http://dx.doi.org/10.7592/EJHR2014.2.3.skalicki

European Journal of
Humour Research 2 (3) 66–85
www.europeanjournalofhumour.org

A statistical analysis of satirical Amazon.com product reviews

Stephen Skalicky
Scott Crossley
Georgia State University, USA

Abstract

A corpus of 750 product reviews extracted from Amazon.com was analyzed for specific lexical, grammatical, and semantic features to identify differences between satirical and non-satirical Amazon.com product reviews through a statistical analysis. The corpus contained 375 reviews identified as satirical and 375 as non-satirical (750 total). Fourteen different linguistic indices were used to measure features related to lexical sophistication, grammatical functions, and the semantic properties of words. A one-way multivariate analysis of variance (MANOVA) found a significant difference between review types. The MANOVA was followed by a discriminant function analysis (DFA), which used seven variables to correctly classify 71.7 per cent of the reviews as satirical or non-satirical. Those seven variables suggest that, linguistically, satirical texts are more specific, less lexically sophisticated, and contain more words associated with negative emotions and certainty than non-satirical texts. These results demonstrate that satire shares some, but not all, of the previously identified semantic features of sarcasm (Campbell & Katz 2012), supporting Simpson's (2003) claim that satire should be considered separately from other forms of irony. Ultimately, this study puts forth an argument that a statistical analysis of lexical, semantic, and grammatical properties of satirical texts can shed some descriptive light on this relatively understudied linguistic phenomenon, while also providing suggestions for future analysis.

Keywords: satire; statistical analysis; online product reviews.

1. Introduction

This paper aims to study the lexical, grammatical, and semantic properties of written satire through a statistical analysis that uses an array of computational and linguistic measurements to account for the linguistic features of satire. While similar studies have examined related forms of ironic language use (Kreuz & Caucci 2007; Caucci & Kreuz 2012; González-Ibáñez et al. 2011; Hancock 2004), the focus here is directed specifically towards satire. A comprehensive literature review found only one other study that has analysed satire via statistical and computational analysis (Burfoot & Baldwin 2009); indeed, satire is relatively understudied in non-literary disciplines in general (Simpson 2003). As such, there is an existing gap in the research that this study addresses. Specifically, a detailed statistical analysis of satirical texts can contribute empirical verification for a theoretical model of satire, while also considering satire in relation to other forms of ironic language use.

To do so, we use a statistical model informed by computational measurements of the lexical, grammatical, and semantic properties of written satire. First, we determine which (if any) linguistic features (represented as numeric measurements) differ between the satirical and non-satirical texts at a level greater than chance (using multivariate analyses of variance); we then use those results to train a statistical model to automatically identify satirical and non-satirical texts (using discriminant function analysis). If our results reach certain statistical thresholds, this suggests that written satire contains measurable linguistic differences, as compared to written non-satire.

The corpus for this analysis is composed of product reviews written on Amazon.com. Because of their recognizable nature (due to media coverage), satirical product reviews from Amazon.com provide a unique and ideal candidate for an analysis of this nature (Popova n.d.; Oremus 2013). Unlike marked forms of satire that are produced by purely satirical outlets such as *The Onion*, satirical product reviews written on Amazon.com co-occur with their non-satirical counterparts in the same environment, allowing for a cross-comparison that holds factors such as genre and medium constant.

1.1. Satire

Defining satire is difficult (e.g. Condren 2012), because it may be conflated with verbal irony or treated as a static type of writing. Typically, satire has been characterised as a literary genre and is studied from the perspective of literary criticism (Simpson 2003; Nilsen & Nilsen 2008). However, this view of satire relies on categorisation of individual texts based on subjective arguments and not any clear criteria conducive to formal linguistic analyses. From a linguistic perspective, Simpson (2003) provides the most cogent definition of satire along with a method for identifying and analysing different instances of satire. While recognizing the role of literary definitions, Simpson (2003) argues that satire is a form of humour that uses ironic means to achieve its goals and has received relatively little scholarly attention from non-literary disciplines.

According to Simpson (2003: 10), satire is a three-stage discursive practice involving three participants: the author, the audience, and the target of a satirical text. A satirical text operates by

evoking a previous discourse event or entity (the prime stage) and then produces a text-internal "collision of ideas" that signals an incongruity (the dialectical stage) between the form of the text and the message of the text. Recognition of this incongruity, which requires specific cultural and genre knowledge, is required for the third stage of the satirical process, the uptake. An uptake that resolves the incongruity between the prime stage and the dialectical stage results in humour, if the audience is sympathetic to the underlying satirical message at the heart of the text.

This study is concerned with the second stage of Simpson's (2003: 9) model, in particular, "the linguistic means used by the satirist to create both prime and dialectical elements of structure in a piece of satire". According to Simpson, two related strategies are used in the satirical method: metonymy and metaphor. Metonymic strategies work to inflate (i.e. saturate) or deflate (i.e. attenuate) perceptions of a satirical target, or to highlight a situation or activity that did *not* happen (i.e. negate), inviting a consideration of the alternative to the negative. Finally, metaphorical strategies involve comparisons to other entities outside of the content domain of the satirical target.

Because the second stage of Simpson's (2003) model is primarily ironic, it is important to consider theoretical understandings of irony and irony processing. Being ironic can serve specific communicative goals and functions (Jorgensen 1996; Gibbs 2000), sometimes in more efficient ways than non-ironic speech (Kreuz et al. 1991). Irony is best defined as a difference between what is said and what is meant, with that difference prompting a hearer to resolve apparent incongruities in an utterance. Attardo (2000) argued for irony as relevant inappropriateness: the meaning behind an ironic utterance is indirect, but the utterance itself still bears some relevance to the context in which it was made (see Colston & Gibbs 2007, for a thorough review of irony).

1.2. Amazon.com product reviews as satire

The corpus for this analysis is composed of product reviews written on Amazon.com. The prototypical communicative purpose of Amazon.com product reviews is to provide experience-based information to potential consumers in order to aid them in making a purchasing decision (Skalicky 2013). This view is complicated by the relatively recent trend of the humorous Amazon review (Popova n.d), of which thousands of examples now exist. Indeed, according to the model of satire put forth by Simpson (2003), the humour in these reviews is satirical, because the reviews create a dissonance between the putative purpose (aid in a purchasing decision) of the review and the function (critique, mock) of the language in the review (Simpson 2003), while simultaneously masquerading as legitimate product reviews. In other words, these funny reviews are satirical because they are pretending to be something they are not; they employ irony to elicit a meaning different from the surface form of the text; and the elicited meaning is humorous because it critiques or mocks some other entity, directly or indirectly. There are, of course, many more product reviews that *are* intended to meet the typical communicative purpose of a product review that are non-satirical. This provides a convenient corpus of satirical and non-satirical texts that are, on their face, members of the same genre, but may differ beneath, in terms of linguistic features. Differences in these linguistic features may help to better explain the second stage of Simpson's (2003) model of satire.

In order to better demonstrate the different ways satire is performed in Amazon.com reviews, examples from our corpus serve well. The following reviews all come from the product

page for a line of *BIC* brand pens that are billed as "for her" and sold in bright pink and purple pastel colours. This product has attracted a large number of satirical reviews due to the themes of traditional gender stereotypes that it assumes (i.e. the presumption that females need a more feminine pen). For instance, consider the review below, an example of metonymic saturation:

> "Finally!,
> Being a girl--well, okay, I'm 50, but I find "woman" to be so militant and feminist--I was all atwitter to see these new Bic for Her pens! I mean, I could barely even lift those big old man pens! With Bic for Her, writing is so much more fun too! Before, anytime I picked up a pen (if I could!) all I could write about was cage fighting, cars and porno. Ew. Now, all I ever write about are flowers, unicorns and Michael Buble. I just wish other companies could be as caring and sensitive as Bic... When will Heinz come out with Ketchup for Her (pink, of course)? Where is Charmin for Her (my lady parts practically shrivel up every time I use unisex toilet paper!). And why is there no iMac for Her? A specially designed computer (pink, of course) with only the features ladies need: shopping, emailing pictures of kittens and visiting michaelbuble.com. Oh well, a girl can dream..." (Review 44S)

The author takes on the familiar trope of finally locating a product that satisfies previously unmet needs, evoking the prime of a legitimate product review. The dialectical is created through the author's exaggerated claims about the topics she was able to write about. By listing other hypothetical "for her" products that all conjure stereotypes related to female delicacy, the author is providing *more of the same* (i.e. saturation) to demonstrate the ridiculous nature of the product and signal incongruity.

The opposite of saturation is attenuation (i.e. deflating by leaving things unsaid), but it was still used to bring to mind the same gender stereotypes as the previous review. In the following excerpt, the author uses both attenuation and negation strategies to create a similar prime of explaining the defects of a particular product, but does so in a way that leaves negative stereotypes unsaid (i.e. that men are better than women at math); incongruity is realised more implicitly here, as the author leaves the reader to associate those stereotypes with the narrative being presented.

> "The pens don't work for Math!,
> I am a female AP and Multivariable Calculus teacher and I prefer to use ink when solving problems. I guess, not surprisingly, these pens cannot be used to do math problems more complicated than 5th grade level. When trying to find a derivative or definite integral, the ball point simply stopped working. I went back and added some numbers and it was fine. I progressed up to solving quadratic equations and the ball point started to "stick" so that I couldn't solve the problem completely…" (Review 40S, excerpt)

Negation is realised in the above example by constructing a reality in which something isn't happening – prompting the reader to consider the alternative to the negative: that any pen should be capable of doing math.

The next review demonstrates metaphoric satire. The prime is the same (a legitimate product review), but the dialectical is realized through connections to advertisements of feminine hygiene products through the use of "all month long" and images of women being active and outdoors without worry.

> "FINALLY!,
> Someone has answered my gentle prayers and FINALLY designed a pen that I can use all month long! I use it when I'm swimming, riding a horse, walking on the beach and doing yoga. It's comfortable, leak-proof, non-slip and it makes me feel so feminine and pretty!" (Review 31S, excerpt)

The satirical interpretation for all these reviews requires the reader to understand that these gender stereotypes are outdated and offensive in modern western society, and that the author is in fact mocking the marketing campaign behind these pens (and the pens themselves), construing *BIC* as a sexist company. Tones of exaggeration, implicit mocking, gender-marked topics, and previous experiences with the products are common to these examples, and they all dress themselves in the trappings of product reviews. If these messages were delivered in a manner different from a product review, they might be considered ironic, or humorous, but never satirical. Without their satirical garb, the specific uses of irony in these texts would change into sarcasm or other forms of related irony. Previous research into sarcasm, satire, and irony helps illuminate the different ways that these related forms of language use have been measured and described from a variety of different research perspectives. One benefit of this previous research is that it provides previous linguistic measurements that we take up in the current analysis.

2. Related research into irony and humour

Research into satire and related forms of irony has primarily occurred through descriptive analysis (corpus and psycholinguistic) or automatic computational detection studies. Descriptive studies have focused on irony in an attempt to locate specific linguistic cues that differentiate irony from other functions of language. The differences in these studies in terms of operational definitions, methods of analysis, and theoretical approaches attest to the difficulty in analysing satire and related forms of irony.

For instance, Hancock (2004) examined differences in verbal irony use between face-to-face and computer mediated communication (CMC) environments, finding that irony (mostly, sarcasm) was more common in CMC settings and primarily signalled through punctuation. Kreuz & Caucci (2007) compiled a corpus of 100 fictional utterances extracted from books (signalled by the phrase "said sarcastically") in order to identify lexical features of sarcasm. Human raters judged the level of sarcasm in the sentences; the sentences were also coded for the presence of intensifiers (e.g. adjectives, adverbs), interjections (e.g. "gee"), question marks, and exclamation points. The results suggested that only interjections played a significant role in raters' perception of sarcasm. Follow-up studies using this data found that combining all of these features increased the likelihood that an excerpt would be rated as sarcastic (Kreuz & Caucci 2008). Whalen et al. (2009) investigated emails between friends for use of what they called non-literal language (a correlate of irony), which they explained to be hyperbole, understatement, rhetorical questions,

sarcasm, and jocularity. They coded the emails for discourse markers based on Hancock's (2004) findings of verbal irony in CMC, finding that 95 per cent of the emails contained non-literal language, with hyperbole being the most common type.

Campbell & Katz (2012) analysed texts created by human participants that were designed to elicit a sarcastic or open interpretation. Using the Linguistic Inquiry and Word Count (LIWC) program, which measures semantic and grammatical properties of individual words (Pennebacker et al. 2007), they found reliable differences in human ratings of sarcasm between texts that invited sarcasm and those that invited an open interpretation for seven out of thirteen LIWC indices. Specifically, sarcastic texts included higher levels of negative emotions, emphasis, clarification, temporal markers, sadness, and causation.

In contrast, automatic detection studies of irony, humour, or satire typically start with a list of indices based on theoretical and a priori assumptions to create and test a model that may be able to automatically differentiate between types of texts. Mihalcea & Strapparava (2006) reported successful results for automatically detecting humorous one-liners (e.g. puns) using stylistic features such as alliteration, antonymy, slang, and content-based features. Their results suggested that irony resulted in humour in approximately half of their data, with other factors such as ambiguity, incongruity, idiomatic expressions, and commonsense knowledge accounting for the remainder of the data.

Only one other study that we know of has explicitly mentioned satire as the target of analysis. Burfoot & Baldwin (2009: 161) sought to examine satire in a corpus of news texts taken from the internet, defining satire as language that "deliberately exposes real-world individuals, organisations and events to ridicule". They focused on the news headlines, inclusion of slang and profanity, and the validity of the news story (measured by comparing frequency of content) and found that these factors were able to classify satirical from non-satirical texts. Classification accuracy ranged from 0.781 to 0.798; the authors argued that subtle cues requiring more "detailed knowledge" likely accounted for the error rates in their model.

Reyes & Rosso (2011) used a corpus similar to the analysis found in this paper (3000 review comments take from five products on Amazon.com). They defined these reviews as ironic rather than satirical. Their model was based on six factors: n-grams (recurrent word combinations), POS n-grams (recurrent part-of-speech combinations), words with semantic characteristics of sexuality or relationships (using WordNet values), positive and negative values of words (using values from the Macquarie Semantic Orientation Lexicon), affective words demonstrating subjectivity (using WordNet values), and pleasantness values of words (using values from the Dictionary of Affect in Language). They explained that these indices were chosen, based on previous research into the automatic detection of humour (e.g. Mihalcea & Strapparava 2006). The authors argued that their model was able to identify the ironic texts satisfactorily (classification scores ranged from 0.703 to 0.782), with the POS n-grams and pleasantness rating being the most important factors.

This review of the literature shows that an interest exists in measuring the linguistic properties of ironic texts, but that developing definitions specifically for irony, sarcasm, and satire has not been universally successful, most likely due to the inherent difficulty in separating these related concepts. In addition to difficulties defining these different forms of ironic text, there has been little research that focuses specifically on satire. This study attempts to add the existing knowledge of satire by adopting a theoretical definition of satire (Simpson 2003) and investigating the importance of previous cues that have been identified as markers of verbal irony

(e.g. punctuation, interjections, semantic properties), as well as new cues related to lexical sophistication (e.g. word frequency, concreteness) and grammar (e.g. verb tenses) in distinguishing satirical from non-satirical texts. The following research questions guide the current study:

RQ 1. Does written satire differ from its non-satirical counterpart based on linguistic measurements when genre-level features are kept the same?

RQ 2. Does written satire differ linguistically from other forms of similar language use, such as irony and sarcasm?

RQ 3. Does the inclusion of new linguistic measurements of lexical sophistication provide a more detailed linguistic definition of written satire?

3. Methods

The purpose of this study is to investigate if lexical, grammatical, and semantic features differ between satirical and non-satirical Amazon.com product reviews, which may allow for a better understanding of the nature of satire and its relationship to other forms of related ironic language. While no study has previously attempted to analyse satire using a statistical model based on a combination of these linguistic features, the review of the literature shows that interest in analysing satire or related forms of language use currently exists. However, only features such as punctuation (Hancock 2004; Whalen et al. 2009) or semantic associations (Campbell & Katz 2012; González-Ibáñez et al. 2011) have been investigated. Automatic detection studies have demonstrated some success using various linguistic features (Burfoot & Baldwin 2009; Reyes & Rosso 2011; Carvalho et al. 2009), but they do not operationalise irony, satire, or the cues for these phenomena in the same way, making it difficult to synthesise their results. In this study, we follow Simpson's (2003) definition of satire, which clearly distinguishes satire from definitions of irony or sarcasm. Recognising the role that irony plays in satire, we incorporate previously identified measures of verbal irony into our model (i.e. semantic and grammatical), while also introducing previously unused measures of lexical sophistication that may account for features specific to satire.

The approach used in this study is different from previous studies such as Reyes & Rosso (2011), who classify all texts as ironic. We argue that satire, sarcasm, and other related forms of language use are similar *because* they are all ironic, but that there are other elements at play distinguishing these forms of language from one another. Those differences are what we hope to begin addressing, by first differentiating satire from non-satire. Additionally, while there is some overlap in corpora between this study and Reyes & Rosso (2011), their study included a large amount of reviews from five separate products, whereas our study includes a smaller amount of reviews from a broader range of products. We opted for a compromise between breadth and depth in order to complement their results. Thus, this study builds off of the previous work by using a corpus of data defined strictly as satire and incorporating a combination of lexical, grammatical, and semantic measurements that are clearly defined and applicable to any text.

3.1. Corpus

The corpus used in this study consists of 750 Amazon.com product reviews extracted from 50 different products from the American and UK editions of Amazon.com. Half of these reviews were satirical and the other half non-satirical (see Appendix A for a complete list of products). Collection started with a list maintained by Amazon.com that identifies funny reviews (Amazon 2013) and was supplemented by internet searches for funny Amazon reviews. Satirical reviews are attracted to products that are either offensive (as seen in the above BIC pen examples or otherwise ridiculous (e.g. a book titled *How to Avoid Huge Ships*); therefore, starting with a product that attracted one or two satirical reviews (through news or social media) provided a convenient way to locate even more satirical reviews.

Satirical reviews were first sorted by choosing to show the most helpful reviews first and then each review was pasted into an individual text file. Each review was read to check for a satirical opposition between the ostensibly helpful purpose of the review genre and the humorous purpose of satire. The first 15 reviews for each product over 100 words (including the title) or more in length with a satirical intent were included. Template information from the reviews (i.e. amount of "helpful" votes, star rating, authors' names) was removed from each file. Due to the nature of how Amazon.com promotes "helpful" reviews, it may be possible that authors of reviews could self-promote their funny reviews ahead of others. However, since the reviews reaching "most helpful" status for these products typically have hundreds to thousands of "most helpful" votes, the larger community of readers is typically responsible for the promotion of these posts rather than the authors themselves.

For the non-satirical reviews, we first located products similar to those that had attracted satirical reviews and then followed the same sorting procedures. Each review was read through in order to ensure that the author was providing a serious (i.e. non-satirical) review. Efforts were made to match products as closely as possible with each other (e.g. book for book). However, this was not possible in all cases, as some products did not have enough reviews to maintain a proper balance between corpora (i.e. fifteen reviews of over 100 words). If a direct match between products was not possible, a match with the overall category of products (e.g. electronics) was made.

The first author gathered the initial corpus and follow-up inter-rater reliability checks were performed to ensure that the texts were satirical or non-satirical by the first author and a trained colleague. Each rater separately coded a duplicated random sampling of 20 per cent of the combined corpus (150 reviews). Samples were coded based on the apparent communicative purpose of the reviews: satirical humour or aiding a consumer in making a purchasing decision (via evaluation, providing information, and so on). To make these decisions, we relied on Simpson's (2003) definition of satire. Ratings were guided by the following question: Is the review creating a dissonance between what is expected from an online review and what is actually being provided? We first determined if each review was humorous or not, and then considered whether the humour was ironic. Following Simpson (2003), we defined ironic humour as indirect humour, where the humorous meaning is not clearly stated and requires resolution of apparent incongruity. Our coding scheme, therefore, left open the possibility for a non-satirical review to still contain humour (e.g. canned jokes, self-deprecating jabs), and for a satirical review to still be indirectly helpful for consumers. Initial rater agreement was satisfactory at 96 per cent.

3.2. Linguistic indices and tools

We used two different text analysis tools in order to obtain linguistic measurements: Linguistic Inquiry and Word Count (LIWC) and the Tool for the Automatic Analysis of Lexical Sophistication (TAALES; Kyle & Crossley 2014). Both programs read each word in a text (or group of words, in the case of n-gram searches) and match that word to reference dictionaries that are loaded into the program's memory. The reference dictionaries are lists of words with values for various features derived from published psycholinguistic and cognitive linguistic results. To use these tools, we converted the corpus into plain text files and analysed each text individually, resulting in an average score for each individual text for each of the indices queried. Appendix B provides a list of the measurements that were used in the analysis, as well as an explanation of each measurement.

Specifically, LIWC is a program designed to measure the "social and psychological meaning of words" (Tausczik & Pennebacker 2009: 30) by assigning words to categories related to emotions, feelings, and other psychological processes. While the LIWC authors have stated that LIWC "…fails to appreciate sarcasm or irony" (Newman et al. 2008: 217, as cited in Campbell & Katz 2012), studies of sarcasm based on LIWC indices demonstrate that sarcasm was detectable using these indices (Campbell & Katz 2012). The TAALES is a program that measures the overall lexical sophistication of a text. Specifically, the TAALES contains psycholinguistic measurements which "[relate] to linguistic properties of words that affect word processing and learnability such as word concreteness, imageability, and familiarity" (Crossley et al. 2010: 3) and indices that calculate word frequency. In sum, LIWC is designed to measure the semantic (based on words' semantic qualities) and grammatical properties of texts, whereas TAALES is designed to measure the level of sophistication of a text (based on the difficulty of a text).

Each of these programs provides results for hundreds of different linguistic measurements; therefore, we carefully selected an initial group of indices tailored to our research questions. For LIWC, we chose to use the indices selected by Campbell & Katz (2012), because their study incorporated and tested a comprehensive list of linguistic cues related to sarcastic irony. Specifically, their indices included words semantically related to negative emotions, inclusion, exclusion, discrepancy, tentativeness, certainty, causation, swearing and sadness, along with words assigned to the grammatical properties of temporal past or present, quantification, and negation. Campbell & Katz (2012) reasoned that these indices were linguistic representations of the pragmatic discourse goals of sarcasm, such as showing negative emotions, clarifying, emphasizing, presence of a victim, and failed expectations. Because satire does not require the negative qualities associated with sarcasm, we added one additional measure (positive emotion words). We also included question marks and punctuation marks, based on studies that indicated punctuation as a marker of verbal irony or sarcasm (Hancock 2004; Whalen et al. 2009; Carvalho et al. 2009).

We supplemented the indices from LIWC with measurements related to lexical sophistication taken from TAALES. Because no previous research has examined the lexical sophistication of ironic or humorous texts, we opted for a set of indices designed to provide a general measurements of lexical sophistication. Therefore, measures of word concreteness,

familiarity, meaningfulness, imageability, age of acquisition, and frequency were included along with the LIWC measures. Word concreteness is a rating of how abstract or concrete the meaning of a word is. Higher concreteness ratings mean that a text is using less abstract language, and vice versa. Word meaningfulness is a rating of how many other words could be associated with a particular word. For instance, the words *food* and *music* have more associations and are, thus, more meaningful than the words *acumen* and *oblique*. Word imageability is a measure of how easily a word triggers a mental image, and word familiarity measures how salient a word is. TAALES measures concreteness, familiarity, and meaningfulness based on the MRC psycholinguistic database (Coltheart 1981). Frequency indices measure how often a word typically occurs in language use (Crossley et al. 2010); the frequency measurement selected for this study is derived from the SUBTLEXus corpus, which calculates frequency scores based on subtitles in film and television (Brysbaert & New 2009). The age of acquisition measurement is derived from Kuperman et al. (2012), who collected rater judgments of ages particular words are learned.

3.3. Statistical analysis

The following section provides details of the statistical tests that were performed in this study. The dependent variables for this study are the measurements of the lexical, grammatical, or semantic qualities of each text, and the independent variable is the type of review: satirical or non-satirical. As two different groups are being compared, a one-way multivariate analysis of variance (MANOVA) was conducted to identify if, based on the indices provided, a significant difference exists between the two types of product reviews. Individual differences for each variable were then examined to see which variables were contributing to the difference.

We followed the MANOVA with a stepwise discriminant function analysis (DFA). For the DFA, we used only those indices that demonstrated significant differences between the satirical and non-satirical reviews in the MANOVA. The DFA generates discriminant functions that can then be used as an algorithm to predict group membership (i.e. whether the texts are satirical or not). We used the DFA first on the entire corpus of reviews. Then, the DFA model reported for the entire corpus was used to predict group membership of the reviews in the corpus using leave-one-out-cross-validation (LOOCV). In LOOCV, one review in turn was left out and the remaining instances were used as the training set (in this case the 749 remaining reviews). We tested the accuracy of the reported DFA model by examining its ability to predict the classification of the omitted instance. The DFA, thus, allows us to test the model on an independent data set (i.e. using data that is not used to train the model). Similar results between the entire set and the *n*-fold cross-validation set provide evidence that the model can be extended to external data sets.

4. Results

4.1. MANOVA

A one-way MANOVA was conducted comparing the effect of product review type (satirical or non-satirical) on various linguistic measures (the dependent variables). Visual inspection of the data followed by square root transformations confirmed that the data was normally distributed, with the exception of the *swear* and *sadness* semantic measures and the *question mark* and *exclamation mark* punctuation measures, which were removed due to extreme positive skewness. After checking the data for multicollinearity using a threshold of $r \geq .70$, two more variables were removed from the model (*imageability* correlated positively with *concreteness* ($r = .93$) and was removed; *age of acquisition* correlated positively with *SUBTLEXus frequency* ($r = .77$) and was removed). Further visual inspection of boxplots for each variable suggested that homogeneity of variance was not violated. While some outliers existed, they were not removed in order to maintain independence of the data.

After checking the statistical assumptions, we were left with a total of sixteen dependent variables. A Levene's test for each variable indicated that homogeneity of variance was violated for two variables (*SUBTLEXus frequency* and *familiarity*), which is most likely attributed to outliers. These variables were removed. Using Pillai's Trace, there was a significant difference between the two review types $V = 0.25$, $F(14, 735) = 17.344$, $p < .001$, with an effect size accounting for roughly 25 per cent of the variance (partial $\eta^2 = .248$). Follow-up between-subject comparisons found significant differences for twelve of the fourteen dependent variables. Table 1 displays descriptive statistics for the measurements of each review type and Table 2 displays a summary of the individual comparisons for each variable, sorted by effect size.

Table 1

Descriptive statistics for selected indices by review type

	Satirical		Non-Satirical	
Index	*M*	*SD*	*M*	*SD*
Concreteness	373.399	21.935	363.976	22.454
Meaningfulness	412.999	13.168	409.334	13.535
Past	1.848	0.711	1.588	0.655
Present	6.397	2.768	8.074	2.601
Quantifier	1.618	0.439	1.865	0.441
Negation	1.162	0.474	1.257	0.454
Inclusion	4.567	1.733	4.716	1.836
Exclusion	1.436	0.530	1.682	0.512
Negative emotion	1.311	0.521	1.014	0.558
Positive emotion	1.746	0.475	1.901	0.439
Tentativeness	1.389	0.502	1.540	0.540
Certainty	1.171	0.481	1.087	0.516
Discrepancy	1.120	0.498	1.175	0.504
Causation	1.248	0.515	1.312	0.516

Table 2

F value, p value, and effect size for selected indices

Index	*F*	*p*	*Partial η²*
Present	73.073	<.001	0.089
Quantifier	59.195	<.001	0.073
Exclusion	41.491	<.001	0.053
Concreteness	33.791	<.001	0.043
Past	26.952	<.001	0.035
Positive emotion	21.531	<.001	0.028
Meaningfulness	14.126	<.001	0.019
Negative emotion	56.505	<.001	0.070
Tentativeness	15.648	<.001	0.020
Negation	7.847	.005	0.010
Certainty	5.160	.023	0.007
Causation	2.879	.090	0.004
Discrepancy	2.225	.136	0.003
Inclusion	1.295	.255	0.002

4.2. Discriminant function analysis

The stepwise DFA automatically selected the variables that best classified the grouping variable (satirical or non-satirical reviews) based on a statistical criterion. The significance level for a feature to be entered into or to be removed from the model was set at $p \leq 0.05$. The stepwise DFA retained seven variables as significant predictors of whether a review was satirical or non-satirical (*present tense, quantification terms, negative emotions terms, exclusion terms, word concreteness, positive emotion terms,* and *certainty terms*) and removed the remaining variables as non-significant predictors.

The results demonstrate that the DFA using the seven indices correctly allocated 538 of the 750 reviews in the total set, $\chi 2$ (df=1, n=750) = 141.738, p < .001, for an accuracy of 71.7 per cent (the chance level for this analysis and all analyses is 50 per cent). For the leave-one-out cross-validation (LOOCV), the discriminant analysis correctly allocated 530 of the 750 reviews for an accuracy of 70.7 per cent (see the confusion matrix reported in Table 3 for results). The measure of agreement between the actual text type and that assigned by the model produced a Cohen's Kappa of 0.435, demonstrating a moderate agreement.

Table 3

Predicted satire type (satirical or non-satirical reviews) from total set and test set for nouns

Actual satire type	Predicted satire type	
Total set	Satirical	Non-satirical
Satirical	**266**	109
Non-satirical	103	**272**
LOOCV Set	Satirical	Non-Satirical
Satirical	**261**	114
Non-satirical	106	**269**

5. Discussion

The purpose of the statistical analyses conducted in this study was twofold. The first analysis (MANOVA) compared the mean values of each of the fourteen linguistic features for both text types in order to determine if any of those mean values differed significantly. A significant result from the MANOVA means that, based on the combination of the fourteen linguistic features, there is a measurable difference between the two text types that cannot be attributed to chance. The MANOVA in this study reported a significance value of less than < .001, meaning that there is a greater than 99.9 per cent chance that the difference between review types was not attributable to chance. Furthermore, the effect size (a measure of the magnitude of the difference) for the MANOVA was 25 per cent, which means that this test was able to account for 25 per cent of the difference or variance between the two text types. In other words, these 14 linguistic features make up approximately 25 per cent of the difference between satirical and non-satirical texts.

We then used the significant variables from the MANOVA as predictors in a stepwise DFA, which attempted to discriminate between two different groups based on certain features (i.e. the fourteen indices). The stepwise DFA ran a series of text selection tasks, where it attempted to correctly identify satire and non-satire based on different samples of the corpus. The DFA results indicated that of seven of the fourteen indices used in the MANOVA best differentiate satirical and non-satirical texts, at an accuracy of 71.7 per cent (whereas chance would be 50 per cent). This accuracy is slightly lower than that reported by Reyes & Rosso (2011) and Burfoot & Baldwin (2009), who reached accuracies of 78 per cent and 72 per cent, respectively. However, our focus in this study is not entirely on the automatic detection of satire, but rather testing the results of the MANOVA using a stepwise DFA. Additionally, the satirical targets in Burfoot & Baldwin (2009) were more readily identifiable (being political figures), which allowed for the inclusion of more constrained automatic detection methods. Because the targets of satire in the current study are less obvious (as they could be the product, the company behind the product, or something tangentially related to the product), the textual indices we included were necessarily less specific.

Overall, three categories of linguistic properties were analysed: measures of semantic associations, lexical sophistication, and grammatical function. All three of these categories

produced at least one significant predictor variable in the DFA. In regard to the semantic features, our results demonstrate that satirical reviews contain more *negative emotion* words while non-satirical reviews have more *positive emotion* words. Satirical texts also have higher levels of *word certainty*, but lower levels of *exclusion words*. For lexical sophistication, only one measure proved to be a significant predictor: *word concreteness*; satirical reviews contained higher levels of concreteness than non-satirical reviews. Finally, for grammatical function, both *present tense* and *quantification* words were significant predictors. The results from the MANOVA and DFA suggest that these seven features are the linguistic characteristics that can best predict if a text is satirical or non-satirical.

The first conclusion that can be drawn from these results is that sarcasm and satire share similar semantic features, but important differences also exist. Both sarcasm and satire have higher levels of *negative emotion* words, but satire tended to contain fewer *exclusion words*. Campbell & Katz (2012) reasoned that higher levels of *exclusion words* signalled the presence of a victim, and they found that sarcastic text did contain higher levels of exclusion than non-sarcastic text. Satire, on the other hand, contained fewer *exclusion words* than non-satire, indicating that satire may not explicitly mention a victim, which is in line with satire's subtle strategy mocking of satirical targets. Furthermore, our results show that satire contains more *certainty words* than non-satire, whereas Campbell & Katz (2012) found that sarcasm contained fewer *certainty words* than non-sarcasm. This suggests that authors of satirical texts use less hedging and may appear more confident. Furthermore, certainty words (e.g. *never*, *always*) can be associated with exaggeration or hyperbole – an identified function of verbal irony (Hancock 2004) *and* relatable to Simpson's (2003) concept of satirical saturation (i.e. exaggerating characteristics of a satirical target). This suggests that while both satire and sarcasm are ironic, they are still separate forms of humour.

The second conclusion that can be drawn is based on satire's higher levels of *word concreteness*. This means that the language in satirical reviews is less abstract than language in non-satirical reviews, suggesting that using words with more specific meanings is a distinguishing characteristic of satire from non-satire. Metonymic and metaphoric strategies of satire were seen in the examples presented earlier, and a preference for one of these strategies may be the reason for higher levels of word concreteness. Brysbaert et al. (2013: 1) define word concreteness as "the degree to which the concept denoted by a word refers to a perceptible entity". Since metonymic attenuation is carried out through underlexicalisation and non-direct references to a perceptible entity (Simpson 2003), the higher levels of word concreteness might suggest less evidence of metonymic attenuation, as direct references to a perceptible entity would serve to overtly lexicalise that entity. The non-satirical reviews are more abstract, suggesting that they employ language at a more sophisticated level than satirical reviews.

The third conclusion to be drawn is that satire employs fewer markers of *present tense* and *quantification* words than non-satire. Along with the results of the MANOVA, which showed a significant difference between *past tense* words (with satirical reviews containing more), this finding suggests that the tense of a review is a significant feature in distinguishing satire. Many of the satirical texts presented a story or narrative related to the product being reviewed, and as such, it makes sense that more markers of past tense were found in satirical reviews. Since satirical reviews are also fictional, creating a fanciful narrative of a past event may be one larger rhetorical strategy of satirical product reviews. The reasons for more *quantification* (e.g. *few*,

many) in non-satirical texts are not as clear, but may be related to reviews that discuss the purchasing and use of a product.

To summarise, the features of the satirical reviews that most reliably set them apart from non-satirical texts were higher levels of *word concreteness*, *negative emotions*, *certainty terms* and lower levels of *present tense*, *exclusion terms*, *positive emotion terms*, and *quantification terms*. The results here suggest that satire shares the negative bite of sarcasm, but is also more certain and does not employ language that indicates the presence of a victim. This provides evidence in support of defining satire as a separate form of ironic language use.

In terms of Simpson (2003) and his model of satire, it may be that satirical Amazon.com reviews favour the satirical strategy of metonymic saturation. Because of the higher levels of word concreteness and word certainty, hyperbole and references to specific entities may work as the ironic means that performs the ironic opposition between the prime and the dialectical (i.e. the purpose of the product review and the meaning behind the product review). None of the other significant linguistic features of satirical Amazon.com reviews align well with attenuation, negation, or metaphor. While there were some linguistic differences noticed between satire and sarcasm (recognising that this study was not a direct comparison), satire and sarcasm may differ more in non-linguistic, and, therefore, less empirically measurable ways. In other words, the genre-specific constrains of satire may be more important to defining satire than the linguistic means that create incongruity.

Therefore, models built around the detection of irony in text may be capable of also detecting satire, but additional contextual information and analysis is required in order to detect satire. This is similar to the conclusion reached by Burfoot & Baldwin (2009) when they concluded that subtle cues of irony were unable to be detected computationally. Our results may provide a starting point at detecting the more subtle cues of satire (i.e. word concreteness, tense, negative emotions), but the problem of definition still persists. As Reyes & Rosso (2011: 123) point out, "irony is one of the most subjective phenomena related to linguistic analysis". This is primarily what statistical investigation such as this and similar studies have to offer humour research: approaches that work to better define what these notoriously subjective forms of language use are.

6. Conclusion

The results of this analysis suggest that there are significant differences in written satire and its non-satirical counterpart, when examining various lexical, grammatical, and semantic features of both types of writing. This supports the argument that satire uses specific linguistic means in order to signal irony and, ultimately, humour (Simpson 2003). However, satire does not appear to differ greatly from sarcasm, suggesting that satirical irony may not be unique in relation to other forms of irony, such as sarcasm. Rather, other non-linguistic features of satire may need to be taken into consideration, making it difficult for linguistic measurements alone to define satire. Nonetheless, the novel inclusion of measures of lexical sophistication in this study indicated that satire uses less abstract language than its non-satirical counterpart, which may help inform research into all forms of ironic language use.

The results of this study provide answers to the three research questions outlined earlier. The first research question asked whether satire differs from non-satire based on linguistic

measurements. The findings indicate that satirical product reviews differ significantly from non-satirical product reviews on several linguistic measures, demonstrating that satire relies in part on specific linguistic strategies. The second research question asked if satire differs linguistically from other forms of related ironic language. While we did not directly compare the two, the findings in combination with previous research support the notion that satire and sarcasm share some semantic features when compared to their non-sarcastic or non-satirical textual counterparts. The third research question asked whether the inclusion of previously unused measures of lexical sophistication would help further define the linguistic properties of satire. The findings report that new indices of lexical sophistication (e.g. *word concreteness*) are significant predictor of satire, providing evidence that satirical language use may be less abstract than non-satirical language. Such a finding can provide future researchers with additional linguistic features to consider when investigating humorous or ironic language.

This research has implications for the field of humour, precisely because satire has been long neglected as a target of study (Simpson 2003). This is partially due to problems of definition, which hinder the ability for researchers to confidently refer to a satirical or sarcastic text as anything less than ironic. This question is brought up by Colston & Gibbs (2007: 4), who consider whether irony should be considered as a broad phenomenon or whether "irony is simply a family of related phenomena that each require their own theoretical approach". The results of this study and those seen in other similar studies (e.g. Reyes & Rosso 2011; Burfoot & Baldwin 2009; Campell & Katz 2012) indicate the family view of ironic language is more theoretically fruitful because it allows for a more nuanced understanding of irony and provides clear definitions based on linguistic measurements that future studies can build from. By extension, satire needs to also be measured against other forms of *humorous* language, because satire is simultaneously humorous and ironic (Simpson 2003). Future studies comparing a corpus of satire, sarcasm, and other forms of ironic language against one another are needed in order to further explain differences among these forms of language, as well as comparisons with other types of humorous texts.

References

Amazon. (2013). *Funny Reviews: Dynamic List*. http://www.amazon.com/gp/feature.html?ie=UTF8&docId=1001250201 (accessed 18 September 2013).

Attardo, S. (2000). 'Irony as relevant inappropriateness'. *Journal of Pragmatics* 32, pp. 793-826.

Burfoot, C. & Baldwin, T. (2009). 'Automatic satire detection: Are you having a laugh?', in *Proceedings of the Association for Computational Linguistics International Joint Conference on Natural Language Processing 2009 Conference: Short Papers (Singapore, 2-7 August 2009)*, pp. 161-164.

Brysbaert, M. & New, B. (2009). 'Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English'. *Behavior Research Methods* 41 (4), pp. 977-990.

Brysbaert, M., Warriner, A. & Kuperman, V. (2013). 'Concreteness ratings for 40 thousand generally known English word lemmas'. *Behavior Research Methods* 46 (3), pp. 904-911.

Campbell, J. & Katz, A. (2012). 'Are there necessary conditions for inducing a sense of sarcastic irony?'. *Discourse Processes* 49 (6), pp. 459-480.

Carvalho, P., Sarmento, L., Silva, M. & de Oliveira, E. (2009). 'Clues for detecting irony in user-generated contents: Oh ...!! It's "so easy" ; -)', in *TSA '09: 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion (Hong Kong, 6 November 2009)*, New York: Association for Computing Machinery, pp. 53-56.

Caucci, G. & Kreuz, R. (2012). 'Social and paralinguistic cues to sarcasm'. *Humor: International Journal of Humor Research* 25 (1), pp. 1-22.

Colston, H. & Gibbs, R. (2007). 'A brief history of irony', in Gibbs, R. R. & Colston, H. (eds.), *Irony in Language and Thought: A Cognitive Science Reader*, New York: Lawrence Erlbaum Associates, pp. 3-21.

Coltheart, M. (1981). 'The MRC psycholinguistic database'. *Quarterly Journal of Experimental Psychology* 33 (4), pp. 497-505.

Condren, C. (2012). 'Satire and definition'. *Humor: International Journal of Humor Research* 25 (4), pp. 375-399.

Crossley, S. A., Salsbury, T., McNamara, D. S. & Jarvis, S. (2010). 'Predicting lexical proficiency in language learner texts using computational indices'. *Language Testing* 28 (4), pp. 561-580.

Gibbs, R. (2000). 'Irony in talk among friends'. *Metaphor and Symbol* 15 (1-2), pp. 5-27.

González-Ibáñez, R., Muresan, S. & Wacholder, N. (2011). 'Identifying sarcasm in Twitter: A closer look', in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (*Portland, Oregon, 19-24 June 2011): Short Papers*, Stroudsburg, PA: Association for Computational Linguistics (ACL), pp. 581-586.

Hancock, J. (2004). 'Verbal irony use in face-to-face and computer-mediated conversations'. *Journal of Language and Social Psychology* 23 (4), pp. 447-463.

Jorgensen, J. (1996). 'The functions of sarcastic irony in speech'. *Journal of Pragmatics* 26 (5), pp. 613-634.

Kreuz, R., Long, D. & Church, M. (1991). 'On being ironic: Pragmatic and mnemonic implications'. *Metaphor and Symbolic Activity* 6 (3), pp. 149-162.

Kreuz, R. & Caucci, G. (2007). 'Lexical influences on the perception of sarcasm', in *FigLanguages '07: Proceedings of the Workshop on Computational Approaches to Figurative Language*, Stroudsburg, PA: Association for Computational Linguistics (ACL), pp. 1-4.

Kreuz, R. & Caucci, M. (2008). 'Do lexical factors affect the perception of sarcasm?' Paper presented at the 18th Annual Meeting of the Society for Text and Discourse. University of Memphis, Memphis, TN, 12-15 July.

Kuperman, V., Stadthagen-Gonzales, H. & Brysbaert, B. (2012). 'Age-of-acquisition ratings for 30 thousand English words'. *Behavior Research Methods* 44 (4), pp. 978-990.

Kyle, K. & Crossley, S. A. (2014). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*.

LIWC, Inc. (n.d.). Linguistic inquiry and word count: Table 1: LIWC2007 output variable information. http://www.liwc.net/descriptiontable1.php (accessed 1 November 2013).

Mihalcea, R. & Strapparava, C. (2006). 'Learning to laugh (automatically): Computational models for humor recognition'. *Computational Intelligence* 22 (2), pp. 126-142.

Newman, M., Groom, C., Handelman, L. & Pennebaker, J. (2008). 'Gender differences in language use: An analysis of 14,000 text samples'. *Discourse Processes* 45, pp. 211-236.

Nilsen, A. & Nilsen, D. (2008). 'Literature and humor', in Raskin, V. (ed.), *The Primer of Humor Research*, New York: Mouton de Gruyter, pp. 243-280.

Pennebaker, J., Booth, R. & Francis, M. (2007). *Operator's Manual: Linguistic Inquiry and Word Count: LIWC2007*. Austin, Texas: LIWC.net http://homepage.psy.utexas.edu/HomePage/Faculty/Pennebaker/Reprints/LIWC2007_OperatorManual.pdf (accessed 1 October 2013).

Popova, M. (n.d.). *Modern Masterpieces of Comedic Genius: The Art of the Humorous Amazon Review*. http://www.brainpickings.org/index.php/2013/07/08/humorous-amazon-reviews/ (accessed 1 September 2013).

Reyes, A. & Rosso, P. (2011). 'Mining subjective knowledge from customer reviews: A specific case of irony detection', in *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (Portland, Oregon, 24 June 2011)*, Stroudsburg, PA: Association for Computational Linguistics (ACL), pp. 118-124.

Simpson, P. (2003). *On the Discourse of Satire*: *Towards a Stylistic Model of Satirical Humor*. Amsterdam & Philadelphia: John Benjamins Publishing Company.

Skalicky, S. (2013). 'Was this analysis helpful?: A genre analysis of the Amazon.com discourse community and its "most helpful" product reviews'. *Discourse, Context & Media* 2 (2), pp. 84-93.

Tausczik, Y. & Pennebaker, J. (2009). 'The psychological meaning of words: LIWC and computerized text analysis methods'. *Journal of Language and Social Psychology* 29 (1), pp. 24-54.

Whalen, J., Pexman, P. & Gill, A. (2009). '"Should be fun – Not!": Incidence and marking of nonliteral language in e-mail'. *Journal of Language and Social Psychology* 28 (3), pp. 263-280.

Appendix A
Amazon.com products used for corpus

Satirical Products	N	Non-Satirical Products	N
Hutzler 571 Banana Slicer	15	Paderno Vegetable Slicer	15
Three-Wolf Moon T-Shirt	15	Carhartt Men's Workwear Pocket T-Shirt	15
BIC Cristal For Her Ball Pen	15	Sharpie Accent Retractable Highlighters	15
Wheelmate Laptop Steering Wheel Desk	15	Dashboard Cell Phone Mount	15
Avery Durable View Binder	15	Wall Mount Pencil Sharpener	15
Tuscan Whole Milk, 1 gal	15	The Switch Sparkling Juice	15
Uranium Ore	15	Weber Chimney Starter	15
Denon AKDL1 Dedicated Link Cable	15	Mediabridge Coaxial Cable	15
Accoutrements Horse Head Mask	15	SecondSkin Full Body Suit	15
How to Avoid Huge Ships	15	How to Read A Book	15
A Million Random Digits with 100,000 Deviates	15	Python Programming: An Introduction	15
Veet for Men Hair Removal Gel Crème	15	Philips Norelco PT730 Electric Razor	15
JL421 Badonkadonk Land Cruiser/Tank	15	Maisto R/C/ Rock Crawler	15
The 2009-2014 Outlook for Wood Toilet Seats in Greater China	15	The Next 100 Years: A Forecast for the 21st Century	15
Parent Child Testing Product	15	LEGO Minecraft	15
BIC Cristal Ball Pen	15	Paper Mate Inkjoy Pens	15
UFO-02 Detector	15	CH Hanson 03040 Magnetic Stud Finder	15
Canned Unicorn Meat	15	Peak Dry Whole Milk Powder	15
Passion Natural Water-Based Lubricant - 55 Gallon	15	Slim Jim Smoke Snack Sticks	15
Fresh Whole Rabbit	15	GMO Free Garbazno Beans	15
Looking For-Best of David Hasselhoff	15	Journey: Greatest Hits	15
Guardian Angel	15	FDL Digital Acupuncture Pysiotherapy Machine	15
Deglong Meeting Knife Set	15	ChicagoCutlery Fusion 18-Piece Knife Set	15
How to Live with a Huge Penis	15	Overcoming Self-Defeating Behavior	15
Microwave for One	15	How to Cook Everything	15
	375		375

Appendix B

List of lexical measures with explanations and examples

Measure	Explanation	Examples*
Concreteness	Average concreteness of content words	
Meaningfulness	Average meaningfulness of content words	
Familiarity	Average familiarity of content words	
Frequency	Average frequency of content words as compared to the SUBTLEXus corpus	
Positive Emotions	Number of words with positive semantic meaning	*love, nice, sweet*
Negative Emotions	Number of words with negative semantic meaning	*hurt, ugly, nasty*
Inclusion	Number of words with inclusive meaning	*and, with, include*
Exclusion	Number of words with exclusion meaning	*but, without, exclude*
Present Tense	Number of words marking present tense	*is, does, hear*
Past Tense	Number of words marking past tense	*went, ran, had*
Certainty	Number of words expressing certainty	*always, never*
Discrepancy	Number of words expressing discrepancy	*should, would, could*
Tentativeness	Number of words expressing tentativeness	*maybe, perhaps*
Causation	Number of words associated with causation	*because, effect*
Quantifier	Number of quantifier words	*few, many, much*
Negation	Number of words marking negation	*no, not, never*

*Examples from LIWC (n.d.)