Undergraduate Journal of Humanistic Studies • Spring 2015 • Vol. I

Sarcasm Detection Using Grice's Maxims

Johanna Maren Hjelle Olsen Carleton College

May 24, 2015

I Introduction

Sarcasm is of interest to researchers because of its ambiguous nature—sarcasm research can provide insight into how we use different types of linguistic and extralinguistic information to process ambiguous language of many types (Rockwell, 2005). We can also use knowledge about how sarcasm functions in conversation to inform our understanding of pragmatic language difficulties present in people with particular brain injuries or developmental disabilities, such as Autism Spectrum Disorders or Specific Language Impairment (Ryder & Leinonen, 2014; Surian, Baron-Cohen, & Van der Lely, 1996). For example, determining which kinds of information people typically use to detect sarcasm can help practitioners assist patients with pragmatic difficulties in identifying and interpreting nonliteral speech.

Constructing a complete explanatory theory of sarcasm interpretation has been a source of controversy since Grice's foundational 1975 paper, "Logic and Conversation," which proposes that conversational partners assume one another to be acting cooperatively within the current speech exchange, and thus often interpret an apparent breach of conversational protocol as an indirect way of implying something which has not actually been said. Two major countertheories emerged soon after—Mention Theory and Pretense Theory—with criticisms of Grice and new ways to conceive of sarcastic utterances altogether. Both theories posit that only the literal meaning of the utterance is represented, and the speaker's attitude toward the utterance is indicated by other methods. These single-meaning models differ from Grice's in that they eliminate the need for an inferred "inverted" meaning of the utterance. A modified version of Grice's original theory accounts for these criticisms as well as newer experimental evidence regarding the processing of nonliteral language. While these theories emphasize the interpretation of sarcasm, more recent sarcasm research focuses on sarcasm detection, specifically which kinds of information hearers utilize when determining whether a given utterance is sarcastic. For instance, while vocal and visual cues may be available to hearers for sarcasm detection in a face-to-face conversation, contextual and lexical cues may become more important in written language.

Couched in Gricean terms and expanding upon previous research on sarcasm cues, I propose a particular set of sarcasm cues that rely on an additional violation of Grice's maxims as a way of directing a conversational partner's attention to the nonliteral nature of a sarcastic utterance. In order to test the validity of these maxim-violation cues, a set of tweets was compiled, half of which were sarcastic and half of which were intended literally. Half of each of those groups contained maxim-violation cues, while half did not. The use of tweets as the source of the utterances was intended to target language which did not make use of more well-studied sarcasm cues such as body language, intonation, and, for the most part, conversational context. Participants were asked whether each of the tweets was sarcastic. Participants also recorded how certain they were of their sarcasm detection for each tweet, providing a built-in tool to disregard noise in the data due to random guesses. The presence of maxim-violation cues predicted a statistically significant amount of variance in sarcasm detection. The effect of sarcastic intent on sarcasm detection was also statistically significant, suggesting that other content-based cues (aside from maxim-violation cues) come into play even in written language with little context.

This paper will be organized as follows: Section 1 will establish a theoretical framework within which to analyze sarcasm detection; section 2 will introduce the concepts of sarcasm detection and discuss various sarcasm cues, most importantly motivating the new set of sarcasm cues which will form the basis of the experiment reported in sections 3 and 4. Section 3 will describe the experimental design, and section 4 will present the results and conclusions of the experiment.

II Arriving at a Theory of Sarcasm

Grice's 1975 paper "Logic and Conversation" fueled a new conversation about how to analyze ironic language by proposing a new model of irony in which the literal interpretation of an utterance is inverted to find its intended meaning. Two major theories appeared shortly after Grice proposed his meaning-inversion model. Mention Theory (also called echoic reminder theory or echo theory and often considered a subdivision of relevance theory) posits—in contrast with Grice—that a sarcastic utterance does not *use* language, but rather *mentions* or *echoes* a belief that is not held by the speaker at the time of the utterance (Kreuz & Glucksberg, 1989; Sperber & Wilson, 1981; Wilson & Sperber, 2002). Pretense Theory attempts to expand on and modify Grice's account, arguing that by speaking sarcastically, one pretends to be an unwise speaker addressing a naïve audience but intends that the hearer will see through the pretense to understand the speaker's true attitude toward the fictional speaker and the belief expressed (H. H. Clark & Gerrig, 1984).

Section 1.1 will introduce Grice's truth-conditional meaning-inversion model before sections 1.2 and 1.3 outline Mention Theory and Pretense Theory. I highlight the disadvantages of both theories in their respective sections and address their concerns with Grice's analysis by proposing a few modifications to the traditional meaning-inversion model. Section 1.4 will then summarize my revised Gricean analysis, and section 1.5 will provide additional empirical support for a Gricean approach.

II.1 Truth-conditional meaning-inversion model

I will analyze sarcasm detection throughout the paper within the framework of Grice's cooperative principle of conversation (Grice, 1975). The cooperative principle is not proscriptive or even generally descriptive of natural language, but is rather a guideline that speakers assume their conversational partners to be following throughout a given conversation: "Make your contribution such as it is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged" (Grice, 1975, p. 43).

According to Grice (1975), the cooperative principle can be divided into four maxims. The maxims are as follows: *quality* (do not say things that are false or for which you lack evidence), *quantity* (do not give too much or too little information), *relation* (say things that are relevant to the conversation at hand), and *manner* (avoid obscurity and ambiguity; be brief and orderly). Speakers are said to be *exploiting* or *flouting* a maxim when they choose not to follow the rules in order to cue the listener in to a hidden meaning, called an *implicature*.

The essential component of a sarcastic utterance is its violation of Grice's maxim of quality in order to express the opposite of what has been said. For example, one might hear a Minnesotan say (1a) on a particularly blustery winter day.

1. (a) It's practically *tropical* outside today!
   (b) It's miserably cold and snowy.

Of course, this utterance seems on the surface to violate the maxim of quality—the literal interpretation is certainly not true. Assuming that their conversational partner is always following the cooperative principle, the hearer will then discount the literal meaning as a possible interpretation of the utterance and search for a different but related attitude that the speaker must intend to express. Here, the hearer will infer that what the speaker really means is something like (1b).

In a natural language setting of the sort Grice assumes when formulating his theory, the hearer is able to look around at the setting of the conversation in order to assess the truth of the speaker's claim—they can see that their environment is not, indeed, tropical. They are also able to rely on the speaker's intonation (emphasis indicated here with italics), their body language (an eye roll or an exaggerated gesture to the snowy outdoors), the broader context of a conversation, or even their relationship with the speaker and their knowledge of the speaker's sarcastic tendencies. However, even in a context-free written form—with none of the above to consider—it seems clear that (1a) is more likely to be used sarcastically than either of the statements in (2).

2. (a) It's not quite as miserably cold and snowy as usual.
   (b) It's nice outside today!

This indicates that some aspect of the literal content of the utterance—or more precisely, the way the utterance is phrased—indicates to the hearer that it may be sarcastic. In section 4, we will see evidence from participants' evaluation of the sarcastic intent of tweets that violations of the cooperative principle such as an excessively formal register cue to hearers that an utterance may be more likely to be sarcastic.

Bach (2005) makes explicit an assumption of Grice's theory: that the implicature is not carried by the sentence itself, but by the utterance of that sentence within its context. For example, (1a) could be uttered on a particularly warm day in the depths of winter as a hyperbolic statement rather than a sarcastic one. This utterance would still seem to violate the maxim of quality, but rather than *inverting* the truth-conditional content of the utterance (the hallmark of sarcasm under Grice's model), the relationship changes such that the implicature is simply a *weakened* version of what is actually said similar to (2a). Thus, speakers can utter the same sentence in different contexts to implicate different meanings.

Additionally, there do seem to be some sentences which entail contradictions when taken literally. Even if the sentence does not inherently implicate a particular interpretation, the only true interpretation of the sentence may be when it is being used to express the implication. For example, A is giving B directions and says to take a right. B turns right, prompting A to say, "Oh, I meant left!" B replies sarcastically, "It's okay, left and right are totally the same thing." Here, the semantic content of the sentence entails a contradiction—left and right are opposites; they are not the same. Said literally, this sentence cannot be true without changing the meanings of the words used. However, it is the *pragmatic* details of the situation which allow the implication to bring the utterance into line with the cooperative principle. This strategy of sarcasm—where the violation of quality is made explicit in the literal meaning of the utterance—is a sarcasm detection cue that I will term *self-contradiction*. It will be discussed in more detail in section 3, and empirical evidence for its use by hearers to detect sarcasm will be provided in section 4.

II.2 Pretense Theory

The Pretense Theory seeks to expand on Grice's underdeveloped claim that speaking ironically involves pretending (1978). According to Clark and Gerrig (1984), the speaker of a sarcastic utterance pretends to be an injudicious speaker addressing an uncritical audience. The speaker intends their actual audience to understand, due to some common ground, that the hypothetical speaker they are impersonating and any audience who would take the utterance at face value are unwise or misinformed. Thus, the speaker communicates their negative attitude toward the satirized speaker and audience as well as the utterance itself. Despite its origin in a clause of Grice's analysis, however, Pretense Theory rejects the meaning-inversion model in which hearers apply a mechanism of inversion to the utterance to uncover its intended meaning. Instead, the utterance has only one meaning—the literal interpretation—and the hearer surmises from the combination of the utterance and the common ground which suggests that the utterance is false that the speaker holds a negative attitude toward the utterance and its supporters (H. H. Clark & Gerrig, 1984).

This theory does seem to account nicely for the role of mutual beliefs and context in sarcasm detection, an issue Grice does not explicitly discuss. Sarcasm is indeed used more frequently among friends (who presumably share common ground) than strangers (Caucci & Kreuz, 2012) and that contextual cues significantly increase a hearer's ability to correctly identify sarcastic utterances (Rockwell, 2000, 2005; Tepperman, Traum, & Narayanan, 2006). However, Clark and Gerrig fail to explain the difference between (1a) and (2) such that (1a) is more likely to be sarcastic, regardless of the lack of context or common ground (as previously discussed in section 1.1). Their theory requires that the hearer rely on context, familiarity with the speaker, or previous knowledge to deduce the speaker's true intentions. However, the fact remains that sarcastic intent can often be determined without any of the above sources of knowledge. Rather, the hearer may be alerted to the presence of sarcasm by intonation[1] and body language, in the case of spoken language (Bryant & Tree, 2005; Kreuz & Roberts, 1995; Rockwell, 2000; Tepperman et al., 2006), or by the use of lexical cues and the maxim-violation cues introduced in section 2.2, in the case of context-free written language like tweets (González-Ibáñez, Muresan, & Wacholder, 2011; Kreuz & Caucci, 2007; Kreuz & Roberts, 1995; Rockwell, 2005). Of course, it is extremely unlikely that any type of purely linguistic sarcasm cue is used in written language alone, so it follows that these cues are also utilized in the natural language environment that Clark and Gerrig wish to address. Furthermore, while Grice does not advocate for contextual knowledge as an essential sarcasm detection cue, his theory does not preclude it. The traditional Gricean analysis of sarcasm focuses entirely on the interpretation of given sarcastic utterances rather than the detection of the presence of sarcasm in an utterance, and so makes no claim regarding which types of information may or may not be used to detect sarcasm.

Even with the simple example in (1), Pretense Theory seems to fall short of a true description of sarcastic language. The speaker who utters (1a) pretends to be a misinformed person—one who thinks that the weather is, in fact, quite summery, or one who believes it to be cold and snowy in the tropics, perhaps. The speaker's hypothetical audience may agree readily, but they intend their actual audience to understand that they are only pretending, and that in reality, the speaker would think that anyone who spoke or agreed to the utterance was quite foolish. Pretense Theory shines when an explicit antecedent for the sarcastic utterance can be found—for example, if the hearer had earlier told the speaker not to bring a jacket or that it would not be very cold outside. Here, the target of the speaker's disdain is clear; they think the hearer was foolish for having suggested the weather would be anything but freezing. The hyperbolic use of the word "tropical" only serves to heighten the sense of absurdity in the hearer's previous remark—the speaker suggests that the hearer, foolish enough to suggest that they not bring a coat, is then foolish enough to believe that the snow constitutes tropical weather.

However, (1) could reasonably be uttered as the initiation of a conversation between two strangers—no context, no possible antecedent is necessary. If no such antecedent exists, we are left with a statement that no one would be foolish enough to agree to, with no explanation for who it is the speaker intends to condescend to—again, Pretense Theory requires that hearers rely on context to interpret sarcastic utterances. In a natural language setting, the combination of clues from the immediate environment with the definition of the word "tropical" assures that the utterance cannot be sincerely agreed upon. Thus, the utterance serves only to express contempt for a logical absurdity—but this account does not accurately reflect the experience of the hearer, who

---

[1] It is worth noting that Clark and Gerrig do argue that their theory can explain the existence of an ironic tone of voice—they claim that this change in register reflects the speaker's impersonation of another (1984). Regardless of whether this explanation provides evidence under their theory for the use of intonation as a sarcasm detection cue (they make no claim regarding its use by the hearer), the hearer is required under pretense theory to utilize contextual information to deduce the speaker's true beliefs whether or not they may use other cues to detect the presence of sarcasm in the first place.

will easily surmise that the speaker means (1b). It is likely that even a stranger reading (1a) from far away (with no sense of what the speaker's actual climate was like) could deduce the speaker's true beliefs from the words used alone. Thus, Pretense Theory offers no way to infer the intended meaning from what is said, and does not account for the interpretation of sarcastic utterances such as tweets which lack context. The Gricean analysis does not require any antecedent or any one particular sarcasm detection cue to be present, and thus avoids the pitfalls that plague Pretense Theory.

II.3 Mention Theory

Proponents of Mention Theory analyze sarcasm with respect to the use-mention distinction (Sperber & Wilson, 1981). Take the sentences in (3), for example. While (3a) *uses* the underlined word to refer to a domesticated feline, (3b) *mentions* the word itself, referring not to a feline but to an abstract object that we can use to talk about a feline. (3c) and (d) illustrate the same concept with a full sentence. (See Jorgensen, Miller, & Sperber, 1984 for a more detailed description of this distinction.)

3. (a) I hate your cat.
   (b) I hate how 'cat' looks in this font.
   (c) I hate cats.
   (d) 'I hate cats' is an awful thing to say.

Advocates for Mention Theory, unlike Grice, propose that sarcastic utterances are not processed quite the same way as sincere ones (Jorgensen et al., 1984; Sperber & Wilson, 1981; Sperber, 1984; Wilson, 2006). While sincere utterances are instances of use, sarcastic ones are instances of mention, where the sarcastic utterance *echoes* some attitude, belief, or idea in a way that makes it clear that the speaker disagrees with the literal content of the utterance (Sperber & Wilson, 1981).

Recall example (1)—under Sperber and Wilson's analysis, the speaker echoes the earlier (incorrect) suggestion of a misinformed friend or weatherman that today's weather would be warm, a hope or expectation that they or their audience had had, or a general cultural norm of some sort, i.e. that we perpetually hope the weather will be warm and sunny. By the time of Sperber and Wilson's later works (Wilson & Sperber, 2002; Wilson, 2006; see also Kreuz & Glucksberg, 1989), the definition of an antecedent becomes so inclusive that it is unclear what utterance would *not* echo some kind of antecedent. The necessity of an antecedent is what causes some researchers to group Pretense Theory and Mention Theory together under an umbrella of "expressivism" (Attardo, 2000; Camp, 2012), but the broader category of echo material and the loss of unwise speaker and audience as targets of contempt differentiate the Mention Theory antecedent from that of the Pretense Theory.

Sperber and Wilson (1981) framed Mention Theory as an alternative to Grice's analysis mainly in that under Grice's analysis, a sarcastic utterance uses language, unlike under Mention Theory. However, meaning-inversion—the crux of Grice's analysis—can occur regardless of whether sarcastic utterances bear an echoic quality as long as Grice's fundamental criterion for sarcasm remains intact: that sarcastic utterances violate quality in order to implicate the opposite of what is said.

II.4 Reimagining Grice's meaning-inversion model

Sperber and Wilson (1981) argue that Grice's meaning-inversion model is flawed in that a violation of the maxim of quality is not necessary nor sufficient for an utterance to be sarcastic. However, Grice is not committed to this violation being sufficient; other types of non-literal language, such as hyperbole, may violate quality without being sarcastic (Grice, 1975). It is specifically the meaning-inversion relationship of literal interpretation and implicature that identifies sarcasm. But even this Grice does not propose as a sufficient criterion—he proposes a second condition for sarcastic utterances, that they must express a negative attitude toward some belief or proposition (Grice, 1978).

It is false, though, that sarcastic utterances must use positive literal interpretations to express negative emotions—this observation builds upon Sperber and Wilson's critique. Grice asserted that an utterance which uses a negative literal interpretation to express a positive sentiment is playful, not ironic, giving the example "What a scoundrel you are!" said to a friend who has done something others find reprehensible (1978, p. 54). Grice uses this example to argue that irony must express a negative sentiment, but the context seems to overcomplicate this example. The following is a more straightforward instance of a sarcastic utterance expressing a positive sentiment. A and B are at A's party. A is worried about how it's going and wails, "This party is going terribly!" B, looking around at the blue sky and the delicious food, tries to comfort A by saying, "Yeah, the weather is horrible, the food is awful..." B thus implicates, "The weather is great, the food is great—and so is the party. It's not so terrible." This utterance clearly fits our definition of sarcasm—stating the opposite of the speaker's true intentions. The difference in "playfulness" Grice observes is only the difference in using nonliteral language with a positive mood versus a negative one, a difference that may be found in literal language, as well. Thus, sarcasm can be used to express either a positive or a negative sentiment, and I reject Grice's second criterion for sarcasm. Still, though, the meaning-inversion model of sarcasm does not require that violating the maxim of quality be a sufficient condition of sarcasm.

Grice's claim that sarcasm is "intimately connected" to emotional expression, however, seems to remain valid (1978, p. 53). Recall once again example (1). Note that while the sentence that was spoken did not contain any explicit linguistic information about the speaker's emotional reaction to the weather, it seems that the implicature could not be positive in this situation. One could say on a warm summer day, "It's practically *tropical* outside today—and I hate it," proving that the positive connotation of the word "tropical" is cancellable, and yet the negative connotation of its opposite (when the sentence is used sarcastically) is not. (1a) cannot be said sarcastically to implicate "It's marvelously freezing outside!" It seems, then, that when the emotional sentiment of a sarcastic utterance is not explicitly communicated, it is involved, nonetheless, in the process of meaning inversion that delivers the intended meaning of the utterance. Thus, the process of meaning inversion must be sensitive to information outside the truth-conditional content of an utterance. However, Grice advocates that his model include only the explicit truth-conditional content of an utterance and that every sarcastic utterance violate the maxim of quality (Grice, 1975, 1978), which would not account for examples like the one above. Sperber and Wilson (1981) are correct to raise the concern that a violation of quality is not necessary for a sarcastic utterance under Grice's analysis. However, we can remedy this issue by expanding what is included in "meaning" with a more inclusive meaning-inversion model (see also Camp, 2012).

The major example Sperber and Wilson use to challenge the notion of a necessary quality-violation is the existence of sarcastic questions (1981). Questions can be used sarcastically despite having an explicit purpose other than expressing a sentiment, and despite often lacking explicit truth-conditional content. For example, a professor announces a surprise test next class that will constitute a large portion of the course grade. After class ends, one student says to her friends, "So how excited are you about that test?" By this she expresses that she is not excited; in fact, she is dreading it—despite the literal interpretation of the utterance not including any information about the speaker's emotional state. The process of meaning-inversion that takes place for the hearer to realign the utterance with the cooperative principle thus includes the implications of the literal interpretation, i.e. "I am excited about the test" or "I assume you're at least somewhat excited about the test." The hearer then reverses these implications to find the intended interpretation: "I'm dreading this test, and I assume you are, too." The friend that responds to the question with "Not at all!" seems to not quite understand the speaker as well as the friend that says, "I know, I can't believe he didn't give us more time!" Thus, the meaning-inversion that occurs in the interpretation of sarcastic utterances must take into account the implicatures of the literal interpretation, making the form of a sarcastic utterance just as important as its content for expressing a particular sentiment as well as for alerting the hearer to its sarcastic nature.

Camp (2012) draws a distinction between examples which function under the traditional truth-conditional model of sarcasm—propositional sarcasm—and those utterances which require the meaning-inversion to include more information—illocutionary sarcasm. It is important to note that many sarcastic utterances do fall into the latter category, but Camp's distinctions between four subsets of sarcasm are largely irrelevant for our purposes. By simply including implicatures of the literal interpretation of a sarcastic utterance in "what is said"—and therefore in the content that is reversed—we can account for the major criticism from Mention Theory within the framework of the meaning-inversion model.

This more inclusive model of meaning-inversion also accounts more easily for the interpretation of sarcastic utterances such as tweets which involve written language with little context. When the meaning of an utterance can include more than what is literally said—namely the implicatures of nonfinal interpretations—it is reasonable to theorize that the implicatures of a non-sarcastic interpretation of the utterance can indicate the presence of sarcasm to a hearer. For example, the presence of hyperbole—as discussed with example (1) in sections 1.1 and 1.2—or another additional maxim-violation (on top of the main quality violation shared by all sarcastic utterances) may signal to the hearer that the entire utterance was meant nonliterally, including the implicatures of the literal interpretation. That is, additional maxim-violations may act as cues for sarcasm detection.

The other question raised by mention theorists is why, under the meaning-inversion model, a speaker would choose to utilize such a seemingly convoluted mode of communication—expressing one thought by saying its opposite (Wilson, 2006). They propose that sarcasm serves to emphasize a discrepancy between reality and expectation or desire, and that the Mention Theory of sarcasm accounts for this emphasis because echoing some antecedent expresses a "dissociative attitude" of the speaker toward the anteceding idea or proposition (Wilson, 2006, p. 1724). Indeed, sarcasm does express this contrast more readily than other forms of nonliteral language like understatement (Colston & O'Brien, 2000), but meaning inversion accounts for contrast with the very mechanism by which the sarcastic interpretation of the utterance is derived. By definition, the sarcastic interpretation bears the utmost contrast to its sincere counterpart. Furthermore, regardless of how sarcastic meaning is derived, sarcasm is simply a useful rhetorical tool, worth whatever processing mechanism it requires. Sarcasm is more intimate and bonding and more face-saving than literal language (H. H. Clark & Gerrig, 1984; Colston & O'Brien, 2000; Pexman & Zvaigzne, 2004), as well as funnier, more effectively criticizing, and more expressive of emotion than other types of nonliteral language (Colston & O'Brien, 2000; Roberts & Kreuz, 1994; Toplak & Katz, 2000). Finally, despite the complex nature of a meaning-inversion mechanism, there exist cues for sarcasm detection (such as intonation, body language, lexical cues, and maxim-violation cues) which signal hearers to process the utterance as a sarcastic one, as we will see in more detail in sections 2.1 and 2.2. In the latter section, I will introduce to the field a new set of sarcasm detection cues (which I will term maxim-violation cues) motivated by the presently modified Gricean model of sarcasm and supported by the experimental results presented in section 4.

II.5 Psycholinguistic evidence for a Gricean approach

As for a more current approach to sarcasm research, psycholinguistic studies regarding the processing speed of sincere versus sarcastic utterances provide empirical evidence for the meaning-inversion model of sarcasm. Both pretense and Mention Theory propose one-step processing models where the only interpretation is the literal interpretation (Camp, 2012; H. H. Clark & Gerrig, 1984; Sperber & Wilson, 1981); the hearer is to process the literal interpretation and thus understand the speaker's attitude toward it. The modified meaning-inversion model, however, requires the hearer to process the literal interpretation and its implicatures before applying the inversion mechanism to arrive at the sarcastic interpretation of the utterance. Thus, while Mention Theory and Pretense Theory would predict that sarcastic utterances can be processed just as quickly as sincere utterances, the meaning-inversion model would correctly predict sarcastic utterances to take longer to process, since two interpretations are represented rather than just one.

Filik and Moxey (2010) found that reading times for sarcastic statements (as measured using eye-movement) were longer than reading times for literal statements, suggesting that more steps are required in the processing of sarcastic utterances than literal ones. Schwoebel, Dews, Winner, and Srinivas found similar results, with slightly more nuance regarding the emotional content of the utterance (2000). In Filik and Moxey's study, participants then read material that pronominally referenced the earlier sarcastic utterance in a way consistent with either the literal or the sarcastic interpretation. Reading times for the pronominal references of both types indicate that both the literal interpretation and the sarcastic interpretation remain equally active during on-line processing, providing evidence that both interpretations are, indeed, represented during the initial processing of the utterance.

Giora et al. (2007) also measured reading times for sarcastic versus literal utterances and found that sarcastic utterances took more time to process. Participants had longer response times to sarcastically-related probe words than to literally-related probe words after reading a story that ended in either a literal or a sarcastic remark, regardless of which story they had been primed with. This suggests, in contrast to Filik and Moxey's study, that the literal interpretation of the utterance may be more readily available than the sarcastic interpretation. However, while Filik and Moxey investigated on-line processing, Giora et al. measured response times to probe words after somewhat of a delay after reading the relevant stories, so their results may represent some sort of decay of the sarcastic interpretation over time. This kind of finding has little to no bearing on the current study, which is focused on real-time processing of sarcastic utterances rather than storage of information represented sarcastically, so does not signify an empirical argument against a Gricean analysis of sarcasm.

The increased processing time for sarcastic utterances and the equally active on-line representations of the literal and sarcastic interpretations suggest that hearers process and represent both the literal and nonliteral interpretations of a sarcastic utterance (Filik & Moxey, 2010; Giora et al., 2007; Schwoebel, Dews, Winner, & Srinivas, 2000; see also Giora & Fein, 1999 for broader discussion of nonliteral language processing).

III Sarcasm Detection

Characteristics of sarcasm like intonation, hyperbole, and an excessively formal register have traditionally been analyzed as symptoms of sarcasm, calling for explanation by a model of sarcasm comprehension (H. H. Clark & Gerrig, 1984; Grice, 1975, 1978; Sperber & Wilson, 1981). More recent contributions to the pretense-mention-inversion debate have done the same (Camp, 2012; Wilson, 2006), but overall, literature surrounding sarcasm research in the last decade has shifted its focus from theoretical models of sarcasm processing to empirical testing of sarcasm detection. This has caused the field to reframe these features of sarcasm as cues used to signal the hearer to apply the meaning-inversion mechanism to the processing of a given utterance, but without much empirical evidence as of yet. In section 1 we determined theoretically how a hearer derives the meaning of a sarcastic utterance from what is said, but the more recent question the literature asks is one step earlier in the process—is this utterance sarcastic?

This reanalysis of sarcastic properties suggests that it may not be necessary to explain the existence of particular traits as a criterion of a sound theory of sarcastic comprehension, but still, we may be able to describe cues which are proven empirically within the framework of the meaning-inversion model and understand how they function. The model itself may also suggest certain cues, as our modified Gricean model does. Section 2.1 introduces well-studied sarcasm detection cues that are available in natural language but not in tweets, and section 2.2 focuses on the detection cues that are available on Twitter and introduces a new set of cues which will be the focus of the investigation in sections 3 and 4: maxim-violation cues.

III.1 Cues for sarcasm detection in natural language

There are three broad categories of sarcasm detection cues which usually do not play a role in channels like Twitter. Vocal cues are the most studied sarcasm cues, perhaps because they are the most obvious category to begin the search for reliable cues and because changes in intonation can be utilized in every natural language setting. Vocal cues which have been recently investigated include a lower pitch, slower tempo, higher volume, heavier stress, nasalization, more frequent changes in pitch, and more pauses surrounding the utterance (Bryant & Tree, 2005; Kreuz & Roberts, 1995; Rockwell, 2000; Tepperman et al., 2006). However, vocal cues are generally an unreliable measure by which to detect sarcasm, suggesting that hearers do not rely as heavily on intonation cues for sarcasm detection as previously thought (Bryant & Tree, 2005; Kreuz & Roberts, 1995; Rockwell, 2000; Tepperman et al., 2006).

Unlike vocal cues, visual cues can only be utilized in face-to-face interactions, but when they are available (as in most natural language settings), they are one of the most reliable tools for detecting sarcasm (Rockwell, 2005). Speakers tend to utilize more frequent and more exaggerated movements of the head, eyebrows, eyes, and mouth when speaking sarcastically, and these cues easily indicate a sarcastic utterance for hearers who are able to utilize such information (Caucci & Kreuz, 2012; Rockwell, 2005).

Contextual cues are also found to be fairly reliable, and are available in some written language as well as spoken language. Tepperman et al. (2006) found that the context surrounding an utterance of the phrase “yeah, right” in a phone conversation was one of the most consistent predictors of whether the utterance was intended sarcastically. Other researchers have found that providing the context preceding an utterance to participants greatly increases the accuracy of their sarcasm detection (Rockwell, 2000, 2005; Tepperman et al., 2006) and that accurate sarcasm detection of written samples is greatly reduced when context is not available (González-Ibáñez et al., 2011). González-Ibáñez et al. (2011) used tweets in their experiment as well and found a greatly reduced sarcasm detection accuracy rate among their participants as compared to other studies where context was provided.

III.2 Cues for sarcasm detection in written language

Cues which rely on the actual content of an utterance are less well-studied. These are the cues which are most relied upon in written contexts with little common ground between speaker and hearer or with negligible contextual information, such as on Twitter or some other social networks, especially other microblogging platforms like Tumblr. Certain lexical cues and other content-related cues have been investigated: interjections, emotional and evaluative language, excessiveness and superlatives, and positive adverb-adjective pairs (like “absolutely lovely”) have all been linked to sarcasm, but their reliability in terms of sarcasm detection by hearers of an utterance remains unclear (González-Ibáñez et al., 2011; Kreuz & Caucci, 2007; Kreuz & Roberts, 1995; Rockwell, 2005).

Some of the lexical cues that have been marginally studied as sarcasm cues—such as hyperbole—easily fit into a Gricean analysis of sarcasm. These cues can fit into a meaning-based linguistic analysis because unlike intonation or body language, lexical cues reflect the actual content of an utterance; their use partially determines the meaning interpretation of the utterance. Recall that the core violation in a sarcastic utterance is that of the maxim of quality. It then follows that if this violation is made explicit within the content of an utterance, it will be easier to identify as sarcastic. When meaning-inversion is made explicit, an utterance contradicts itself—I will call this feature *self-contradiction.* This is the first of three content-based cues I will investigate in sections 3 and 4.

These cues may be especially important in the absence of extralinguistic cues like tone and body language, or with a lack of context—for example, if the hearer of (1) was in a different state from the speaker and was not aware of the weather conditions in the speaker's state, it might be more difficult to recognize the speaker's sarcastic intent. Note that any cue utilized in a more restrictive environment is also available in a less restrictive environment (i.e. all sarcasm cues are available in face-to-face interaction, while the fewest number of cues are available in a context-free written setting), so cues that are available in the most limiting circumstances are in fact the most universal indicators of sarcasm. Thus, the investigation presented in sections 3 and 4 provides evidence that these cues are used not only in written language, but also in a natural language setting.

IV Experimental Design

IV.1 Maxim-violation cues

The content-based cues I investigated rely on violations of the cooperative principle. There may be many cues within the category of maxim-violation, but I will focus on three: self-contradiction, hyperbole, and manner violation. If the chief violation of quality is made explicit in a sarcastic utterance, it will be easier to recognize as sarcasm. Thus, speakers may do so in some cases to cue to the hearer that the utterance is sarcastic. This is the first of three maxim-violation cues I examine in the remainder of the paper and will be termed *self-contradiction.* Self-contradiction differs from the other maxim-violation cues in that it does not introduce an additional violation of the cooperative principle to the utterance. Instead, it makes the essential violation (that of quality, a violation which all sarcastic utterances share) explicit in the language used, rather than the violation residing in the contradiction between, for example, utterance and context. Self-contradiction can be split into two sub-types: lexical contradiction and sentimental contradiction. Lexical contradiction describes an utterance in which the words used necessarily contradict one another: recall our earlier driving example, "Left and right are totally the same thing." "Left" and "right" cannot be the same; this sentence is contradictory based upon the definitions of the words used regardless of context. Sentimental contradiction, however, relies on common social knowledge of the positive or negative connotation of a particular situation. This type is based on the work of Riloff et al. (2013), who focus their entire paper on sarcastic utterances utilizing this cue. For example, take "I love when my car stalls." While none of the words here necessarily contradict one another, "love" is a positive sentiment, and a stalled car is generally considered to be a negative situation. Thus, sentimental contradiction is somewhat weaker than lexical contradiction, due to its reliance on socially-shared connotations and its lesser accuracy rate. (While lexical contradiction certainly indicates a false statement, sentimental contradiction may not: it is possible to love when one's car stalls, but it is not possible for left and right to be the same.)

If making an existing maxim-violation explicit indicates sarcasm to a hearer, it may follow, then, that an additional violation of the cooperative principle may draw attention to the utterance and verify that it is not sincere, causing the hearer to reconsider the literal interpretation and apply the meaning-inversion mechanism to reach the intended sarcastic meaning. These additional maxim-violations, mentioned briefly earlier in the paper, include hyperbole (an additional violation of quality) and an excessively formal register or some other type of manner violation. Hyperbole is mentioned in the literature, but not yet studied in great detail. Hyperbole violates the maxim of quality, but in a different way than sarcasm itself, as discussed in section 1.1. Thus, hyperbole which is found in sarcastic utterances adds a second quality violation to the utterance. This second violation serves to cue to the hearer that the utterance cannot be interpreted at its literal value, triggering the meaning-inversion mechanism. Of course, hyperbole can be used in non-sarcastic utterances, but when it is present, may signal hearers to look for alternative interpretations of the utterance, regardless of whether they might find one or not. Sarcastic utterances such as "I can't even tell you how much I love the winter" utilize hyperbole to express a sarcastic statement. Kreuz & Roberts (1995) determined that lexical hyperbole (phrases such as "wonderfully perfect") acted as effective cues to sarcasm, but did not test examples like the one given here. The current study examines hyperbole similar to the example above as well as more situational hyperbole as in (1a). Camp (2012), Utsumi (2000), and Wilson (2006) also mention hyperbole as a common feature of sarcastic utterances.

The third and final sarcasm cue I will investigate is manner-violation. This may be the most broadly-defined of the three. Repetition violates manner, specifically the sub-maxim "be brief," as in this sarcastic utterance: "I'm not mad. Nope. Not mad at all. Not even a little" when presumably the same amount of information is conveyed in the much more concise, "I'm (not) mad." Another construction commonly found in sarcastic utterances is the "it's not like… (or anything)" construction—"It's not like I was waiting for three hours or anything"—which seems to violate briefness as well. "I was waiting for three hours" would convey the message more briefly, but the manner violation signals to hearers that they ought to pay attention for the true meaning of this utterance. One of the more common devices on Twitter, it seems, is using an overly formal register for posting on Twitter, which violates manner much the same way as above. For example, one might say to a friend who has playfully insulted them, "Oh, aren't you quite a dear," taking on an excessively formal register to indicate their lack of seriousness.

In order to determine whether maxim-violation cues are, indeed, utilized by hearers in order to detect sarcasm, it is necessary to present them to hearers in a setting where they are forced to rely on the content of the utterance alone to determine sarcastic intent. Asking participants to evaluate the sarcastic intent of tweets eliminates their ability to use visual or vocal cues, as well as contextual cues to some degree. Of course, broader societal context is impossible to eliminate, and to attempt to do so would be detrimental: indeed, sentimental contradiction almost always relies on a culturally-determined notion of what is considered undesirable. This background knowledge is a necessary part of almost any talk exchange; no utterance can occur in an isolated vacuum. Participants may also be aware of a prominent event to which a tweet refers, giving the utterance some context based on the participant's understanding of the event. The potential interference of cultural context with our results will be discussed in detail in sections 4.1 and 4.2. By eliminating visual, vocal, and contextual cues to the extent possible, we leave participants only content-based cues with which to determine sarcastic intent. While this may somewhat diminish the accuracy of participants' sarcasm detection as found by González-Ibáñez et al. (2011), it will reduce the risk that participants will rely on other cues, minimizing the noise to signal ratio of the study's results.

IV.2 Participants and methods

Forty-five participants between the ages of 18 and 27 completed an online survey. Participants were required to be native English speakers and to have never taken a linguistics course in order to qualify for the study. 80% of participants live in Minnesota most of the time; 20% live primarily elsewhere in the US. 24 of the participants never use Twitter, 10 use Twitter a few times a month or less, and eight use Twitter a few times a week. Only three participants use Twitter every day.

Before the study began, participants were asked whether they agreed with the following definition of sarcasm: sarcasm is when a person "say[s] the opposite of the truth, or the opposite of their true feelings in order to be funny or to make a point" (BBC, 2013). A space was provided for participants to provide their own definition if they did not agree with the given definition. Four participants provided their own definitions, but none was substantially different from the given one.[2]

In a procedure loosely based on that of Kreuz & Caucci (2007),[3] participants were then shown a series of 28 tweets and answered two questions for each tweet:

- Is the tweet sarcastic? (yes or no)

- How certain are you?

Participants rated their level of certainty on a scale of 1 to 4, where 1 was not at all certain, 2 was somewhat uncertain, 3 was somewhat certain, and 4 was very certain. The 28 tweets were randomly chosen for each participant from a pool of 56 tweets total and were displayed in a random order.

At the end of the survey, a short demographic section asked participants whether they considered themselves well-informed about particular topics (sports, politics, and movies) that were relevant to the subject matter of many of the tweets, as well as whether they had watched the 2015 State of the Union address, which many of the tweets referenced. Participants were also asked for their age, where they live most of the time, and how often they use Twitter.

At the end of the survey, a short demographic section asked participants whether they considered themselves well-informed about particular topics (sports, politics, and movies) that were relevant to the subject matter of many of the tweets, as well as whether they had watched the 2015 State of the Union address, which many of the tweets referenced. Participants were also asked for their age, where they live most of the time, and how often they use Twitter.

IV.3 Collection and coding of tweets

There were 56 tweets, 14 in each of the following four categories: intended sarcasm with and without any maxim-violation cues (henceforth called MV cues), and no intended sarcasm with and without any MV cues.[4] Thus, half the tweets were sarcastic and half were not, and half the tweets included MV cues and half did not, in a 2x2 factorial design. Table 4 below shows examples for each of the four categories.

---

[2] Two participants mentioned a predisposition of sarcasm to be negative, while one mentioned a sarcastic tone of voice. The fourth participant included irony within the umbrella of sarcasm but did not elaborate on the distinction between the two.

[3] Kreuz and Caucci (2007) asked participants to rate the likelihood of sarcastic intent for excerpts (including context and utterance) from published works including the phrase "said sarcastically," with "sarcastically" removed. Participants rated these utterances as more likely to be sarcastic than the control utterances, which did not include "sarcastically" in their original form. The exerpts were rated on three potential sarcasm cues: presence of adjectives and adverbs, presence of interjections, and use of punctuation. Based on statistical analysis of the data, participants seemed to rely most heavily on the presence of interjections in the utterances as a cue.

[4] A few additional pieces of information were recorded for each tweet, including the presence of negative phrasing in each tweet and the presence of interjections, which some argue to be a sarcasm cue (González-Ibáñez et al., 2011; Kreuz & Caucci, 2007; Rockwell, 2005).

	Sarcastic	Non-sarcastic
Contains MV cues	“2 hours of sleep is cool, I mean it’s not like I have an interview and 4 classes today” • Sentimental contradiction: Getting such a small amount of sleep is a negative experience, but “cool” expresses a positive sentiment. • Manner violation: “It’s not like” constructions violate briefness.	“Can’t wait till the in laws come! Missed them a lot” • Hyperbole: The speaker says they “can’t wait” when in reality they can wait, but do not want to. • Sentimental contradiction: In-laws visiting is culturally considered a negative experience, but the utterance expresses a positive sentiment.
Lacks MV cues	“So good to have the person you need stand by you when you feel down!” • Does not contain self-contradiction, hyperbole, or manner-violation	“What a great way to start off the day! Great session with an amazing woman!” • Does not contain self-contradiction, hyperbole, or manner-violation

Table 4: Example tweets for each category.

All sarcastic examples were culled from Twitter’s hashtag #sarcasm, and all non-sarcastic utterances with maxim-violation cues were collected from the hashtag #notsarcasm. The hashtags were removed before being presented to participants. All non-sarcastic utterances without cues were found by searching topics prevalent in the body of sarcastic tweets and gathering tweets from those topic search pages. Tweets that referenced photos or external links were disqualified from inclusion in the study. Extraneous hashtags and links, emoticons, and indications of tone using capitalization were removed from the tweets along with the relevant hashtag (#sarcasm or #notsarcasm) where applicable. This prevents the interference of extralinguistic information in sarcasm detection based on MV cues.

IV.4 Expected results

We expect to find that the presence of MV cues predicts a significant amount of the variance in sarcasm perception from actual sarcastic intent. That is, if MV cues were the only type of cue available to hearers for detecting sarcasm, and if MV cues were infallible, we would expect every utterance with MV cues to be perceived as sarcastic 100% of the time, and for every utterance without MV cues to be perceived as sincere 100% of the time. We do expect some variance in the data explained by factors other than the presence of MV cues, however, since there are likely other cues available to hearers in this setting (i.e. interjections, emotional and evaluative language, excessiveness and superlatives, and positive adverb-adjective pairs, as described in section 2.2), and since an utterance can contain maxim-violations without being sarcastic (i.e. a hyperbolic but non-sarcastic utterance, discussed in more detail in sections 1.1 and 3.1). Due to the natural occurrence of these other cues in sarcastic utterances more often than in non-sarcastic ones (hence their usefulness as indicators of sarcasm), and because tweets were not chosen nor excluded based on the presence of other potential cues, we expect that actual sarcastic intent will explain some of the variance in sarcasm perception, as well.

To a lesser extent, we expect that each of the MV cues individually—self-contradiction, hyperbole, and manner violation—will explain some of the variance in sarcasm detection. However, the study was designed to identify whether hearers use MV cues in general to detect sarcasm, so more specific findings may not be substantial. Due to the small number of tweets with a given cue (between 4 and 9 tweets per factorial subset), as well as the overlapping of cues within a single tweet, results for individual cues may not be as straightforward as for the maxim-violation category of cues as a whole.

Certainty scores were mainly used as a method of reducing noise in the data, but they also provide additional insights into the results. If tweets that contain MV cues are more often perceived as sarcastic than those that do not, we expect to find that tweets with MV cues will have higher certainty scores than those without. This trend would suggest that MV cues make sarcasm more easily identifiable to hearers.

V Results and Discussion

The reported findings of this study exclude all responses for which the participant rated their certainty at a 1 or 2 ("not at all certain" and "somewhat uncertain," respectively) except for reports of average certainty scores and where stated otherwise. This procedure serves to reduce noise in the data by eliminating responses generated by random guesses. On average, 26.5% of responses from each of the four tweet categories were excluded by this constraint. The total mean certainty score was 3.05 out of 4, barely above "somewhat certain." The average certainty scores for each of the four categories ranged from 2.87 (sincere tweets lacking MV cues, with the highest percentage of excluded (low certainty) tweets at 32.5%) to 3.27 (sarcastic tweets containing MV cues, with the lowest percentage of tweets excluded at 20.1%). This tight range indicates that none of the four categories prompted participants to be either extremely certain or extremely uncertain about their judgments.

Throughout this section, I will continue to use the term "speaker" for the utterer of a sarcastic or sincere utterance and "hearer" for the interpreter of that utterance for the sake of consistency throughout the paper and following the conventions of the literature. In the context of this study, however, the "speaker" is the author of the tweet, and the "hearer" is anyone who might read the tweet.

V.1 Presence of maxim-violation cues versus intent as explanatory variables

The major comparison of interest in the study is the perception of sarcasm in tweets with MV cues versus those without: Table 5 summarizes these results. Note that based on *intent* rather than *perception*, the true proportion for each cell below is 0.5. This would also be the expected proportion of perceived sarcasm if participants were able to correctly identify sarcastic intent and if MV cues did not affect perception of sarcasm.

	Perceived as sarcasm	Perceived as sincere
Tweets with maxim-violation cues	0.7848361	0.2151639
Tweets without maxim-violation cues	0.3310345	0.6689655

Table 5: Proportion table for sarcasm evaluations conditioned on the presence of maxim-violation cues.

As expected, tweets with MV cues were perceived as sarcastic significantly more often than those without (p-value = 0.0001, alpha = 0.05). Tweets with MV cues were perceived as sarcastic 78.5% of the time, while tweets without MV cues were perceived as sarcastic only 33.1% of the time. This difference shows that participants did seem to be using MV cues to determine whether a tweet was sarcastic and provides more general evidence that hearers rely on MV cues to interpret sarcastic intent in a context-free, written setting. As discussed in section 2.2, cues utilized in a restrictive language environment are likely to be utilized in less restrictive environments, as well—thus, it is likely that hearers in a natural language environment (which is the subject of our modified Gricean model of sarcasm) utilize MV cues in their detection and interpretation of sarcasm.

MV cues explain a significant amount of the deviance of perceived sarcasm from sarcastic intent, but other intervening factors must have contributed to the overall pattern of results. If MV cues were the only explanatory factor for participants' judgments, tweets with MV cues would be perceived as sarcastic close to 100% of the time and those without, almost never, as discussed in section 3.4. Since the observed proportions differed substantially from these extremes, participants must have utilized some cues other than MV cues. As discussed in section 3.4, these other cues are expected to pattern with sarcastic intent and can thus be roughly measured by the influence of sarcastic intent on the variance of sarcasm perception. Indeed, intent explained a significant amount of variance in perceived sarcasm (p-value = 0.0003, alpha = 0.05). The explanatory effects of both intent and presence of cues can be seen in Table 6, which presents the total proportion of tweets perceived as sarcastic within each of the four categories of tweets.

	Perceived as sarcasm	Perceived as sincere
Tweets with maxim-violation cues	0.958175	0.582222
Tweets without maxim-violation cues	0.537445	0.105769

Table 6: Total proportion of tweets perceived as sarcastic for each category; sarcastic intent versus presence of maxim-violation cues.

Here, we can see that the presence of MV cues has a substantial effect on sarcasm perception, as tweets with MV cues are perceived as sarcastic considerably more often than those without for both tweets intended sarcastically and those intended sincerely. Likewise, tweets intended sarcastically are perceived as sarcastic more often than those intended sincerely for tweets with and without MV cues. Since sarcasm cues presumably derive their usefulness from a higher prevalence in sarcastic utterances than in sincere ones, we can assume that sarcastically-intended tweets with MV cues contain the most cues, and sincerely-intended tweets without MV cues contain the fewest. Intent and the presence of MV cues thus have a cumulative effect such that the former category is perceived as sarcastic over 95% of the time and the latter only 10.6% of the time. Even these results have variance, of course, due to the difficulties of detecting sarcasm, especially in written language with no context—after all, the most dependable indicators of sarcasm (visual and contextual cues) are not available to the hearer (Rockwell, 2005).

In fact, each of these more straightforward categories (those where intent and presence of MV cues predict the same results) has only one significant outlier. (7a) is a sarcastically-intended tweet with MV cues (sentimental contradiction and a manner violation due to overly formal register). While the other 13 sarcastic tweets with MV cues were perceived as sarcastic between 81.2% and 100% of the time, (7a) was perceived as sarcastic exactly half of the time.

4. (a) Anyone else excited about the impending Spruce Grove-St. Albert byelection? (0.500)[5]

---

[5] Following each example in section 4 is the proportion of participants who perceived the tweet as sarcastic,

(b) The speech last night president Obama's Was so wonderful for us (0.583)

Despite the special considerations for sarcastic questions examined in section 1.4, it doesn't seem that this anomaly is due to the fact that the utterance is a question: a similar sarcastic question (also beginning with "Anyone else…") that did not contain MV cues was perceived to be sarcastic a small minority of the time, as predicted by the lack of MV cues. It seems that the wide disagreement between participants is circumstantial rather than reflective of an intrinsic feature of such sarcastic questions: the presence of the self-contradiction cue depends on a belief that byelections are generally considered unexciting. If one does not hold that belief, the self-contradiction is no longer present, reducing the tweet's likelihood of being sarcastic. This explanation is supported by the difference in sarcasm perception between participants who claimed to be well-informed about politics versus those that did not. Because so many responses were excluded due to low certainty scores (70.6%), I used the full set of data, including all responses, to compare the proportion of perceived sarcasm for participants informed about politics to the proportion for those not informed about politics. Participants who were informed about politics, and who presumably then find elections to be more exciting, perceived the tweet to be sarcastic 26.7% of the time. Participants less informed about politics perceived the tweet to be sarcastic almost twice as often, at 47.4%. This difference can explain much of the outlier status of this particular tweet, illustrating that the irregular nature of the tweet is not due to any particular quality inherent in sarcastic questions.

The tweet in (7b) was perceived as sarcastic over half of the time (58.3%), despite being a sincere tweet with no MV cues. Most other tweets in this category were perceived as sarcastic less than 10% of the time. This tweet may have been perceived as sarcastic due to a belief that most people found Obama's State of the Union address, which took place just before results were collected, to be less than wonderful, thus implicating a sentimental contradiction. The outlier status of this tweet may also be due to the fact that it appears to have been written, perhaps, by a non-native English speaker. The syntax is broken and the phrasing awkward, potentially leading to some ambiguity regarding its intent—indeed, almost half (45.5%) of the responses were excluded due to low certainty scores, demonstrating participants' confusion. Of course, notwithstanding the two outliers dissected in this section, the overall effect of both intent and the presence of MV cues on sarcasm perception were quite strong.

V.2 Evidence for non-maxim violation cues

Returning to consider the influence of intent versus that of the presence of MV cues, the specific results that can be explained by former but not the latter are tweets whose intent was correctly identified more often than expected given the presence or lack of MV cues: that is, sarcastic tweets with no MV cues and sincere tweets with MV cues. For both groups, their middling proportion of perceived sarcasm was due not to great disagreement *within* each tweet, but more to varied proportions of perceived sarcasm *between* tweets. This suggests that for some tweets, like the examples in (8) and (9), sarcasm detection was based largely on the presence of MV cues, while for other tweets, non-MV cues indicated to participants the true intent of the tweet. The examples in (8) are of sarcastically-intended tweets that were largely perceived as sincere due to their lack of MV cues. Likewise, (9) presents a sample of sincerely-intended tweets that were generally perceived as sarcastic due to the presence of MV cues.

5. (a) So nervous about the game tonight!!! (0.000)
   (b) So good to have the person you need stand by you when you feel down! (0.0870)
   (c) The weather is so pretty here in Corpus today. (0.182)

---

excluding low-certainty responses. Citations for all tweets used as examples can be found in Appendix A.

In (8), the lack of MV cues in each tweet indicated to hearers a sincere intent over 92% of the time on average. In (9), participants perceived sarcastic intent over 98% of the time on average due to the presence of MV cues (self-contradiction in all three as well as hyperbole in (9a)).

6. (a) Can't wait to be productive watching college cheer videos all day tomorrow. (1.00)
   (b) Field is an ice sheet and it's pouring rain?! #idealconditions (0.958)
   (c) This chapter for English started with a quote by Nietzsche "Every word is a prejudice"... Well this is gonna be fun! (1.00)

However, many participants detected the true intent of tweets in both categories, regardless of MV cues. As for sarcastic tweets with no MV cues, four of the fourteen tweets were correctly identified as sarcastic by a majority of participants, found in (10). There are a number of non-MV cues which may account for these outliers, or it may be possible that some participants found MV cues in these tweets which were not originally considered.

7. (a) Solid start to the week.. (0.944)
   (b) The world totally needs more Ben Stiller movies... (0.882)
   (c) Just what I wanted to hear, John Feinstein's opinions on Wrigley Field. (0.929)
   (d) Glad to see Obama start his bipartisanship off strong. (1.00)

The perception of sarcasm in (10a) may be due to the punctuation, an important sarcasm cue in written language, particularly social communication such as tweets (González-Ibáñez et al., 2011). In such communication, it is often easy to appear less enthusiastic than intended, so overt markers of positivity are required, where an unpunctuated sentence may be read as curt or otherwise negative. Thus, a sincere version of this tweet would likely have included an exclamation point or even an emoji to indicate a positive mood, while the lack thereof combined with the pseudo-ellipsis suggests a negative reading. The ellipsis in (10b) may have the same function, or it may be that participants believe Ben Stiller to be a widely disliked actor, suggesting a sentimental contradiction. Kreuz and Caucci (2007) proposed lexical hyperbole such as "totally" as a sarcasm cue, but strictly lexical hyperbole was not included in the current study. This may have contributed to the insignificance of hyperbole as an explanatory factor for sarcasm perception discussed in section 4.3.

The perception of sarcasm in (10c) identifies an idiosyncratic variety of sarcasm cues. Despite only 24% of participants reporting that they are well-informed about sports, all but one participant who evaluated this tweet correctly identified its sarcastic intent, suggesting that contextual cues cannot account for the anomaly. While the tweet does not contain any MV cues, it does contain what may be a grammaticalized sarcasm cue: "just what I wanted to hear." While one can say sincerely, "That's just what I wanted to hear!", any utterance that begins with "just what I wanted to hear" prefacing a description of some talk exchange sounds unquestionably sarcastic. The concept of grammaticalized sarcasm was introduced by Wilson, who used the example "fat chance" as an instance of sarcasm that has become lexicalized (2006, p. 1723). Further, while not quite lexicalized, some words or phrases do seem to be used fairly often in sarcastic utterances, such as "yeah, right" (discussed in more detail by Tepperman et al., 2006), as well as the "no, not at all" repetition and the "it's not like... (or anything)" construction, both discussed in 3.1. Another phrase that seems to appear more often in sarcastic tweets is "(so) glad"—found in (10d) as well as in three other sarcastic tweets and no sincere tweets. The use of these phrases, along with other cues, could indicate sarcastic intent to a hearer. The perception of sarcasm in (10d), similarly to (10b) and especially to (7b), may also be due to a potential sentimental contradiction if participants believed (or assumed) that Obama's 2015 State of the Union address had not satisfactorily displayed bipartisanship. As illustrated by (10b) and (d), the somewhat subjective nature of sentimental contradiction leads to difficulty in quantifying its effects.

V.3 Individual maxim-violation cues as explanatory variables

There were four noteworthy outliers as well for sincere tweets that contained MV cues, found in (11). Over 94% of participants presented with these tweets correctly identified them as sincere.

8. (a) Gonna be a good next few days, gonna be a good weekend, gonna be a good month, gonna be a great fucking year (0.0500)
   (b) Cant lie, 2015 has been good to me so far (0.0526)
   (c) Stressed through the roof that the weight of the rest of my life sits on the next 6 months of preparation (0.000)
   (d) Wikipedia is the greatest compendium of human knowledge ever compiled. (0.118)

(11a) was the only tweet in the study that contained repetition (a subset of manner violations) and no other MV cue. It is entirely possible, though not verifiable with our data, that repetition does not actually act as a sarcasm cue to hearers. It is also possible that repetition is usually a sarcasm cue, but that this particular brand of repetition thwarts that interpretation for some unknown reason. We simply do not have enough data to support or reject the inclusion of repetition under manner violation cues. (11b) through (d) all contain hyperbole. Again, it may be that hyperbole does not, in fact, act as a sarcasm cue—indeed, hyperbole was not found to explain a statistically significant amount of variance in sarcasm perception when compared to the entire body of tweets that did not contain hyperbole (p-value = 0.1421, alpha = 0.05). However, to build on the discussion in section 3.4, hyperbole was the cue with the fewest number of tweets in each category (11 total), which may have affected its significance level. Our results suggest that hyperbole is not used as a sarcasm cue, but further research is needed in this area. Just as with repetition, it may be that we do not have enough information to reject the use of hyperbole as a sarcasm cue.

9. (a) Stressed through the roof about preparing for this
   (b) The weight of the rest of my life sits on the next 6 months of preparation

The increased ease with which a sarcastic interpretation can be found for the utterances in (12) as opposed to (11c) can be explained by a corresponding decrease in complication for a meaning-inversion mechanism. That is, there are too many possible meaning-inverted interpretations of (11c); it is unclear which clause or subsection thereof requires meaning-inversion. In (12), there is a more obvious main thrust of the utterance which can be inverted: a sarcastic interpretation of (12a) suggests that the speaker is not stressed; (12b) may suggest that very little depends upon the speaker's preparation.

Thus, it is unclear whether hearers utilize hyperbole as an indicator of sarcastic intent. However, both self-contradiction and manner violations individually affect sarcasm perception at a statistically significant level (respectively: p-value = 0.0001, alpha = 0.05; p-value = 0.0391, alpha = 0.05). While these results may not be as robust as the overall effects of maxim-violation cues due to sample size, they do provide sufficient evidence that hearers use both self-contradiction and manner violations to identify sarcastic utterances as such.

V.4 Certainty scores

While examining participants' certainty scores alone would not lead to theoretically interesting discoveries, they can supplement our knowledge about the trend of sarcasm perception across categories and lend additional evidence for the importance of MV cues. On average, participants' confidence in their sarcasm assessments remained relatively stable regardless of the sarcastic intent in the tweet they were evaluating. However, tweets containing MV cues had significantly higher average certainty scores than tweets without MV cues (p-value = 0.024, alpha = 0.05). This suggests that the presence of such cues provides ample support for a sarcastic interpretation in a hearer's analysis of an utterance, but the absence of maxim-violation cues is not evidence *against* a sarcastic interpretation, but rather a deficiency of information. Hearers find sarcasm detection to be a more ambiguous task in the absence of MV cues, demonstrating that processing MV cues is an essential step in sarcasm detection in written language.

V.5 Limitations and areas for further research

It is also possible that some of the difference in average certainty scores is due to the fact that participants were expecting to be presented with sarcastic utterances, and thus felt more comfortable indicating a tweet to be sarcastic than sincere. The recruiting materials, initial questions regarding the definition of sarcasm, and the survey instructions all primed participants to interpret sarcastic tweets. Thus, participants would be expected to perceive more sarcasm on average than if the purpose of the study had been obscured, i.e., with filler questions like "Is the tweet negative, positive, or neutral?" (as suggested by González-Ibáñez et al., 2011). This imbalance did not seem to skew participants' actual sarcasm evaluations, however—on average, participants indicated that 51.8% of the tweets presented were sarcastic, which is to be expected if participants' evaluations are to be explained either by actual sarcastic intent or by the presence of MV cues. Thus, it seems unlikely that the transparent nature of the study negatively affected its results.

A few other logistical concerns may have negatively impacted the study's results: for instance, because all tweets were presented on the same page of the survey, participants were able to return to previous tweets and change their answers, which may have clouded initial judgments for some participants. The collection of sincere tweets with no MV cues was also not entirely reliable, in that their designation as tweets with sincere intent was based solely on my judgment, since this was the only category I did not collect from a particular hashtag. As discussed in section 4.2 regarding (10), the process of coding the tweets along the dimensions of the three MV cues was somewhat subjective, as well, particularly concerning sentimental contradictions. What constitutes a sentimental contradiction cannot be impartially determined since the cue itself relies on a notion of what is culturally considered a negative situation, which differs between people. Thus, what one person considers an unambiguous example of a sentimental contradiction, another might consider perfectly reasonable. The subjectivity of both factors discussed above can be mitigated by conducting a trial before the study proper which involves multiple participants determining which tweets are intended non-sarcastically (or actually producing non-sarcastic tweets), or in the latter case, indicating which cues they believe each tweet contains. Similar procedures can be found in González-Ibáñez et al. (2011) and in Kreuz and Caucci (2007).

Finally, the more thorough exclusion of confounding variables may have led to cleaner results. The exclusion altogether of punctuation could have minimized interference, but also may have caused some tweets to appear less positive than intended (as explained in section 4.2), skewing the results in the opposite direction. It may also have been beneficial to exclude, or somehow control for, tweets that reference particular background information. For example, the tweets with the highest certainty scores (and some of the highest proportions of perceived sarcasm) within the category of sarcastic tweets including MV cues were those that required no background knowledge, i.e. tweets that referenced getting an insufficient amount of sleep rather than those that referenced particulars of the State of the Union address. Thus, the omission of more specific tweets may have led to more robust results, as they would have completely eliminated contextual cues. Alternatively, an in-depth analysis of the impact of background knowledge even in limited-context settings such as Twitter may elucidate the role of contextual cues in sarcasm detection more generally. Although interjections did not seem to play a role in sarcasm perception in the current study, it may be wise to control for the presence of interjections as an additional potential explanatory variable in future research.

There are numerous directions with which to take future research in the area of sarcasm detection based on the current study. Experiments focused on only one of the three MV cues could find more targeted results regarding their individual significance as sarcasm cues. In particular, a more focused study could clarify the role of hyperbole or repetition as sarcasm cues, as the results of the current study were fairly inconclusive (and in the case of hyperbole, contradictory to the results of other, more robust studies). The inclusion of lexical hyperbole in future studies may also illuminate any intricacies in the use of hyperbole to detect sarcasm. The particular cues require further research in order to understand their role in other speech settings, as well, such as natural face-to-face interactions. Potential grammaticalized and lexicalized cues, such as "just what I wanted to hear" and "(so) glad," are missing almost entirely from the literature as it stands, so these cues are an area of research that remains wide open. Lastly, perhaps the most theoretically intriguing research question suggested by the current study concerns exploring the interaction of the meaning-inversion mechanism with variables such as utterance length, the amount of information in an utterance, the number of clauses or claims the utterance makes, etc. as a way of gaining a more detailed understanding of the meaning-inversion mechanism itself. Through these theoretically-based research questions, both theoretical and experimental linguistics can shed light on a model of sarcasm detection and interpretation.

Appendix

Table 4

Julz [idreaminchoc]. (2015, January 20). "Can't wait till the in laws come! Missed them a lot #notsarcasm" [Tweet]. Retrieved from https://twitter.com/idreaminchoc/status/557805478596857858

Kate [katejoanne93]. (2015, January 20). "So good to have the person you need stand by you when you feel down! #sarcasm #alone #depressing" [Tweet]. Retrieved from https://twitter.com/katejoanne93/status/557622427229097984

Takahata, J. [ciaociao808]. (2015, January 21). "2 hours of sleep is cool, I mean it's not like I have an interview and 4 classes today #sarcasm" [Tweet]. Retrieved from https://twitter.com/ciaociao808/status/557900185456873474

The Secret Psychic [SecretPsychic]. (2015, January 21). "What a great way to start off the day! Great session with an amazing woman! http://psychic.bitwine.com/a/44547r #bitwine" [Tweet]. Retrieved from https://twitter.com/SecretPsychic/status/557927823806300160

Example 7

(a) Wojtaszek, D. [phendrana]. (2015, January 22). "Anyone else excited about the impending Spruce Grove-St. Albert byelection? #amirite #sarcasm #abvote" [Tweet]. Retrieved from https://twitter.com/phendrana/status/558330317413441536

(b) Jean-Miche, S. [sintianiej]. (2015, January 21). "The speech last night president Obama's Was so wonderful for us #SOTU2015 !" [Tweet]. Retrieved from https://twitter.com/sintianiej/status/557933253135716353

Example 8

(a) Wolverine [ben_defbball40]. (2015, January 20). "SO nervous about the game tonight!!! 😁😂😁😂#sarcasm @MLeljedal @CalebMadl @Eggers_25 @Jezusfreek97 @John_Cuomo1" [Tweet]. Retrieved from https://twitter.com/ben_defbball40/status/557629227680661504

(b) See Kate (2015) in Table 4.

(c) Kayla [deleonk05]. (2015, January 22). "The weather is so pretty here in Corpus today. 😒😂#Sarcasm" [Tweet]. Retrieved from https://twitter.com/deleonk05/status/558323618786983936

Example 9

(a) Boulding, T. [tinaboulding]. (2015, January 19). "Can't wait to be productive watching college cheer videos all day tomorrow. #notsarcasm" [Tweet]. Retrieved from https://twitter.com/tinaboulding/status/557398825242218496

(b) Less Than Ultimate [LessUltimate]. (2015, January 19). "Field is an ice sheet and it's pouring rain?! #idealconditions #notsarcasm #HellYeah" [Tweet]. Retrieved from https://twitter.com/LessUltimate/status/557401609102434305

(c) Diver, E. J. [EmilieDiver]. (2015, January 19). "This chapter for English started with a quote by Nietzsche "Every word is a prejudice"... Well this is gonna be fun! #notsarcasm" [Tweet]. Retrieved from https://twitter.com/EmilieDiver/status/557427113494458368

Example 10

(a) Olson, L. [BigLeifDaddy]. (2015, January 20). “Solid start to the week.. #Sarcasm” [Tweet]. Retrieved from https://twitter.com/BigLeifDaddy/status/557635959937040384

(b) Pike, S. [sallypike12]. (2014, December 20). “The world TOTALLY needs more Ben Stiller movies... #sarcasm” [Tweet]. Retrieved from https://twitter.com/sallypike12/status/546566080143949824

(c) Sean [AngryDisneyNerd]. (2015, January 20). “Just what I wanted to hear, John Feinstein’sopinions on Wrigley Field. #sarcasm” [Tweet]. Retrieved from https://twitter.com/AngryDisneyNerd/status/557619506181582850

(d) Forsyth, D. [dforsyth47]. (2015, January 20). “Glad to see Obama start his bipartisanship off strong. #sarcasm” [Tweet]. Retrieved from https://twitter.com/dforsyth47/status/557739958686208003

Example 11

(a) mootzadell sticks [oheeyItsMike]. (2015, January 6). “Gonna be a good next few days, gonna be a good weekend, gonna be a good month, gonna be a great fucking year #notsarcasm #fullofpositivity” [Tweet]. Retrieved from https://twitter.com/oheeyItsMike/status/552672858078715904

(b) Ramos, J. [Bosco0812]. (2015, January 22). “Cant lie, 2015 has been good to me so far 😃 #NotSarcasm” [Tweet]. Retrieved from https://twitter.com/Bosco0812/status/558262480825483264

(c) High Quality H2O [KoryMiller3314]. (2015, January 11). “Stressed through the roof that the weight of the rest of my life sits on the next 6 months of preparation #ilovemylife #notsarcasm” [Tweet]. Retrieved from https://twitter.com/KoryMiller3314/status/554208247742078978

(d) Chung, Griffith [grifter1910]. (2015, January 9). “@fmanjoo Just like Wikipedia is the greatest compendium of human knowledge ever compiled. #notsarcasm” [Tweet]. Retrieved from https://twitter.com/grifter1910/status/553595264603537410

Works Cited

Attardo, S. (2000). Irony as relevant inappropriateness. *Journal of Pragmatics*, *32*, 793–826. doi:10.1016/S0378-2166(99)00070-3

Bach, K. (2005). The top 10 misconceptions about implicature. In B. Birner & G. Ward (Eds.), *Drawing the Boundaries of Meaning: Neo-Gricean Studies in Pragmatics and Semantics in Honor of Laurence R. Horn* (pp. 21–30). John Benjamins Publishing Company.

BBC. (2013). Being sarcastic. Retrieved November 20, 2014, from http://www.bbc.co.uk/worldservice/learningenglish/radio/specials/1210_how_to_converse/page13.shtml

Bryant, G. A., & Tree, J. E. F. (2005). Is there an ironic tone of voice? *Language and Speech*, *48*(3), 257–277.

Camp, E. (2012). Sarcasm, Pretense, and The Semantics/Pragmatics Distinction. *Nous*, *46*(4), 587–634.

Caucci, G. M., & Kreuz, R. J. (2012). Social and paralinguistic cues to sarcasm. *Humor*, *25*(1), 1–22. doi:10.1515/humor-2012-0001

Clark, H. H., & Gerrig, R. J. (1984). On the pretense theory of irony. *Journal of Experimental Psychology: General*, *113*(1), 121–126. doi:10.1037/0096-3445.113.1.121

Colston, H. L., & O'Brien, J. (2000). Contrast and pragmatics in figurative language: Anything understatement can do, irony can do better. *Journal of Pragmatics*, *32*, 1557–1583. doi:10.1016/S0378-2166(99)00110-1

Filik, R., & Moxey, L. M. (2010). The on-line processing of written irony. *Cognition*, *116*(3), 421–436. doi:10.1016/j.cognition.2010.06.005

Giora, R., & Fein, O. (1999). On understanding familiar and less-familiar figurative language. *Journal of Pragmatics*, *31*, 1601–1618. doi:10.1016/S0378-2166(99)00006-5

Giora, R., Fein, O., Laadan, D., Wolfson, J., Zeituny, M., Kidron, R., … Shaham, R. (2007). Expecting Irony: Context Versus Salience-Based Effects. *Metaphor and Symbol*, *22*(2), 119–146. doi:10.1080/10926480701235346

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* (pp. 581–586). Portland, Oregon: Association for Computational Linguistics.

Grice, H. P. (1975). Logic and Conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and Semantics 3: Speech Acts* (pp. 41–58). New York: Academic Press.

Grice, H. P. (1978). Further Notes on Logic and Conversation. In P. Cole (Ed.), *Syntax and Semantics 9: Pragmatics* (pp. 13–27). New York: Academic Press.

Jorgensen, J., Miller, G. a., & Sperber, D. (1984). Test of the mention theory of irony. *Journal of Experimental Psychology: General*, *113*(1), 112–120. doi:10.1037/0096-3445.113.1.112

Kreuz, R. J., & Caucci, G. M. (2007). Lexical influences on the perception of sarcasm. In *Proceedings of the Workshop on Computational Approaches to Figurative Language* (pp. 1–4). Rochester, New York: Association for Computational Linguistics. doi:10.3115/1611528.1611529

Kreuz, R. J., & Glucksberg, S. (1989). How to be sarcastic: The echoic reminder theory of verbal irony. *Journal of Experimental Psychology: General*, *118*(4), 374–386. doi:10.1037/0096-3445.118.4.374

Kreuz, R. J., & Roberts, R. M. (1995). Two cues for verbal irony: Hyperbole and the ironic tone of voice. *Metaphor and Symbolic Activity*, *10*(1), 21–31.

Pexman, P. M., & Zvaigzne, M. T. (2004). Does Irony Go Better With Friends? *Metaphor and Symbol*, *19*(2), 143–163.

Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., & Huang, R. (2013). Sarcasm as Contrast between a Positive Sentiment and Negative Situation. *EMNLP*, 704–714.

Roberts, R. M., & Kreuz, R. J. (1994). Why do people use figurative language? *Psychological Science*, *5*(3), 159–163. doi:10.1111/j.1467-9280.1994.tb00653.x

Rockwell, P. (2000). Lower, slower, louder: Vocal cues of sarcasm. *Journal of Psycholinguistic Research*, *29*(5), 483–495. doi:10.1023/A:1005120109296

Rockwell, P. (2005). Sarcasm on television talk shows: Determining speaker intent through verbal and nonverbal cues. In A. V. Clark (Ed.), *Psychology of Moods* (pp. 109–122). New York City: Nova Science Publishers, Inc.

Ryder, N., & Leinonen, E. (2014). Pragmatic Language Development in Language Impaired and Typically Developing Children: Incorrect Answers in Context. *Journal of Psycholinguistic Research*, *43*, 45–58. doi:10.1007/s10936-013-9238-6

Schwoebel, J., Dews, S., Winner, E., & Srinivas, K. (2000). Obligatory Processing of the Literal Meaning of Ironic Utterances: Further Evidence. *Metaphor and Symbol*, *15*(1&2), 47–61. doi:10.1207/S15327868MS151&2_4

Sperber, D., (1984). Verbal irony: Pretense or echoic mention? *Journal of Experimental Psychology: General*, *113*(1), 130–136. doi:10.1037/0096-3445.113.1.130

Sperber, D., & Wilson, D. (1981). Irony and the Use-Mention distinction. Retrieved from http://discovery.ucl.ac.uk/1331930/

Surian, L., Baron-Cohen, S., & Van der Lely, H. (1996). Are children with Autism deaf to Gricean maxims? *Cognitive Neuropsychiatry*, *1*(1), 55–71. doi:10.1080/135468096396703

Tepperman, J., Traum, D., & Narayanan, S. (2006). "Yeah right": Sarcasm recognition for spoken dialogue systems. In *Interspeech* (pp. 1838–1841). Pittsburgh: ICSLP.

Toplak, M., & Katz, A. N. (2000). On the uses of sarcastic irony. *Journal of Pragmatics*, *32*, 1467–1488. doi:10.1016/S0378-2166(99)00101-0

Utsumi, A. (2000). Verbal irony as implicit display of ironic environment: Distinguishing ironic utterances from nonirony. *Journal of Pragmatics*, *32*, 1777–1806. doi:10.1016/S0378-2166(99)00116-2

Wilson, D. (2006). The pragmatics of verbal irony: Echo or pretence? *Lingua*, *116*, 1722–1743. doi:10.1016/j.lingua.2006.05.001

Wilson, D., & Sperber, D. (2002). Relevance theory. *UCL Working Papers in Linguistics*, *14*, 249–290. doi:10.1075/pbns.37