rate of approximately 70% (Sosamphan et al. 2016).

NLTK

Another Python tool we used extensively is the Natural Language Toolkit (http://www.nltk.org), or simply NLTK. The packages we used provide text tokenization (such as splitting the text into sentences and words), operations on words (stemming, lemmatization), parts-of-speech (POS) tagging and a WordNet (http://wordnet.princeton.edu/) interface.

Extraction of first person pronouns

To count the number of first person pronoun occurrences in our texts, we generated a new table in our SQLite database named "Word forms". We populated this table with counts for each word form[12], for each condition. Note that it is also possible to use POS-tagging directly to extract and count first person pronouns, but instead we analyzed the first person pronoun counts directly from our table of all word forms.

[12] A word form is any unique sequence of letters with a meaning, separated from other such sequences by one or more space characters on either side. For example, 'empty' is one word form, and 'emptiness' is another word form. Similarly, 'dog' and 'dogs' are two separate word forms.

Extraction of Verbs and Tenses

To count the past tense verbs in our text, we used the POS-tagger developed within the NLTK platform. This tagger allowed us not only to extract verbs, but also to extract particular tenses of verbs. We counted the occurrences of different verb forms, and used counts of past simple verbs as measures of past tense occurrences in our analysis.

Extraction of Nouns

To extract the nouns in each review, we generated an SQLite table of "Nouns", looped through each review, tagged all part of speech occurrences using the NLTK POS-tagger, and added all nouns with their corresponding number of occurrences values into the table. We then counted the number of nouns in each review and added those values to the previously created table "Reviews".

This exercise yielded a final list of 3070 noun entries in the table. Among these entries, we spotted a few POS-tagging mistakes (some words were recognized as nouns by the NLTK POS tagger, but were not nouns in fact). We corrected this issue by checking whether a word is indeed a noun[13], before adding it to the table. Rerunning the code with this correction left us with a final count of 2937 unique nouns in the full corpus of reviews.

Spelling correction is one of the most time-consuming functions in our code. Thus, to make our process efficient, we completed the data cleaning for all the reviews and then proceeded with analyzing the pre-processed texts[14].

2. Definition and Extraction of Unique Words

Another feature of reviews we considered in this work is what we called 'uniqueness'. The hypothesis in the work was that the usage of unique words in authentic reviews is more substantial than in fictitious reviews or in reviews with a uniqueness clue. In order to test this hypothesis we first constructed the distributions of word forms in each of the three conditions

[13] To check if a word is a noun or not we used the WordNet dictionary (Princeton 2010),

[14] Future improvements to text cleaning: There are many limitations to the spell checking algorithm which we described above. To address some of them one could design a "smarter" cleaning algorithm by taking care of specific types of issues individually, such as dealing with numerical symbols, abbreviations, slang and repeating characters (such as in the following example of spelling of the word 'love': 'loooove').

Most importantly, one could invest effort in designing a much smarter spelling correction algorithm. One important improvement would be taking into account keyboard layout (Deorowicz and Ciura 2005). The basic assumption of this approach is that people often simply mistype characters, i.e. they replace the character they would actually wish to type with another one located next to it on the keyboard. One could introduce a weighing system, so that the characters located closer to the mistyped one would have a higher probability to substitute it (unlike in edit distance based algorithms, where all characters are considered equivalent). Combining such a weighing method with edit distance based algorithm (by also assigning weights to different edit distance values) could give a much better accuracy of spelling correction.

Of course keyboard-typing errors are not the only possible spelling mistakes. One could simply make orthographical errors and we could further improve the algorithm if we know the common spelling mistakes. A tentative list of improvements one could do to make the spellchecking more accurate can be found for instance on Peter Norvig's website (Norvig 2016).

over the frequency of those word forms[15]. The different word forms distributed according to a power law (the factor was close to -1.65).

Analysis of the Pattern of Recurrence of All Word Forms

We developed a code in Python which runs through all the reviews and counts the recurrence of different word forms in the entire corpus. We found a certain number of word forms that occur only once in the whole corpus, a certain number of word forms that recur twice, etc. We then counted how many occurrences of each frequency are observed in each condition. Subsequently, we compared the occurrences of 'rare' words in the authentic, fictitious, and unique-words-clue conditions, as follows. We used Matlab to perform most of the analysis and calculations beyond text pre-processing and generating the original SQLite database with word form counts as described above.

Definition of 'rare' words

For the analysis to be meaningful one has to define how exactly one is going to 'measure' the 'rarity' of a particular word form. A preferable way is to define rarity in a condition-independent way. Then one can easily compare between different conditions. Namely, one could take a certain corpus (it could be any corpus in fact) and define the rarity of each word based on its occurrence in this corpus, divided by the number of words in the corpus. To have a more robust definition one should try to have a larger corpus. Using a larger corpus as the basis to assign a value of frequency to every word form will lead to a finer step size between the discrete frequencies possible. Note that defining the frequency of a word form in this way, one obtains a range of possible frequencies with the most rare words appearing only once (corresponding frequency value is 1/N) in the whole corpus[16], second most rare appearing exactly two times (corresponding frequency value is 2/N) etc. Thus, the size of the step between adjacent frequencies will be 1/N, where N is the number of words in the whole corpus. The larger the corpus the finer the structure of 'rarity', as the minimal difference in frequencies that this way of assigning frequency values resolves is 1/N and this number becomes smaller as N grows.

It also makes sense to use a corpus relevant to the topic at hand. For instance, rather than using a larger, more general corpus, it might be preferable to use a corpus relevant to hotel or hotel reviews, given the type of data we explore in this work. If we use a corpus outside of our domain, it may have a vocabulary that is distinct from ours; thus, words which are rare in that corpus could be common in our domain, and vice versa.

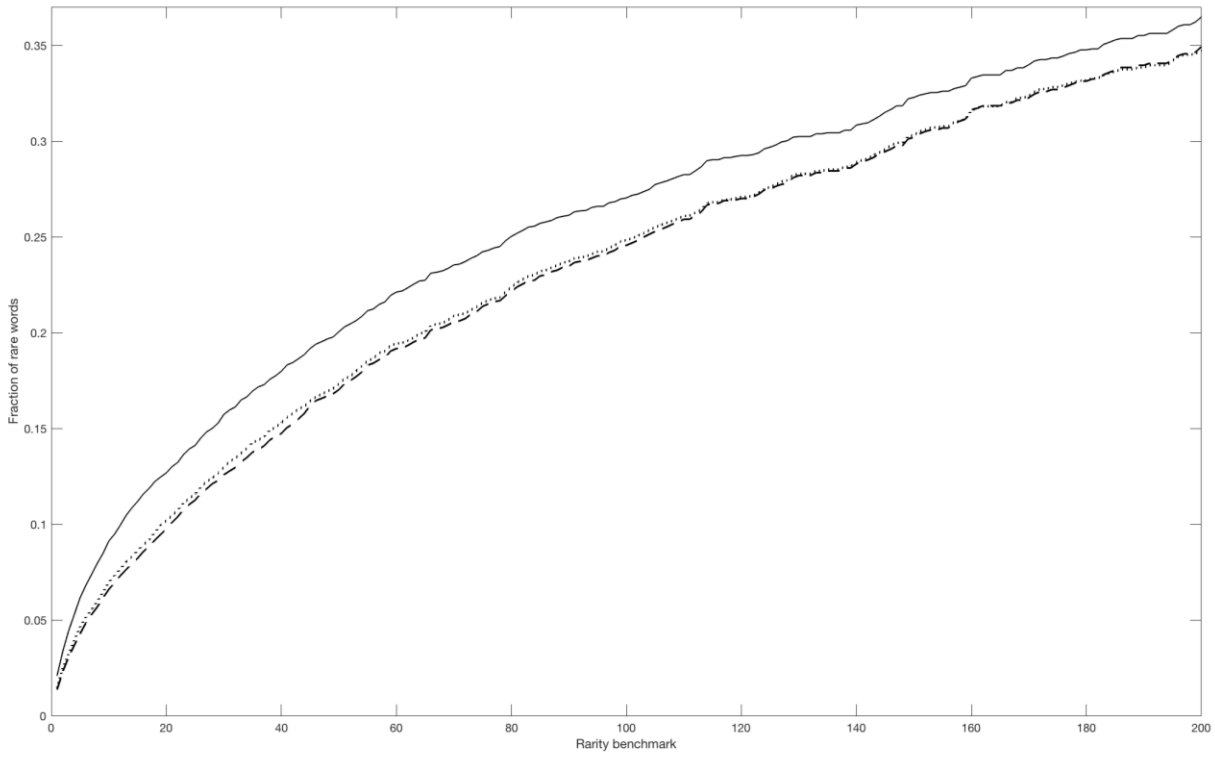
Given this potential issue, we concluded that it was most reasonable for us to use our entire corpus collected in study 1 as the basis for defining the way of assigning a value to the rarity of each word form. We apply this same corpus in defining the rarity, across all conditions.

There is no particular value of frequency which can be determined as the only possible benchmark to distinguish between rare and not rare words. The choice of such a benchmark is to

[15] We can easily expand this analysis onto any subset of the complete corpus, for instance limit it to any particular part of speech. For example, we separately considered only nouns and then all word forms. Results were similar and somewhat stronger for nouns than for all word forms. This difference was not the focus of this paper, however.
[16] Note that it is a common feature of texts that a large portion of the words appear in the corpus only once (e.g. Dunning 1993).

a certain extent arbitrary and up to the researcher. Our goal was to test how many words in each condition were rarer than a certain benchmark value of rarity. To determine the benchmark for rare words, we conducted the following calculation: There were 181,144 words in the complete corpus, and among them there were 6820 unique word forms. Therefore on average each unique word form repeated approximately 27 times in the whole corpus. We considered word forms which repeated less than 27 times in the complete corpus as relatively rare, and words that repeated 27 times or more as frequent. Figure 1 portrays the fraction of rare word forms as a function of the rarity benchmark. The dotted lines represent the two fictitious reviews conditions and appear to be significantly lower than the solid line of the authentic condition.

Figure 1: Fraction of Rare Word Forms as a Function of the Rarity Benchmark



Following this quantitative definition of rarity for each word form, we counted the proportion of rare words in each condition[17]. That is, in each condition we counted the proportion of word forms that occurred less than 27 times in the whole corpus. We found that there was a significantly larger proportion of rare words in authentic reviews than in fictitious reviews. To evaluate if the difference in proportions between the conditions is significant, we

[17] We did not divide the frequencies we found by the number of reviews for the following two reasons: 1. One could divide by the number of reviews within each condition, however one would have to divide both the vocabulary size and the total word count, so the ratio of total number of words to the vocabulary size will not be affected. Thus, the division will not matter for the validity of the conclusion. 2. While dividing the total number of words by the number of reviews does make sense (it gives the average word count per review), dividing the vocabulary size does not have much meaning (it is not equal to the average vocabulary size for review.) Generally while the total count of words increases linearly with the number of reviews, the vocabulary size does not.

calculated the Z-score for that difference. Results for this comparison are reported in the main paper.

Notably, we noticed the following pattern: the vocabulary size in the authentic condition was smaller than the vocabulary size in both fictitious conditions, while the opposite was true for the total number of terms (be it nouns or word-forms). This implies that the language used in authentic reviews was more diverse, whereas in the fictitious reviews the vocabulary size was smaller and there were more repetitions of the same terms.

Potential Improvements for Analyzing Unique Words

Another observation we made was that for many word forms, there were noticeably more instances of that word form in the fictitious condition than in the authentic reviews. Nonetheless, for some other word forms we find the opposite pattern, with those word forms being more frequent in the authentic conditions. Additional research could explore the specific types of word forms for which this opposite pattern exists, as there may be a clear thematic distinction between these word forms and others in our corpus. Also, one could analyze other parts of speech (e.g. verbs, adjectives) and compare results for these language features to our findings for nouns. Another improvement on our current method can be performed by lemmatizing every word (rather than stemming words), so that all different words are considered unique, rather than different word *forms*[18].

Definition and Calculation of Review Concreteness

In this section we describe how we defined the concreteness of a given text and how we calculated a concreteness score for each review. We then compare the distributions of review concreteness values within and between conditions. For technical reasons which will be explained later in this section, we exclusively focused on nouns for the definition and calculation of concreteness.

Defining the depth of a noun and calculating text concreteness

We relied on previous research (e.g. Changizi 2008; Iliev and Axelrod 2017; Nelson 2016) to define the 'depth' of each noun in each review. Specifically, we leveraged the location of each noun in the hierarchy of nouns within the WordNet Dictionary (Princeton 2010). WordNet 3.1 classifies every noun into a hierarchy ranging from one single ancestor, which is the word "entity". *Entity* sits at the very top of the hierarchy as the most general term in the dictionary of nouns. At lower levels of the hierarchy, there are more specific descendant terms. Terms higher than a given noun are known as hypernyms and terms that are lower in the hierarchy are known as hyponyms.

We limit our analysis to nouns because the hierarchical structure in WordNet is particularly consistent and well defined for nouns. Specifically, there is one single word ("entity") at the top of the hierarchy (as opposed to complications such as the presence of several

[18] Lemmatization uses each configuration of a word as a unique term, whereas stemming collapses all forms of a stem into one. For example, the words "ugly", "uglier" and "ugliness" would be represented as 3 different terms after lemmatization, but as a single term after stemming.

superior hypernyms in the verbs hierarchy, e.g. Richens 2008). As a result, for a given noun we had a straightforward way to define its depth. Namely, we defined this value as the distance in the hierarchy from the most general term ("entity") to that particular noun, with each level of descendant counting as one additional unit of depth. Higher values of depth correspond to deeper (or more specific) nouns, and lower values correspond to more general nouns (less steps from the most general word "entity").

Next, we defined the concreteness of a given text (i.e. a collection of words, such as a review) based on the individual depth values of each noun in that text. We proceeded using a statistical approach which expands the logic of depth to a collection of words. Specifically, we used the combinatorial formula (1) below to calculate the number of texts that we could create using hypernyms (more general nouns) in the WordNet hierarchy, in place of the given text. The formula is as follows:

(1) $$\prod \binom{d+f}{f} = \prod \frac{(d+f)!}{f!d!}$$

where d is the depth of the noun, f represents the number of times the noun occurs in the text and the product iterates over all identified nouns in the given text. The result of this calculation is the number of all possible texts that are more general than the given text, using alternative hypernyms in the WordNet hierarchy.

Next, to handle the large values we received, we took the natural logarithm of the result in formula 1 (see formula 2). This quantity is similar to entropy as defined in complex systems. We defined this entropic quantity as the concreteness of the text. We used the result of these calculations in our analyses.

(2) $$D = log \prod \binom{d+f}{f}$$

Possible Extensions for the analysis of concreteness

There are a number of possible ways to extend this analysis. Additional research could explore the inclusion of other parts-of-speech (POS) in the analysis. The hypernym structure of verbs in WordNet has been studied in other research, and there have been proposals on how to make this structure more consistent (Richens 2008). Subsequent research could also define the concreteness of a text in a different way, as there is mounting research defining concreteness in various ways. However, we did not consider the different definitions of concreteness in this work, as this is beyond the scope of this work.

Technical Appendix References

Deorowicz, Sebastian, and Marcin G. Ciura. (2005) "Correcting spelling errors by modelling their causes." *International journal of applied mathematics and computer science*, 15, 275-285.

Norvig, Peter (2016). http://norvig.com/spell-correct.html (accessed on June 26 2017).

Richens, Tom. (2008). "Anomalies in the WordNet verb hierarchy." *Proceedings of the 22nd International Conference on Computational Linguistics*, Volume 1, pp. 729–736. Association for Computational Linguistics, 2008. (accessed online on June 20 2017: http://www.aclweb.org/anthology/C08-1092).

Sosamphan, Phavanh, Veronica Liesaputra, Sira Yongchareon and Mahsa Mohaghegh (2016). Evaluation of Statistical Text Normalisation Techniques for Twitter. *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, 413-418, 2016, Porto, Portugal.