UNIVERSITY OF CALIFORNIA
SANTA CRUZ

EVIDENCE FOR DIVERSE FORMS OF SARCASM

A dissertation submitted in partial satisfaction
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

PSYCHOLOGY

by

J. Trevor D'Arcey

December 2020

The dissertation of J. Trevor D'Arcey is approved:

_____
Professor Jean E. Fox Tree, chair

_____
Professor Nicolas Davidenko

_____
Professor Leila Takayama

_____
Quentin Williams
Acting Vice Provost and Dean of Graduate Studies

TABLE OF CONTENTS

List of Tables & Figures

Abstract

Evidence for Diverse Forms of Sarcasm

J. Trevor D'Arcey

Sarcasm is a difficult concept to define accurately and completely and is similarly hard to identify in natural communication. In three works, this dissertation develops a deeper understanding of sarcasm, both as a concept and as a phenomenon. Using a computational approach, I describe novel markers of sarcasm that function as signals for the reader to pause, wait, or slow down, and show that they are often copresent with sarcastic content. Using experimental methods, I show that although sarcasm can be elicited in the laboratory at high levels, agreement on what statements are sarcastic is extremely unreliable. Finally, using a qualitative approach, I ask participants to modify statements to make them sarcastic, and consider the strategies they report using, as well as their success at doing so. Overall, my results suggest that sarcasm is an extremely diverse phenomenon, and that future research should adjust its focus to use a broader view of sarcasm, rather than using a specific definition.

Acknowledgements

Countless collaborators made this work possible, and without all of their support it just would not have happened. I feel great gratitude for all the time, energy, and thought that they poured into this work. I am especially indebted to the eighteen research assistants whom I was fortunate to work with over the past five years. They are Ian Bicket, Joaquin Canizales, Alea Casanova, Paige Collazo, Ericka Elphick, Sara Eslami, Sean Gardner, Jasmin Granke, Bronwyn Hassall, Emilie Kovalik, Nicole Laflin, Valerie Muñoz, Sonia Perez Lemus, Justin Siegel, Emily Truong, Emma Wexler, Elizabeth Williams, and Evelyn Yap. You are all wonderful, and I cannot express how much I enjoyed learning and growing with you.

Thank you to my labmates Jackson Tolins, Kris Liu, Alina Larson, Alicia Hammond, Yasmin Chowdhury, Allison Nguyen, Andrew Guydish, Vanessa Oviedo, and Elise Duffau for adding your friendship, advice, and humor to the difficult moments that are inherent to the graduate school experience.

To my committee members, Nick Davidenko, Jean E. Fox Tree, Jason Samaha, Leila Takayama, and Steve Whittaker, thank you for the amazing conversations and for all the time you have put into making my work better. You helped me refine my exciting but vague ideas into focused, clear theses that I became even more excited about. You all inspire me to focus my attention, to find efficient and effective paths to success, to stand up for my beliefs, to not be afraid to dig into complex problems, and to use kindness to encourage greatness.

To my coauthors Jean E. Fox Tree, Alicia Hammond, Alina Larson, and Shereen Oraby, thank you for collaborating to make this amazing work come to life. Taking an idea from concept to publication is a much longer process than I expected it to be, and I'm so grateful for your willingness to continue working on projects even years later.

To Travis Seymour, Doug Bonett, and Alan Kawamoto, who gave tremendous amounts of their time to help with various elements of my projects, including ensuring methodological rigor, planning statistical analyses, and helping me learn various amounts of four and a half programming languages.

To my Master of Arts program advisor Mike Ennis, and committee members Penelope Kuhn and Lawrence Herringer, who encouraged me to apply to graduate school in the first place, taught me about psychophysiology, neuroscience, emotion, and personality, and showed me the joy of pursuing and interrogating new knowledge in methodical ways.

A huge thank you to Katherine Montano, Paul Sosbee, and Allison Land, whose friendly help navigating the complex university systems and requirements directly allowed me to finish this work in a timely manner while maintaining some sanity.

To Jean E. Fox Tree, who has been a more amazing PhD advisor than I believed could exist. Jeannie, thank you from the bottom of my heart for your mentorship, guidance, and friendship. Words will never begin to express how lucky I

feel to have worked with you over the past five years. I learned a tremendous amount about who I want to be, and I will always be grateful for that.

To my family and friends, who brought me here and supported me through it, even by allowing me the latitude to disappear for extended periods of time. Thank you for tolerating many extremely late replies to phone calls and messages without question.

To Elena, who consistently demanded I keep applying to graduate schools, and who cheerfully put my work ahead of everything else when needed. The time you spent to make this dissertation happen is invisible to the reader, but you should definitely be a coauthor by the amount of support you've given me.

Thank you to all participants who participated in these many studies over the last five years, without whom this work could not exist. And finally, thank you to Taqueria Los Pericos, for sustaining me across all five levels of Maslow's Hierarchy of Needs.

The text of this dissertation includes reprints of the following previously published material:

D'Arcey, J. T., Oraby, S., & Fox Tree, J. E. (2019). Wait Signals Predict Sarcasm in Online Debates. *Dialogue & Discourse*, *10*(2), 56–78.

Fox Tree, J. E., D'Arcey, J. T., Hammond, A. A., & Larson, A. S. (2020). The Sarchasm: Sarcasm Production and Identification in Spontaneous Conversation. *Discourse Processes*, 57(5–6), 507–533. https://doi.org/10.1080/0163853X.2020.1759016

The co-author listed in this publication (Jean E. Fox Tree) directed and supervised the research which forms the basis for the dissertation.

EVIDENCE FOR DIVERSE FORMS OF SARCASM

1 Introduction

Misunderstandings in communication are so common that entire industries have sprung up in order to prevent, identify, and remedy them. Not only does legal experts' work often center on interpreting others' language and coming to agreement on what was meant, but therapists, arbitrators, mediators, negotiators, and authority figures of various types also share the responsibility of interpreting both sides of a story and helping people in different situations begin to share common ground. Why do misunderstandings occur so frequently, after millennia of language development? No doubt language's natural evolution is responsible for at least some misunderstandings, especially between people of different generations (e.g., a high schooler may refer to something as "sick" to make a statement of positive sentiment, which could be misinterpreted by an elderly person as a statement of negative sentiment) and different cultures (e.g., a football coach may describe a political candidate's behavior as being "in the red zone", suggesting he is close to scoring a goal, while a car racer might interpret this statement as suggesting he is close to destroying his engine). But language's lack of universal comprehensibility also serves an important function as a way to facilitate social group creation, both by establishing closeness as well as by establishing distance from others (Gallois et al., 2005).

When people share similar views of the world, they are more likely to be friends (Parkinson et al., 2018). Language's diversity, therefore, may be used as an interpersonal thermometer for people to gauge how well they will get along with

someone, as part of rapport-building processes. An important part of rapport building is showing similarity between one's interlocutor and oneself (Duffy & Chartrand, 2015), and similarity in language use may act as a cue toward similarity in worldview. It is this researcher's hypothesis that people actively test worldview similarity through various mechanisms, some of them linguistic. Among the most potent of these mechanisms is sarcasm.

Sarcasm in conversation creates a strong test of worldview similarity precisely because its success as a form of humor relies on its ambiguity. Whether dry or dripping, the question, "Are you being sarcastic?" regularly follows sarcasm. And this is for good reason, as evidence suggests that being unable to resolve the ambiguity is more common than not (Fox Tree et al., 2020). However, when interlocutors do understand each other's sarcasm, it creates the opportunity for them to connect on a much deeper level than just knowing that they both like the same type of cake -- they share a similar worldview. Likewise, interlocutors with a different worldview are unlikely to get the joke. It is the subtle display of sarcastic intent only to those who are meant to understand, and actually do understand, that makes it a powerful way to build rapport. In a sense, it affirms that interlocutors share something special that is not shared by others.

Since sarcasm's success as a social group facilitator is contingent upon its ability to be accessible exclusively to those with a similar worldview, it is no surprise that its expression is just as diverse (if not more so) than other linguistic phenomena -- expressing a sentiment so that it will be accessible to one worldview but not to

another requires diverse thinking. And as a result, research on sarcasm never quite seems to hit the bullseye: researchers have doggedly studied it from the perspective of its linguistic and pragmatic attributes, attempting to pin it down with definition after definition only to find their definitions generally accurate but insufficient to show a commonality across all sarcasm.

In this dissertation, I present three chapters that together encourage understanding sarcasm from many viewpoints, including considering the social functions it serves. Chapter 1 uses a computational approach to show evidence of novel markers of sarcasm in online discussion boards (D'Arcey, Oraby, & Fox Tree, 2019). It presents a psycholinguistic theory for why these markers co-occur with certain types of sarcasm. Chapter 2 discusses the connections between sarcasm and rapport-building, demonstrates a novel rapport-based procedure for eliciting sarcasm in the laboratory, and discusses why sarcasm may be misunderstood more commonly than it is understood (Fox Tree, D'Arcey, Hammond, & Larson, 2020). Chapter 3 presents mixed methods work showing that sarcasm too often reflects what researchers make of it, and advocates for a more inclusive idea of sarcasm -- one which leverages public knowledge of sarcasm in order to form a more complete picture of its use (D'Arcey & Fox Tree, under review). In the discussion, I discuss the impact of each of these three approaches on the field of sarcasm research, and argue for novel, creative, and contextual approaches to the continued study of sarcasm.

2 Chapter 1: Wait Signals Predict Sarcasm in Online Debates

2.0 Pre-Introduction

In this chapter, I present results from a published paper, Wait Signals Predict Sarcasm in Online Debates (D'Arcey & Fox Tree, 2019), in which we show several textual patterns that tend to co-occur with sarcasm in debate forums online. The connecting force between many of these patterns is that they all ask the reader to *wait* -- something that, on its surface, seems useless in an already asynchronous communication medium. Much as one would not send an email to a fire department to notify them of a fire, writing a post on a forum online is similarly a poor way to ask someone to wait. We argue that writers provide a sarcasm cue to readers by leveraging the pragmatic incongruity inherent in requesting a delay using an asynchronous medium.

2.1 Abstract

We examined the predictive value of wait signals for sarcasm in online debate forums. In Study 1, we examined the word frequency of *um* and *uh* across six corpora. In general, there were far more of these fillers in spoken corpora than written corpora. We also found that the proportion of *um*s to *uh*s varied by corpus type. In Study 2, we tested whether the inclusion of *um* or *uh* at the beginning of online debate forum posts led to higher probability of those posts being classified as sarcastic by Amazon Mechanical Turk workers. We found that posts beginning with these items were twice as likely to be labeled sarcastic. In Study 3, we tested fillers and ellipses in the middle of posts. We found that posts including these items were approximately

three to five times more likely to be labeled sarcastic. We compared results to other signals like the word *obviously* and quotation marks. Signals that indicate delay in written communication cue readers to non-literal meaning.

2.2 Introduction

Non-literal language use is common in communication, both in speech (Gibbs, 2000; Glucksberg, Gildea, & Bookin, 1982) and writing (Whalen, Pexman, & Gill 2009; Walker, Fox Tree, Anand, & King, 2012). One form of non-literal language is sarcasm, in which people's intended meaning contrasts with the literal, semantic meaning of their words. People can use sarcasm to mock or to be funny (Kreuz, Long, & Church, 2009), to affirm and modify social relationships (Seckman & Couch, 1989), and to help a friend save face (Jorgensen, 1996). Fluency with sarcasm and other forms of humor is an important social skill that predicts a variety of positive social outcomes such as peer reputation in children (Masten, 1986) and ability to cope with stress in adults (Overholser, 1992). Creating tools with the ability to recognize sarcasm would have wide-reaching benefits for these groups.

Yet identification of sarcastic content is notoriously elusive for both people (Rockwell, 2000; Burgers, Van Mulken, & Schellens, 2011) and machines (Reyes & Rosso, 2014; Riloff, Qadir, Surve, De Silva, Gilbert, & Huang, 2013; Felbo, Mislove, Søgaard, Rahwan, & Lehmann, 2017). A number of cues to sarcastic content have been identified, but one that has not been fully explored is the use of fillers like *um* and *uh* and spontaneously written versions of spoken pauses like ellipses. *Um* and *uh* (*er* and *erm* in British English) have been shown to be used by speakers to notify

interlocutors of an upcoming delay in speech (Smith & Clark, 1993), and ellipses typically indicate an omission or pause in writing. We propose that these phenomena are used as wait signals in writing. These wait signals operate to change the pacing at which a text is read, thereby introducing novel pacing in the reader's mind and potentially delaying delivery for dramatic effect. Dramatic pacing can be observed in the following: "The watch-word here is 'big': big guitar-licks, big melodic surges, big-hearted words and, erm … big blokes" (from the British National Corpus, CK5/3128). Wait signals and sarcasm can be observed in the following: "Yeah, I'll ....uh keep that in mind dude....trust me!" (from the Internet Argument Corpus, Walker et al., 2012). In this report we document the use of wait signals as indicators of sarcasm in writing.

2.3 Identifying Sarcasm

We begin our discussion of sarcasm by noting that it may be futile to try to experimentally differentiate sarcasm from irony, regardless of whether raters are trained to do so (Attardo, Eisterhold, Hay, & Poggi, 2003). Irony is using language to mean something other than what the words literally express, such as saying, "I'll keep that in mind," while meaning, "I most definitely will not keep that in mind." Sarcasm is often thought of as adding a negative connotation to the irony, such as by targeting a victim; for example, by saying "nice hair" to someone with a bad haircut (Cambpell & Katz, 2012, p. 460). Despite these definitions, most researchers are in agreement that the two concepts are difficult to differentiate. To further complicate matters, the word sarcasm may be becoming more prevalent as a replacement for irony (Nunberg,

2001), suggesting that to the layperson, the concepts may be interchangeable. When we use the term irony in this work, it is because the research we are referencing uses this term. For all other instances we use the label sarcasm because it is more readily understood (Bryant & Fox Tree, 2002), while acknowledging the fact that researchers generally agree they are separate constructs. To this end, in our present research, we were explicit in defining sarcasm for participants as:

> 1: a sharp and often satirical or ironic utterance designed to be humorous, snarky, or mocking.
> 2: a mode of satirical wit depending for its effect on bitter, caustic, and often ironic language that is often directed against an individual or a situation.

Participants were also given examples of statements with and without sarcasm:

> **With sarcasm**: "Yes, you are 100% correct. Criminals would be sure to pay the tax on their illegally owned pistol, just like they pay income tax on drug money. Oh, wait they don't pay tax on their drug money. Most criminals break the law you see."
> **Without sarcasm**: "The article said very little about his observations and almost nothing about his methods."

Our goal was to be as clear as possible, although it is well known that defining these concepts is difficult.

2.3.1 Human Sarcasm Identification

Perhaps anticipated by the challenges in defining sarcasm, people have a hard time agreeing on whether statements are sarcastic. Individual (Akimoto & Miyazawa, 2017; Ivanko, Pexman, & Olineck, 2004; Rockwell & Theriot, 2001) and regional (Dress, Kreuz, Link, & Caucci, 2008) variations in the conception of sarcasm exacerbate this challenge. For example, political beliefs can affect how satire from late night comedy routines is interpreted (LaMarre, Landreville, & Beam, 2009).

Even if individual, regional, and political backgrounds are held constant, interpretation of sarcasm can vary based on the context presented with the sarcastic utterance. Context can make an originally sincere utterance appear sarcastic and vice versa (Bryant & Fox Tree, 2002). Although there are challenges to identifying sarcasm, under some circumstances, people can be quite good at detecting it. In a study of tweets originally marked with #sarcasm compared to those which were not, people could correctly identify which were marked sarcastic about 70% of the time when the hashtags were removed (Kovaz, Kreuz, & Riordan, 2013).

Raters' misidentification of sarcasm has led researchers to develop explicit, rigorous procedures to achieve high inter-rater reliability on ratings of sarcasm and irony. One such method, the Verbal Irony Procedure (Burgers et al., 2011), found high reliability for film reviews by asking raters to engage in a four step process: first to read the entirety of the review and determine the author's overall stance, second to remove purely descriptive utterances (which, it is assumed, never contain verbal irony), third to remove utterances that have a literal evaluation that fits with the overall stance, and fourth to construct scales of evaluation for the remaining (possibly ironic) utterances in which the literal evaluation of each utterance can be compared to the rater's perception of the writer's intent. Utterances which contrast are coded as ironic. With this procedure, the authors achieved very strong agreement (97.3%) between two coders (Burgers, Van Mulken, & Schellens, 2011). However, this method may not apply as well to less explicitly evaluative texts — film reviews are, by their nature, usually quite expressive.

2.3.2 Machine Sarcasm Identification

On the other hand, methods to computationally identify sarcasm are improving as deep learning techniques are put into broader use. Nonetheless, the best models are still unable to agree with people on what's sarcastic, whether it is spoken or written. One issue is that the rates of sarcasm in corpora are generally sparse, hovering around 10% (e.g., Gibbs, 2000; Walker, et al., 2012), leading to more difficulty in measuring classifier success in natural language processing research.

In the field of natural language processing, many researchers studying imbalanced classification problems like sarcasm identification measure their models' success using two metrics: The first, recall, is defined as the percentage of sarcastic occurrences that the model correctly identifies. For example, in a set of 1,000 internet posts, 100 may include sarcasm. If the model identifies 80 of the 100 sarcastic posts, its recall is .8. The second measure, precision, is defined as the percentage of model-identified sarcastic occurrences that are actually sarcastic. So, if the aforementioned model correctly identified 80 sarcastic posts, but it also incorrectly labeled another 80 posts as sarcastic, its precision is .5. These are important as separate constructs in imbalanced classification tasks because both a large rate of false positives and a large rate of false negatives are important to understanding the model's performance. Recall and precision are frequently combined into a single measure of performance, F1, defined as the harmonic mean. The harmonic mean is used to avoid rewarding models in which either recall or precision is close to perfect, at the expense of the other (Koehrsen, 2018).

State-of-the-art classifying models (see, for example, Felbo, Mislove, Søgaard, Rahwan, & Lehmann, 2017; Ghosh & Veale, 2016; Poria, Cambria, Hazarika & Vij, 2016) achieve a wide range of F1 scores depending on the text being analyzed and the model being used. Felbo et al. (2017) developed the DeepMoji model for sarcasm detection using publicly released debate forums data from Walker et al. (2012) and Oraby et al. (2016) using around 1,000 training examples and achieving F1 scores of 0.69 and 0.75, respectively. Ghosh and Veale (2016) performed sarcasm classification in the Twitter domain. They first constructed a dataset of 39K tweets, 18K sarcastic and 21K non-sarcastic. They collected the sarcastic class by using "positive markers of sarcasm" — hashtags such as #sarcastic and #yeahright (Ghosh & Veale, 2016, p. 3). The non-sarcastic class were tweets lacking these hashtags. Training was performed on tweets after the relevant hashtags were removed. Ghosh and Vale (2016) achieved an F-score of 0.92 on their test set using a convolutional neural network (CNN) model with a Long Short-Term Memory layer (LSTM) and deep neural network (DNN). They also verified their model on other existing datasets: for Riloff et al.'s (2013) test set of 3K tweets, they reported an F1 of 0.88 as compared to the baseline result of 0.51, and for Tsur et al.'s (2010) test set of 180 product reviews, they reported an F1 of 0.90 — which was higher than the previously reported best result of 0.83. In addition to these models, Poria et al. (2016) reported F1 scores using a deep convolutional neural network on two datasets from Ptacek et al. (2014): 0.98 F1 on a balanced dataset of 100K tweets, and 0.95 F1 on an unbalanced dataset (25K sarcastic and 75K non-sarcastic).

2.4 Cues for Sarcasm in Speech and Writing

Although it is challenging to identify, sarcasm is common in speech and writing. In a study of 62 10-minute conversations, for example, Gibbs (2000) and two student judges agreed that at least 289 utterances were ironic (8% of the corpus), of which 80 were deemed to be sarcastic. In another study of communication under irony-inducing conditions — describing badly-dressed celebrities and planning meals for a disliked guest — over 10% of the turns produced were ironic (Hancock, 2004). Of forty dyads who communicated in face-to-face and instant messaging conversations, only one dyad did not produce any ironic utterances (Hancock, 2004). Irony induction is not necessary to observe sarcasm in writing. In a study of 105 people's emails to close friends, almost all contained non-literal language, averaging almost three per email (Whalen, Pexman, & Gill 2009). Even in academic genres, where one might expect to find more straightforward language use, sarcasm is frequent and widespread (Lee, 2006).

2.4.1 Cues for Sarcasm in Speech

Despite the fact that it is a common phenomenon, cues for spoken sarcasm are not particularly straightforward. Pop culture places emphasis on prosody to convey sarcasm, but a close examination of prosodic tone did not show consistent patterns for an ironic tone of voice (Bryant & Fox Tree 2005), although people could differentiate between talk radio utterances that were originally produced as sarcastic and those that were originally produced as sincere (Bryant & Fox Tree 2002). Noting that a target utterance contrasts with surrounding talk is another key way people determine what is

sarcastic (Attardo et al., 2003). Some specific cues that can be used in spoken communication include facial cues like smiles, laughter, and slow nods, which are more likely to co-occur with sarcastic utterances (Caucci & Kreuz, 2012), and air-quotes gestures, which can be used to indicate irony or sarcasm (Lampert, 2013). Eye gaze towards (Caucci & Kreuz, 2012) or away (Williams, Burns, & Harmon, 2009) from addressees can also predict sarcasm, perhaps depending on how prepared the sarcastic utterances are (they were not prepared in advance in Caucci & Kreuz but they were in Williams et al., although there may be other differences). In the Switchboard corpus of spoken dialogue, 23% of occurrences of the phrase yeah right indicated sarcasm (Tepperman, David, & Narayanan, 2006). Contextual cues and regional differences may influence generation and perception of spoken sarcasm as well: common ground between interlocutors led to more sarcasm (Caucci & Kreuz, 2002; Clark, 1996), and Southern U.S. participants' viewed sarcasm as more hurtful (Dress et al., 2008).

2.4.2 Cues for Sarcasm in Writing

Written sarcasm cannot take advantage of the multimodal auditory, facial, and bodily cues that can help in identifying spoken sarcasm. But written sarcasm does have textual cues that are not available for spoken sarcasm, such as exclamation points and question marks, laughter expressions like *lol*, emoticons like *:)*, and quotation marks, all of which contribute to irony detection (Carvalho, Sarmento, Silva, & De Oliveira, 2009; this study was done in Portuguese). As with speaking, contrast is also important with writing. A frowning emoticon matched with an

apparently positive message conveys sarcasm (Derks, Bos, & Von Grumbkow, 2008b). Contrast between positive and negative sentiment can also be used to identify sarcasm (Riloff et al., 2013).

Many words and phrases that indicate sarcasm have also been identified. In a corpus of written arguments, the phrases included *let's all* (62% sarcastic), *I love it when* (56% sarcastic), *oh really* (50%) and *I'm shocked*/*amazed*/*impressed* (42%; Oraby, Harrison, Reed, Hernandez, Riloff, & Walker, 2016). In writing, the inclusion of the word *really* in online debate posts made the probability that the post-response pair will be perceived as sarcastic approximately double (Walker et al., 2012). Other markers of sarcasm included *you mean* and *so* (Walker et al., 2012). In a study of 100 lines from books that were introduced by "said sarcastically," the presence of interjections such as *well* and *uh* were predictive of sarcastic content (Caucci & Kreuz, 2012). Sarcastic lines in books marked by *said sarcastically* and sarcastic tweets marked by #sarcasm had more positive emotion words than non-marked lines (Kovaz et al., 2013). This converges with the idea that sarcasm is more commonly used to evaluate a negative situation with positive affect (Clark & Gerrig, 1984).

2.4.3 Contrast Between Sarcasm in Speech and Writing

Some of the words identified as sarcastic in writing are not generally associated with sarcasm when they are spoken. *Well*, for example, is understood to mean that there is a mismatch between what follows and what's expected (Blakemore, 2002; Jucker, 1993), suggesting a less-obvious interpretation (Fox Tree, 2010), such as a dispreferred interpretation (Holtgraves, 2000). This meaning of *well*

aligns well with its potential use in sarcasm, as sarcasm is intended to represent something beyond the literal meaning. When coupled with *said sarcastically*, and grouped with other interjections, written *well*s found in books were predictive of sarcasm (Caucci & Kreuz, 2012). However, when 605 turn-initial *well*s in spontaneously written debates were compared to unmarked turns, those marked by *well* were not more sarcastic (Fox Tree, 2015).

Um and *uh* are also not generally associated with sarcasm when they are spoken. *Um* and *uh* indicate upcoming delay in speaking (Clark & Fox Tree, 2002; Fox Tree, 2001). Delays marked by *um* are different from silent pauses without *um*s; marked delays are more indicative of speech production trouble and are also associated with lack of comfort with the topic under discussion and dishonesty (Fox Tree, 2002). But it is important to note that the *um* itself does not mean that speech production difficulty, discomfort, or deception will necessarily follow. In a study of 35 people's self-assessment of the meaning of *um* and *uh*, for example, no one indicated it meant deception (Fox Tree, 2007). No one indicated it meant sarcasm either (Fox Tree, 2007). Like in speaking, *um*s and *uh*s in writing can also indicate a kind of delay, a need to think, such as to answer a question (Fox Tree, Mayer, & Betts, 2011; Fox Tree, 2015) as in "So, uh... what movie is everyone talking about? I don't think I've seen any previews" (from the Internet Argument Corpus, Walker et al., 2012). But also, as with speaking, in none of these prior studies were spontaneously written *um*s or *uh*s proposed to indicate sarcasm.

Why do specific n-grams (textual patterns of variable length, here defined as patterns of one or more words) like *let's all*, *really*, and *you mean* contribute to sarcastic perceptions? One potential explanation is that they are used to call attention to an incongruity, with incongruity being one way nonliteral language is flagged. For example, the incongruity between the body and last lines of a news story suggests that it is satire (Rubin et al., 2016). It could be that the incongruity is flagged by the words themselves, if the words are uncommon contextually. For example, slang is not expected in news stories, and has been shown to indicate satire (Burfoot & Baldwin, 2009). As another example, transforming written quotes to the face-to-face modality as air-quotes gestures sets up an incongruity (because quotes are typically written, not enacted), and it could be this incongruity that flags the sarcasm that air-quotes suggest (cf. Lampert, 2013). Similarly, using a quote for a single word in writing (e.g., thanks for the "advice") may also indicate sarcasm because it is a noncanonical usage (in writing, quotes are usually used to indicate a direct report of speech, which is usually more than one word long). That is, a word out of context may cue non-literal meaning.

2.5 Fillers and Ellipses as Signals of Sarcasm

Many signals appropriate for speech are not as useful for written communication. A hand gesture may be communicative when directed at a driver who cuts off other drivers but is less likely to be communicative when directed at a forum troll who belittles an argument. In addition to gestures, another group of signals that may not have as much value among asynchronous writing is requests to

wait for production to continue. Unlike face-to-face communication with a waiting addressee, spontaneous writing often takes place asynchronously. The composition process is not observed keystroke by keystroke, as we observe speakers phoneme by phoneme. Instead, writers generally finish their messages prior to sharing their product. In writing, as opposed to speaking, there are usually fewer costs to lack of timeliness (Fox Tree, 2015). This asynchrony means that there are not as many reasons to ask addressees to wait or to inform them of an upcoming pause when writing. In a sample of 44 students' spoken and text conversations, *um*s and *uh*s were nine times more common in speaking (Fox Tree, 2015).

But although they were less common, *um*s and *uh*s and other signals of time did still occur in writing. We define wait signals in writing as tools used by writers to pace readers' consumption of information. They include *um*s and *uh*s (which can be spelled in numerous ways), ellipses, parentheses that indicate asides, em-dashes, and other markers. In asynchronous writing, wait signals should be expected to be less prevalent than fillers and pauses in speaking. But their lack of prevalence may imbue them with additional significance when they are used. Whereas wait signals are not traditionally associated with sarcasm in speech, we propose that wait signals suggest to readers that they take more time with the information that follows them, with the additional time leading to non-literal interpretations. One definition of wit from the Oxford English Dictionary is, "A natural aptitude for using words and ideas in a quick and inventive way to create humour" (Oxford English Dictionary, n.d.). Our hypothesis is that the use of traditional wait signals in contexts where wait signals

have limited use for signaling a pause constitutes one form of wit, or using words in inventive ways. As hearing *um* at the beginning of a turn leads listeners to consider that the speaker is having production trouble, discomfort with the topic, or is preparing a dishonest answer, so too can reading *um* suggest that writers are intending something different from what they've literally written, such as that they are being sarcastic. It is both the unexpectedness of the wait signal in writing as well as the extra processing suggested by the wait signal that drives the sarcastic interpretation. Whalen, Pexman, and Gill (2009) suggested something similar for non-filler wait signals: "Hyphens, parentheses, and ellipses could be construed as a category of 'text-separators,' used to segment portions of the text to assist the reader in detecting those portions that are to be interpreted non-literally" (pp. 275-276).

In support of this hypothesis, we note that the highest predictor of sarcasm in a study of a variety of textual cues to sarcasm was *oh wait*, at 87% (Oraby et al., 2016). While *oh* on its own has been linked to sarcasm and negative emotion in writing (Abbott et al., 2011; Fox Tree, 2015), the predictiveness of *oh wait* is much higher than *oh* in combination with other words such as *oh really* or *oh yeah* (both 50%, Oraby et al., 2016). Not all *oh*s are sarcastic. In speaking they can indicate arrival at revised interpretations or state change (Heritage, 1984) which can be used strategically, such as to politely show newsworthiness in comparison to responding with a *yes* (Fox Tree & Schrock, 1999). The revised interpretations can also be used sarcastically to imply that something is newsworthy when it is not (Fox Tree, 2015). *Oh* has both attitudinal and cohesive functions, functions that differ from temporally

sensitive markers like *um* and *uh* which are much more common in synchronous communication (Fox Tree, 2015). The rate of *oh* production is similar in spontaneous speech and spontaneous writing (Fox Tree, 2015). We think the high predictiveness of *oh wait* comes from both the revision-predictiveness of the *oh* (which violates expectations of no revision) and the wait-signaling of the *wait*, although there may be other factors or interactions; the predictiveness of *oh right* as a signal of sarcasm was also high, 81% (Oraby et al., 2016).

We predict that the unexpectedness of written fillers plus fillers' basic meaning of waiting will lead to increased ratings of sarcasm when assessing debate posts that have fillers. Similarly, the unexpectedness of ellipses in asynchronous writing (which allows producers time to plan) plus ellipses' basic meaning of waiting will also increase sarcasm ratings for debate posts with ellipses. Importantly, we do not propose that wait signals in writing only cue sarcasm. We propose that when asked to evaluate sarcasm, the unexpectedness of the wait signal in an asynchronous form of communication coupled with the signal to wait will suggest sarcasm.

2.6 Current Research

We tested the hypothesis that contextually unexpected text patterns are cues for sarcasm, and in particular that wait signals — which prompt taking time in assessing upcoming information — are cues to sarcasm. In a corpus comparison, we tested the rate of filler production across a range of spoken and written corpora. Although others have observed more fillers in speaking than in writing (e.g. Fox Tree, 2015), we wanted to confirm this across a wide range of corpora, as well as explore

the proportions of *um*s to *uh*s across corpora. In Studies 1 and 2, we tested the hypothesis that online posts that included a wait signal, defined as fillers or ellipses, would be rated as more sarcastic than online posts without them. Because a pause in a spoken conversation has no single written equivalent (periods, ellipses, dashes, em dashes, semicolons, and commas all may qualify), it is challenging to identify whether any particular pause is meant to convey sarcastic meaning. However, ellipses (...) specifically suggest "an omission (as of words) or a pause" (Merriam-Webster's Online Dictionary, n.d.) and so may be most likely to be linked to sarcasm when readers are asked about sarcasm.

2.7 Hypotheses

We began by verifying that fillers are contextually unexpected text patterns, comparing across spoken and written American and British corpora:

H1: There are more fillers in speaking than in writing (Study 1).

We then tested whether the presence of wait signals in an unexpected context increased sarcasm ratings. We tested fillers at the beginning of turns:

H2: The presence of a filler at the beginning of written turns will suggest sarcasm at a higher than baseline rate (Study 2).

And in the middle of turns:

H3: The presence of a filler in the middle of written turns will suggest sarcasm at a higher than baseline rate (Study 3).

As well as ellipses, which most often occur in the middle of turns:

H4: The presence of an ellipsis in the middle of written turns will suggest sarcasm at a higher than baseline rate (Study 3).

An alternative to the hypothesis that wait signals suggest sarcasm when readers are asked about sarcasm (H2, H3, H4) is that wait signals are a stylistic device to make written language feel more like spoken talk, without any implication for conveying sarcasm.

In general, we predict that contextually unexpected patterns can be cues to sarcasm, such as fillers in writing or quotes (air-quotes) in speaking. But beyond contextual inappropriateness, we predicted that cues to wait would enhance ratings of sarcasm, as they suggested deeper thought — with deeper thinking possibly leading to alternative interpretations from the literal words expressed. We compared fillers to words we thought might indicate sarcasm:

H5: The words *obviously*, *surely*, *no doubt*, and *clearly* will suggest sarcasm at a higher than baseline rate.

H6: Fillers will be more effective at suggesting sarcasm than the words *obviously*, *surely*, *no doubt*, and *clearly*.

As an alternative to H5, *obviously*, *surely*, *no doubt*, and *clearly* may not suggest sarcasm at higher than baseline rate. As an alternative to H6, fillers may suggest sarcasm less than or to a similar degree as the words *obviously*, *surely*, *no doubt*, and *clearly*. We also compared fillers to a device we thought might indicate sarcasm:

H7: Quotation around a single word will suggest sarcasm at a higher than baseline rate.

H8: Fillers will be more effective at suggesting sarcasm than quotation around a single word.

As an alternative to H6, quotation around a single word may not suggest sarcasm at higher than baseline rate. As an alternative to H7, fillers may suggest sarcasm less than or to a similar degree as quotation around a single word.

2.8 Study 1: Comparing Corpora

In Study 1 we investigated the frequency of the fillers *um* and *uh* across several corpora of both spontaneous communication and planned communication. Working with transcripts of spoken conversation can be challenging because across corpora, transcribers generally do not follow the same transcription rules. In addition, it is often not possible to access the original audio conversation to determine how transcription was done. This is especially problematic when examining word frequencies for discourse markers and fillers, as transcription rules vary especially widely on whether to include words like *so*, *I mean*, and *uh*. Furthermore, frequency of these markers may show large variance across different contexts. For example, if one corpus is made up of unscripted conversations from radio and television shows (e.g. Simpson, Briggs, Ovens, & Swales, 2002), there may be fewer fillers due to television and radio personalities being more likely to have received speech training to avoid using them. Likewise, when performing a difficult communication task over the phone (e.g. Liu, Fox Tree, & Walker, 2016), one may expect the frequency of fillers to be higher on average because people may be more likely to produce delays, and therefore the fillers that indicate delays (Clark & Fox Tree, 2002). For these

reasons, we chose to analyze several different corpora from both spoken and written sources and examine their differences and similarities.

2.8.1 Method

Word frequencies were calculated from several publicly available corpora. We include short explanations of and examples from each corpus to contextualize word frequencies in each.

The Michigan Corpus of Academic Spoken English (MICASE) is a 1.8-million-word corpus that consists of transcripts from colloquia, dissertation defenses, sections, lectures, office hours, seminars, study groups, and similar academic situations. Its close to 200 hours of transcribed audio were recorded at the University of Michigan in Ann Arbor (Simpson, Briggs, Ovens, & Swales, 2002). Fillers in MICASE appear to generally be quite spontaneous. For example, "okay. then that's... that is that's one thing to figure out um but that's probably too much work it's not worth that" (from the Michigan Corpus of Academic Spoken English, LEL565SU064; Simpson, Briggs, Ovens, & Swales, 2002).

The Corpus of Contemporary American English (COCA) is the largest corpus used in this analysis, at more than 520 million words. The spoken component of the corpus contains over 109 million words transcribed from unscripted TV and radio conversations over 26 years. The four written portions of the corpus are each of similar size to the spoken portion and are taken from fictional works, magazines, newspaper articles, and academic journals (Davies, 2008). Unfortunately, because audio is no longer available for the COCA and transcription methods are unknown, it

is difficult to interpret word frequencies for fillers, which frequently are left out of transcription instructions. In the written component, many fillers are within direct quotations, but some exist outside of them, for instance, “Samantha, Samantha, Samantha. What to say about Sa-Man-Tha? Um, okay. This is what I’m going to say about Samantha. Nothing” (from the Corpus of Contemporary American English; Davies, 2008).

The British National Corpus (BNC) is 100 million words divided into spoken (10%) and written (90%) components. The written portion samples newspapers, fictional works, academic books, and other texts, while the spoken portion is entirely made up of informal conversations, “recorded by volunteers selected from different age, region and social classes in a demographically balanced way” (British National Corpus Consortium, 2007). An example taken from the written part of the corpus is, “It’s very nice of you to ask me — erm — but I’ve got a lot to do when I get back to England — erm — I’d like to have a lie down … and there’ll be piles of washing … and I haven’t got a hairdresser …” (from The British National Corpus; British National Corpus Consortium, 2007).

SubtlexUS consists of 50 million words of “spoken-like” language of English-language subtitles from television and film (Brysbaert & New, 2009). Because this corpus generally contains scripted speech, we treat it as written — but we acknowledge that the nature of improvisation and acting may allow for more fillers, as in “Pardon me, please. Yeah. The, uh ... the man-eating wolves are on a, um ... ski vacation” (from SubtlexUS; Brysbaert & New, 2009).

The Internet Argument Corpus consists of about 73 million words of debate posts taken from a popular online debate forum (Walker et al., 2012). It should be noted that this corpus is different from the other written corpora we cite in that it consists of work that has not been published in the traditional sense of the word — that is, all the other written corpora draw from newspapers, magazines, books, and other written works that are likely heavily edited prior to being published. An internet forum, on the other hand, has relatively simple mechanisms for revising a work prior to publishing it. In addition, whereas more traditional written works tend to be monologic, the Internet Argument Corpus consists almost entirely of dialogue. These differences are frequently apparent in the corpus, as in "First you lie about what I said, then you quote me to prove it's a lie. That was, um, helpful of you" (from the Internet Argument Corpus; Walker et al., 2012).

The Artwalk Corpus contains about 500,000 words transcribed from mobile cell-phone conversations that took place while participants collaborated on a naturalistically situated referential communication task that also involved a wayfinding component (Liu, Fox Tree, & Walker, 2016). Although Brysbaert and New (2009) suggest that corpora must be 1-3 million words in order to get reliable estimates of high-frequency words, we also included the Artwalk corpus in our analysis for two reasons: First, we believe it represents an important type of naturalistic conversation that is not represented by the other corpora. Second, Brysbaert & New's operationalization of high frequency was "over 20 words per million." Because there is a difference of several orders of magnitude between this

conceptualization of high-frequency and the frequency of our target words in the Artwalk corpus (over 9,000 words per million), we believe that the additional information from Artwalk is interesting enough to warrant inclusion. An example from the corpus is, “The the computer for the directions it says we have eight minutes to find each um like we’re finding statues and like art pieces um” (from Artwalk; Liu, Fox Tree, & Walker, 2016).

Interpreting raw differences between spoken and written frequencies may be inequitable due to higher lexical diversity in written media. With more words to choose from, the rate of any particular word would be lower. For this reason, we multiplied frequencies originating from written corpora by the constant 2.05, the highest ratio of lexical diversity between spoken and written reported in Johansson (2009). Because we hypothesize that the frequency of our target words should be lower in written communication, this adjustment creates a more conservative estimate.

2.8.2 Results

Table 1 reports raw word frequencies for *uh*, *um*, *er*, and *erm* (British forms of *uh* and *um*) across spoken and written corpora, and written frequencies when corrected for the difference in lexical diversity between the two media.

With the exception of the COCA corpus, the rates of spoken *uh*, *um*, *er*, and *erm* are many times higher in spoken corpora than written corpora. The average rate of *um*s and *uh*s in the spoken MICASE and Artwalk corpora was 9,802 instances per million words compared to 430 instances per million words in the written IAC and

SubtlexUS corpora, adjusted for lexical diversity, a difference of 23 to 1. For COCA, this relationship was 0.44 to 1: there were more written *um*s and *uh*s than spoken. In the spoken corpora, the ratio of *um*s to *uh*s was 1.07 to 1 for MICASE, 1.24 to 1 for Artwalk, 0.71 to 1 for the BNC, and 0.47 to 1 for COCA. That is, in the American conversational spoken corpora, there were more *um*s than *uh*s, and in the British corpus and the American television and radio corpus, there were more *uh*s than *um*s. In the written corpora, the ratio of *um*s to *uh*s was 0.52 to 1 for COCA, 0.94 to 1 for the IAC, 0.12 to 1 for SubtlexUS. The ratio of *erm*s to *er*s was 0.18 to 1 for the BNC and 0.12 to 1 for the IAC. There were no written *erm*s in SubtlexUS. That is, across all written corpora there were more *uh*s and *er*s than *um*s and *erm*s.

Spoken Corpora				
	MICASE	*Artwalk*	*COCA*	*BNC*
uh	9,043.13	9,174.45	13.29	8,542
um	9,644.20	11,377.18	6.26	6,029
Written Corpora				
	IAC	*SubtlexUS*	*COCA*	*BNC*
uh	19.68 (40.71)	717.24 (1,470.33)	14.16 (29.02)	11 (22.56)
um	18.67 (38.28)	86.69 (177.71)	7.36 (15.09)	2 (4.1)

Table 1: Frequency in words per million for filler words in corpora that were either spoken or written. Lexical diversity adjusted written frequencies in parentheses. Note that the BNC and COCA have both written and spoken components. Frequencies are reported for each component. For the BNC, British equivalent fillers (*er* and *erm*) were substituted.

2.8.3 Discussion

The difference between filler use in spoken and written corpora was stark, with far more fillers in spoken corpora. COCA's rates were much lower than the

other corpora we examined. Because we could not ascertain whether *um*s or *uh*s were included in transcription instructions for COCA's spoken corpora, we leave it out of our analysis entirely, merely noting that when we took a closer look at the COCA's instances of *um* and *uh* that occurred in writing, we found the majority of them to be direct quotations. When excluding COCA, the rate of fillers across spoken to written settings was 23 to 1. Extrapolating this data to estimate how likely language users are to encounter fillers across settings suggests that for every filler a person reads in written conversation, a person could be expected to hear, conservatively, 23 fillers in spoken conversations.

In American conversational corpora, there were more *um*s than *uh*s. In British conversational corpora and American television and radio corpora, there were more *er*s/*uh*s than *erm*s/*um*s. In both American and British written corpora there were more *er*s/*uh*s than *erm*s/*um*s.

The strongest outlier in these data was the spoken component of COCA. Our best explanation of this difference is that although COCA's spoken component is made up of unscripted conversations (such as interviews and debates) from television and radio programs, transcribers of these programs may not have concerned themselves with transcribing fillers. Additionally, speakers in television and radio may be more likely to have been trained against the use of fillers in speech. Television personalities may also have spoken quickly, which Clark and Fox Tree (2002) showed is inversely correlated with the frequency of fillers.

The BNC has fewer spoken fillers in comparison to both American English corpora, MICASE and Artwalk. But the BNC also has far fewer written fillers in comparison to the IAC and in SubtlexUS. The BNC displays a more than five-hundred-fold difference in frequencies for *er* and *erm* across spoken and written formats, in comparison to the twenty-three-fold difference in MICASE, Artwalk, IAC, and SubtlexUS for spoken and written *uh* and *um*. One interpretation is that the words *er* and *erm* are just more commonly spoken than written in British English. Additionally, *er* and *erm* may be just far less commonly written in British English. Because American English generally uses *uh* and *um*, *er* and *erm* frequencies are predictably low for corpora featuring American English. Nonetheless, *er* is actually more common in the IAC and SubtlexUS than in the British English corpus.

More convincing than the overall differences between spoken and written contexts may be that the highest rate of fillers, lexical diversity-corrected, for written corpora was in Subtlex, the corpus most clearly meant to emulate spoken dialogue. In summary, fillers are used more frequently in speech than in writing, although they do occur in both contexts. This result supports Hypothesis 1. In most spoken and written corpora investigated here, there were more uhs than ums. The exception was American conversational corpora where there were more ums than uhs. In Study 2, we turn to the test of whether fillers in writing indicate sarcasm.

2.9 Study 2: Wait Signals at the Beginning of Turns

In Study 2, we examined whether posts to online debate forums were more likely to be perceived as sarcastic if they began with a filler. Previous researchers

showed that the probability of Mechanical Turk workers rating a post-response pair from the Internet Argument Corpus as sarcastic was approximately 12% (Walker et. al., 2012). We also examined three other words and a phrase which we thought may also be used to indicate sarcasm: *obviously*, *surely*, *no doubt*, and *clearly*. If these phenomena indicate sarcasm, the probability that Mechanical Turk workers will rate posts as sarcastic should be higher than 12%. We tested the beginning of the turns because that is the likely location for fillers in writing (Fox Tree et al., 2011).

2.9.1 Method

In this section, we discuss the participants, materials, and procedure for Study 2.

2.9.2 Participants

Mechanical Turk workers were required to have an overall approval rate of at least 95%, to have completed at least 500 tasks, and to have an IP address originating from an English-speaking country (including Australia, Canada, New Zealand, Great Britain, and the United States). Workers were paid $0.80 for rating 20 post-response pairs.

2.9.3 Materials

We used regular expressions to collect a set of stimuli from the Internet Argument Corpus. We then performed additional filtering by limiting our set to posts that had parent posts (contained a quote from a previous post) and contained between 10 and 150 words. For example, “Wouldn’t this be contrary to the popular convention that sexuality is innate and orientation is permanent?” is a parent post to, “No, but it

would be contrary to your false premise that sexuality is a dichotomy and that orientation is uh... pardon the expression... rigid." We collected all the post-response pairs in the Internet Argument Corpus which contained one of the following six textual patterns in the response: *um* (at the beginning of the response), *uh* (at the beginning of the response), *obviously*, *surely*, *no doubt*, and *clearly*. The last four of these textual patterns were included as contrasts to *um* and *uh*. The stimuli selected were others that had the potential to indicate sarcasm. All of them could be considered to belong to the category of "adjectives or adverbs used to exaggerate or minimize a statement" (Hancock, 2004, p. 453), which have been shown to be related to judgements of irony, although the set we selected was not specifically mentioned in Hancock (2004). Obviously was noted by several researchers as a marker of sarcasm (Burgers, Van Mulken, & Schellens, 2012; Oraby et al., 2016; Whalen et al., 2009). *Surely* and *clearly* were selected because of their similarity to *obviously*. *No doubt* is also similar to *obviously* and was part of a sarcastic sample in Whalen et al. (2009). *Er* and *erm* were not used, as they were not frequent enough in the corpus to analyze. We randomly selected 166 - 168 posts with each textual pattern from the results to be used as our stimuli.

2.9.4 Procedure

Once the final set of posts were selected, we then created a Human Intelligence Task (HIT) on Amazon Mechanical Turk that asked workers whether any part of the response contains sarcasm. Five workers rated each post-response pair and

posts were marked as sarcastic if the majority of workers (three out of five) agreed that the response contained sarcasm. A total of 233 workers accepted the tasks.

2.9.5 Results

Given that most sarcasm annotation tasks of this type find low reliability on sarcasm ratings, we expected low reliability (e.g., Walker et. al., 2012; Swanson, Lukin, Eisenberg, Corcoran, & Walker, 2017; Davidov, Tsur, & Rappoport 2010). Indeed, many studies avoid this problem by focusing on text that includes the explicit #sarcasm or #irony hashtags, common on Twitter (Peng, Lakis, & Pan, 2015; Liebrecht, Kunneman, & van Den Bosch, 2013; Abercrombie & Hovy, 2016; Riloff et. al., 2013; González-Ibánez, Muresan, & Wacholder, 2011) rather than have humans hand-annotate text. Davidov et al., (2010), when using Fleiss's kappa with two categories (the fewer categories, the higher the κ), achieved a reliability of .34 for Amazon reviews and .41 on Twitter tweets, indicating fair reliability at best. Even when including relatively clear cases of sarcasm, Swanson et. al., (2017) found an alpha of only .387. They argue that though this is usually considered low, the subjectivity of sarcasm may mean that it should be treated differently. Several researchers argue that these low agreements are a result of the fact that there is wide variation in how people use and understand sarcasm (Walker et. al., 2012; Swanson et al., 2017; Davidov, Tsur, & Rappoport, 2010). Low inter-rater reliability on manual ratings of sarcasm seems to be an unfortunate corollary of studying forms of sarcasm that don't contain explicit textual flagging.

Sure enough, our Krippendorff's alpha for the workers' ratings of sarcasm was $\alpha = .17$. As a part of the process of preparing the rating task, several researchers and assistants tested our HITs. Not one of our researchers or assistants were able to complete the task in fewer than five minutes. However, 39 of our 250 tasks were completed in under five minutes, 17 in under three minutes, and one in 14 seconds (as reported by Mechanical Turk). On the opposite end, 44 workers were reported as spending over 30 minutes on the task. Although we cannot be certain about the large discrepancy in times, a plausible explanation is that the short duration workers skimmed or ignored the post-response pairs, and the long duration workers took breaks while working on the task. Another possibility is that short duration workers considered their answers prior to accepting the task, leaving them with the trivial task of filling them in once they accepted the task, and the work time counter began. It could also be that some participants put little effort into the nontrivial cognitive task of sarcasm comprehension.

Although inter-rater reliability was practically nonexistent for our participants, there were still reliable differences between stimuli that contained our cues and those that did not. Comparing the rate of sarcasm across conditions is still valuable in spite of the high variability in participants' rating behavior. We ran chi-squared tests of independence to determine if the rates of sarcasm in our post-response pairs were significantly different from the baseline rate of 12% that Walker et al. (2012) found using stimuli from the same corpus and an identical HIT procedure on Mechanical Turk. See Table 2. Post-response pairs starting with the word *uh* at the beginning of

the post had higher rates of sarcasm, $\chi 2(1) = 23.1$, $p < .001$, $\Phi = .08$, and we can be 95% confident that between 18.1% and 31.3% of IAC post-response pairs that start with the word *uh* would be rated as sarcastic by a majority of mTurk workers. Post-response pairs starting with the word um also had higher rates of sarcasm, $\chi 2(1) = 43.2$ $p < .001$, $\Phi = .11$ and we can be 95% confident that between 22.6% and 36.5% of IAC post-response pairs that start with the word *um* would be rated as sarcastic by a majority of mTurk workers. In addition, post-response pairs including the word *obviously* had higher rates of sarcasm than baseline, $\chi 2(1) = 11.7$, $p = .001$, $\Phi = .06$ and we can be 95% confident that between 14.8% and 27.1% of IAC post-response pairs that include the word *obviously* would be rated as sarcastic by a majority of mTurk workers, post-response pairs including the word *surely* had higher rates of sarcasm than baseline, $\chi 2$ (1) = 18.6 $p < .001$, $\Phi = .08$ and we can be 95% confident that between 16.9% and 29.8% of IAC post-response pairs that include the word *surely* would be rated as sarcastic by a majority of mTurk workers, and post-response pairs including the word *clearly* had higher rates of sarcasm than baseline, $\chi 2$ (1) = 6.2, $p < .013$, $\Phi = .04$ and we can be 95% confident that between 12.6% and 24.3% of IAC post-response pairs that include the word *clearly* would be rated as sarcastic by a majority of mTurk workers. Post-response pairs including the phrase *no doubt* did not have higher rates of sarcasm than baseline, $\chi 2$ (1) = 1.4, $p = .239$, $\Phi = .02$ and we can be 95% confident that between 9.6% and 20.5% of IAC post-response pairs that include the phrase no doubt would be rated as sarcastic by a majority of mTurk workers.

Our two best candidates, *um* and *uh*, each displayed rates of sarcasm more than double the baseline rate in the corpus.

Study 2: Comparison of Sarcasm Rates to 12% Baseline						
cue	*Regex*	*Annotated*	χ^2	p	ϕ	95% *CI*
uh (beginning)	^uh+\b	166	23.1	<.001	.08	18.1%, 31.3%
um (beginning)	^um+\b	166	43.2	<.001	.11	22.6%, 36.5%
obviously	\bobviously\b	167	11.7	<.001	.06	14.8%, 27.1%
surely	\bsurely\b	167	18.6	<.001	.08	16.9%, 29.8%
clearly	\bclearly\b	168	6.2	.013	.04	12.6%, 24.3%
no doubt	\bno doubt\b	166	1.4	.239	.02	9.6%, 20.5%

Table 2: Textual cues of interest, the regular expression used to isolate post-response pairs with that pattern, the total number of post-response pairs annotated, and results of the χ^2 analysis comparing the number of sarcastic ratings to the baseline frequency of sarcastic ratings using this procedure.

2.9.6 Discussion

Writing *um* or *uh* at the beginning of a turn suggested to readers that the writers were being sarcastic at more than twice the base rate of sarcasm for the Internet Argument Corpus. This result supports Hypothesis 2. *Um*s and *uh*s were more predictive than a number of other words tested, including words others identified as related to sarcasm. This result supports Hypothesis 6. Of the conventional words hypothesized to be related to sarcasm — *obviously*, *surely*, *no doubt*, and *clearly* — only *no doubt* did not have a higher rate of sarcasm ratings than baseline. This result partially supports Hypothesis 5.

One possibility is that only wait signals at the start of turns will affect sarcasm perception. The start of a turn is more noticeable, and indeed others have tested the role of written discourse markers in turn initial position precisely because of the salience of this location (Abbott et al., 2011). In Study 3, we assessed whether wait signals in the middle of turns also influenced sarcasm perception.

2.10 Study 3: Wait Signals in the Middle of Turns

In Study 3, we examined whether posts to online debate forums were more likely to be perceived as sarcastic if they contained a filler or an ellipsis that was not at the beginning of a turn (referred to henceforth as *uh* (within) and *um* (within)). We also examined quotation marks encapsulating single words, which we thought may indicate higher sarcasm as a textual equivalent of the air-quotes gesture (Lampert, 2013). Once again, if fillers, ellipses, or quotes around a word indicate sarcasm, Mechanical Turk workers should rate posts including them as sarcastic at a rate higher than 12%.

2.10.1 Method

Methods for Study 3 were identical to Study 2 with two exceptions: First, we used different textual patterns to collect post-response pairs from the Internet Argument Corpus, and second, we recruited a smaller set of workers who already had experience rating sarcastic content in online debate posts, in an attempt to achieve higher inter-annotator agreement.

2.10.2 Participants

Mechanical Turk workers were recruited from a pool of workers who had previously been ranked as providing reliable ratings of sarcasm in textual stimuli according to the conditions specified in Oraby et. al., (2016). All workers were also required to have an overall approval rate of at least 95%, to have completed at least 500 tasks, and to have an IP address originating from an English-speaking country

(including Australia, Canada, New Zealand, Great Britain, and the United States). Workers were paid $0.80 for rating 20 post-response pairs.

2.10.3 Materials

As in Study 2, we used regular expressions to match specific textual patterns within the Internet Argument Corpus, using the same constraints as before, selecting only posts that had between 10 and 150 words, and included a quote from a previous post.

We randomly selected sets of approximately 200 posts per pattern. Due to possible limitations of our scripts combined with relative scarcity of cues within the corpus, only 159 *uh* (within) posts were identified. Further, some posts were manually removed because upon inspection the posts fell into categories we did not want to examine and also believed we could computationally control for in future studies. The categories that made a post-response pair eligible for exemption from our set of stimuli were: (1) The post-response pair was a duplicate post-response pair to one that already existed in the set (1 removed), (2) The post-response pair included the matched pattern as part of a URL (19 removed), (3) The response did not include fillers or ellipses (15 removed) and (4) The post-response pair was not written in English (1 removed). This process afforded us a set of 154 *uh* (within) posts, 182 *um* (within) posts, 184 ellipses posts, and 292 quoted word posts (posts with quoted words were relatively plentiful within the IAC, and so were used to fill our quota for the HIT). Quotations were included as contrasts to *um* and *uh*. Quotations have been

argued to express sarcasm both in speaking, as air-quotes (Lampert, 2013), and in writing (Carvalho et al., 2009).

2.10.4 Procedure

Once the final set of posts were selected, we then created a HIT (Human Intelligence Task) on Amazon Mechanical Turk that asked workers whether any part of the response contains sarcasm. Five workers rated each post-response pair, and posts were marked as sarcastic if the majority of workers (three out of five) agreed that the response contained sarcasm. A total of nine workers accepted the tasks.

2.10.5 Results

As in Study 2, we expect a low worker reliability due to the challenges presented in Section 3.2. For Study 3, the Krippendorff's alpha for the workers' ratings of sarcasm was $\alpha = .32$, which was higher than the alpha of .17 in Study 2, but still far under common thresholds for fair reliability (Krippendorff, 2004). We attribute this to the higher quality of our workers. The Krippendorff's alpha was also higher than the alpha of .22 for the original sample of 3,158 post-response pairs. We attribute this boost in reliability to the fact that our set of posts contained strong predictors of sarcasm (*um*, *uh*, or ellipses), so the sarcasm should be less ambiguous, leading people to agree on it in more cases. This explanation fits with the higher alpha (.39) achieved in another study in which researchers included unambiguous sarcastic/non-sarcastic post-response pairs (Swanson, Lukin, Eisenberg, Corcoran, & Walker, 2017).

Despite the low reliability between workers, comparing sarcasm ratings across conditions is still valuable. While reliability detects the agreement of workers, the following analyses detect differences between overall proportion of post-response pairs rated as sarcastic. We ran chi squared tests of independence to determine if the rates of sarcasm in our post-response pairs were significantly different from the baseline rate of 12% that Walker et al. (2012) found using stimuli from the same corpus and an identical HIT procedure. See Table 3. Post-response pairs including the word *uh* had higher rates of sarcasm, $\chi 2(1) = 363.7$, $p < .001$, $\Phi = .33$, and we can be 95% confident that between 60.1% and 74.9% of IAC post-response pairs that include the word *uh* would be rated as sarcastic by a majority of mTurk workers. Post-response pairs including the word *um* also had higher rates of sarcasm, $\chi 2\ (1) = 309.8$, $p < .001$, $\Phi = .30$ and we can be 95% confident that between 52.2% and 66.4% of IAC post-response pairs that include the word *um* would be rated as sarcastic by a majority of mTurk workers. And post-response pairs including ellipses had higher rates of sarcasm, $\chi 2\ (1) = 122.6$, $p < .001$, $\Phi = .19$, and we can be 95% confident that between 33.6% and 47.9% of the IAC post-response pairs that include ellipses would be rated as sarcastic by a majority of mTurk workers. In addition, post-response pairs including quotations had higher rates of sarcasm, $\chi 2\ (1) = 195.2$, $p < .001$, $\Phi = .24$, and we can be 95% confident that between 36.5% and 47.8% of the IAC post-response pairs that include quotations would be rated as sarcastic by a majority of mTurk workers.

Study 3: Comparison of Sarcasm Rates to 12% Baseline						
cue	*Regex*	*Annotated*	χ^2	p	ϕ	95% *CI*
uh (within)	[^\w]uh[^\w]	154	363.7	<.001	.33	60.1%, 74.9%
um (within)	[^\w]um[^\w]	182	309.8	<.001	.30	52.2%, 66.4%
ellipses	[a-zA-Z]\.\.\.	184	122.6	<.001	.19	33.6%, 47.9%
quoted word	\"(?:[A-Za-z]{3,})\"	292	195.2	<.001	.24	36.5%, 47.8%

Table 3: Textual cues of interest, the regular expression used to isolate post-response pairs with that pattern, the total number of post-response pairs annotated, and results of the χ^2 analysis comparing the number of sarcastic ratings to the baseline frequency of sarcastic ratings using this procedure.

2.10.6 Discussion

Writing *um*, *uh*, or using ellipses in the middle of a turn suggested to readers that the writers were being sarcastic at 4.5 times the base rate of sarcasm for the Internet Argument Corpus. These results support Hypotheses 3 and 4. The lowest rate found, for ellipses, was still over triple the baseline rate of sarcasm in the corpus. This result supports Hypotheses 7 and 8.

As observed in prior work (Carvalho et al., 2009; Lampert, 2013), quotations around single words were also indicative of sarcasm, at over triple the baseline rate. In speech, people reported that they used direct quotation (which would be expressed with quotation marks if written) to be entertaining (Blackwell & Fox Tree, 2012). Direct quotes were also used to report thoughts (Fox Tree & Tomlinson, 2008), and were often accompanied by vocal and bodily demonstrations, such as moving the mouth and neck up as if howling and using a howling voice to imitate a dog's behavior (Blackwell, Perlman, & Fox Tree, 2015). Being entertaining, reporting thoughts, and adding vocal and bodily information might all contribute to a relationship between spoken quotation and sarcasm. This relationship may be alluded to with written quotations. Written quotations may also act as text-separators to

highlight non-literal content (Whalen et al., 2009, p. 275). As text-separators they could potentially contribute to the pacing of information consumption which in turn may be suggestive of sarcasm, as proposed for um, uh, and ellipses.

2.11 General Discussion

Sarcasm has been studied across speech and writing and in synchronous and asynchronous settings. In the current series of studies, we documented the prevalence of fillers across spoken and written corpora and tested how likely fillers were to suggest sarcasm when they fell at the beginning of turns and in the middle of turns. We predicted that fillers would be more frequent in speaking than in writing (H1). We also predicted that fillers would suggest sarcasm because they are uncommon in writing (H2, H3), and that fillers and ellipses would suggest sarcasm because they communicate the need to wait in a context where waiting isn't necessary (H4). We thought that seeing elements typical of spoken speech in writing (fillers and a written representation of spoken pauses, ellipses) would suggest to readers a need to think more deeply about what the writer was communicating.

We also predicted that *obviously*, *surely*, *clearly*, and *no doubt* would indicate the presence of sarcasm (H5), although at lower rates than fillers and ellipses (H6), and that quotation around a single word would indicate sarcasm (H7), also at lower rates than fillers and ellipses (H8), because fillers and ellipses indicate delay, further prompting readers to consider the material they were reading more deeply. We predicted that the search for deeper meaning would lead listeners to consider writers' non-literal goals in using fillers and ellipses, such as the production of sarcasm.

Across corpora, we demonstrated that fillers are more common in speech than in writing. We also documented differences in preferences for *er*/*uh* versus *erm*/*um* across corpora and settings, with more *um*s in conversational American English corpora, and more *er*s/*uh*s in a conversational British corpus, a television/radio American corpus, and all written corpora. In two studies, we showed that fillers and ellipses reliably indicated sarcasm to readers and to a greater extent than other sarcasm-predicting devices.

These data are indicative of a broader pattern in which writers use incongruent language to express sarcasm (Clark & Gerrig, 1984; Kovaz et al., 2013; Rubin et al., 2016). Another way to view incongruence is by noting language that is used more frequently in one medium than another. Because fillers and pauses are not necessary in asynchronous written communication, such as online forums, the use of fillers and pauses are contextually inappropriate — their use contrasts with their medium. We suggest that this contrast is what enables *um*, *uh*, ellipses, and likely other phenomena, to cue non-literal meaning, including sarcasm.

One next step with this research is to examine whether these patterns exist for more types of computer-mediated communication, including testing varying levels of synchronicity. More synchronous communicative methods, like instant messages and text-messages, could be expected to have lower rates of sarcasm co-occurrence with fillers, because fillers would be more likely to be used in these media for their time-noting functions; for example, communicators using text chat might write *um* to indicate that their response will be delayed (although text chat programs that contain

blinking ellipses to indicate that the respondent is writing may obviate the need for an *um*). More asynchronous communicative methods, like Reddit and other message boards, could be expected to have higher rates of sarcasm co-occurrence with fillers, much like we observed here with an online debate forum. Other phenomena that might be explored include other wait signals, such as words like *wait* or *hang on*, characters like em-dashes, typographic behavior such as spacing out words, like *t h i s*, or elongations like *thiiiis*. Like fillers, elongations have been shown to indicate upcoming problems in speech (Fox Tree & Clark, 1997). Their interpretation in writing may be similar to fillers as well.

Another next step with this research is to assess the role of wait signals on other kinds of inferences readers can make beyond sarcasm. For example, wait signals may influence assessments of politeness or evasion. Hearing *um*s at the beginning of the spoken turns affected listeners' judgements of speakers' production difficulty, comfort, and honesty (Fox Tree, 2002). But this wasn't because the *um*s were a leaked symptom of difficulty, discomfort, or dishonesty. The assessments were a product of the *um*s' basic meaning — announcing an upcoming delay — coupled with the requirements of the task. Listeners were, in essence, asking themselves why a speaker would need to delay right then, and, if thinking about honesty, conclude that the speaker needed time to come up with a deceptive answer.

Finally, it would be interesting to determine whether there is a difference in how sarcastically *um*s are viewed as opposed to *uh*s. In spoken communication *um*s lead to longer pauses than *uh*s on average (Clark & Fox Tree, 2002). Since wait

signals in online forums seem to be able to cue sarcasm through their inappropriateness, a longer pause could be viewed as more inappropriate than a short one. It is possible, therefore, that the longer pauses implied by *um* lead to higher ratings of sarcasm than *uh*. Although our data trends toward *um*s at the beginnings of posts being rated as sarcastic more frequently than *uh*s, it trends in the opposite direction for *um*s and *uh*s in the middle of posts. It's also important to note that frequency does not necessarily imply intensity, so it would be interesting to use a more nuanced rating of sarcasm to check for differences between wait signals.

As we achieve a better understanding of mechanisms and cues of non-literal language, both in writing and in speech, we will be able to train computers to flag sarcasm in language more and more accurately, leading to better tools to assist those who could benefit from them. One group who could benefit are people with hearing difficulties. Deaf children show slower development in recognizing sarcasm than hearing children, and although native sign language signers' performance appears to eventually catch up to hearing persons', late signers (those from hearing families) continue to show reduced performance in sarcasm recognition into adulthood (O'Reilly & Peterson 2014). Another group who could benefit are people on the autism spectrum, who struggle with sarcasm identification (Kaland, Møller-Nielsen, Callesen, Mortensen, Gottlieb, & Smith, 2002; Peterson 2012). A third group who could benefit are second language learners, who also struggle with sarcasm identification, such as identifying satirical news (Prichard & Rucynksi, 2019). And there are others who could benefit, such as anyone who has trouble differentiating

satirical news reports from real stories, or who has trouble recognizing the satire behind a deadpan delivery. Technology with the ability to recognize sarcastic intent could inform readers of non-literal meaning as they read, bridging gaps in communication.

2.12 Acknowledgements

This work was supported in part by NSF IIS-1302668, and by the Federico and Rena Perlino Award for research related to deafness. We thank Marilyn Walker for her support of this project. We thank Brian Schwarzmann for assistance with Amazon Mechanical Turk.

2.13 References

Abbott, R., Walker, M., Anand, P., Fox Tree, J. E., Bowmani, R., & King, J. (2011, June). How can you say such things?!?: Recognizing disagreement in informal political argument. In *Proceedings of the Workshop on Languages in Social Media* (pp. 2-11). Association for Computational Linguistics.

Abercrombie, G., & Hovy, D. (2016). Putting sarcasm detection into context: The effects of class imbalance and manual labelling on supervised machine classification of twitter conversations. In *Proceedings of the ACL 2016 Student Research Workshop* (pp. 107-113).

Akimoto, Y., & Miyazawa, S. (2017). Individual Differences in Irony Use Depend on Context. *Journal of Language and Social Psychology*, 36(6), 675-693. https://doi.org/10.1177/0261927X17706937

Attardo, S., Eisterhold, J., Hay, J., & Poggi, I. (2003). Multimodal markers of irony and sarcasm. *Humor*, 16(2), 243-260. https://doi.org/10.1515/humr.2003.012

Blackwell, N. & Fox Tree, J. E. (2012). Social factors affect quotative choice. *Journal of Pragmatics*, 44, 1150-1162. https://doi.org/10.1016/j.pragma.2012.05.001

Blackwell, N. L., Perlman, M., & Fox Tree, J. E. (2015). Quotation as multimodal construction. *Journal of Pragmatics*, 81, 1-7. https://doi.org/10.1016/j.pragma.2015.03.004

Blakemore, D. 2002. Relevance and Linguistic Meaning: The Semantics and Pragmatics of Discourse Markers. Cambridge: Cambridge University Press.

British National Corpus Consortium. (2007). British National Corpus version 3 (BNC XML edition). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Retrieved February 13, 2012.

Bryant, G. & Fox Tree, J. E. (2002). Recognizing verbal irony in spontaneous speech. *Metaphor and Symbol*, 17(2), 99-117. https://doi.org/10.1207/S15327868MS1702_2

Bryant, G. A., & Fox Tree, J. E. (2005). Is there an ironic tone of voice? *Language and Speech*, 48(3), 257-277. https://doi.org/10.1177/00238309050480030101

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), 977-990. https://doi.org/10.3758/BRM.41.4.977

Burfoot, C., & Baldwin, T. (2009, August). Automatic satire detection: Are you having a laugh?. In *Proceedings of the ACL-IJCNLP 2009 Conference* (pp. 161-164). Association for Computational Linguistics.

Burgers, C., Van Mulken, M., & Schellens, P. J. (2011). Finding irony: An introduction of the verbal irony procedure (VIP). *Metaphor and Symbol*, 26(3), 186-205. https://doi.org/10.1080/10926488.2011.583194

Burgers, C., van Mulken, M., & Schellens, P. J. (2012). Type of evaluation and marking of irony: The role of perceived complexity and comprehension. *Journal of Pragmatics*, 44(3), 231-242.

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

Carberry, S. (1989). A pragmatics-based approach to ellipsis resolution. *Computational Linguistics*, 15(2), 75-96.

Carvalho, P., Sarmento, L., Silva, M. J., & De Oliveira, E. (2009, November). Clues for detecting irony in user-generated contents: oh...!! it's "so easy" ;-). In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion* (pp. 53-56). ACM.

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

Clark, H. H. (1996). Using language. 1996. Cambridge University Press: Cambridge.

Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, 84(1), 73-111. https://doi.org/10.1016/S0010-0277(02)00017-3

Clark, H. H., & Gerrig, R. J. (1984). On the pretense theory of irony. http://dx.doi.org/10.1037/0096-3445.113.1.121

Davidov, D., Tsur, O., & Rappoport, A. (2010, July). Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning* (pp. 107-116). Association for Computational Linguistics.

Davies, M. (2008-). The Corpus of Contemporary American English (COCA): 520 million words, 1990-present. Available online at http://corpus.byu.edu/coca/.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14(2), 159-190. http://dx.doi.org/10.1075/ijcl.14.2.02dav

Derks, D., Bos, A. E., & Von Grumbkow, J. (2008). Emoticons and online message interpretation. *Social Science Computer Review*, 26(3), 379-388. https://doi.org/10.1177/0894439307311611

Dress, M. L., Kreuz, R. J., Link, K. E., & Caucci, G. M. (2008). Regional variation in the use of sarcasm. *Journal of Language and Social Psychology*, 27(1), 71-85. https://doi.org/10.1177/0261927X07309512

Ellipsis, (n.d.). In Merriam-Webster's online dictionary. Retrieved from https://www.merriam-webster.com/dictionary/ellipsis

Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., & Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *ArXiv Preprint* ArXiv:1708.00524.

Fox Tree, J. E. (2001). Listeners' uses of um and uh in speech comprehension. Memory and Cognition, 29(2), 320-326.

Fox Tree, J. E. (2002). Interpreting pauses and ums at turn exchanges. *Discourse Processes*, 34(1), 37-55. https://doi.org/10.1207/S15326950DP3401_2

Fox Tree, J. E. (2007). Folk notions of um and uh, you know, and like. *Text & Talk*, 27(3), 297-314. https://doi.org/10.1515/TEXT.2007.012

Fox Tree, J. E. (2010) Discourse markers across speakers and settings. *Language and Linguistics Compass*, 3(1), 1–13. https://doi.org/10.1111/j.1749-818X.2010.00195.x

Fox Tree, J. E. (2015). Discourse markers in writing. *Discourse Studies*, 17(1), 64-82.

Fox Tree, J. E., & Clark, H. H. (1997). Pronouncing "the" as "thee" to signal problems in speaking. *Cognition*, 62(2), 151-167. https://doi.org/10.1016/S0010-0277(96)00781-0

Fox Tree, J. E., Mayer, S. A., & Betts, T. E. (2011). Grounding in instant messaging. *Journal of Educational Computing Research*, 45(4), 455-475. https://doi.org/10.2190/EC.45.4.e

Fox Tree, J. E., & Schrock, J. C. (1999). Discourse markers in spontaneous speech: Oh what a difference an oh makes. *Journal of Memory and Language*, 40(2), 280-295. https://doi.org/10.1006/jmla.1998.2613

Fox Tree, J. E. & Tomlinson, J. M., Jr. (2008). The rise of like in spontaneous quotations. *Discourse Processes*, 45, 85-102. https://doi.org/10.1080/01638530701739280

Ghosh, A., & Veale, T. (2016). Fracking sarcasm using neural network. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 161-169).

Gibbs, R. W. (2000). Irony in talk among friends. *Metaphor and symbol*, 15(1-2), 5-27. https://doi.org/10.1080/10926488.2000.9678862

González-Ibánez, R., Muresan, S., & Wacholder, N. (2011, June). Identifying sarcasm in Twitter: a closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2* (pp. 581-586). Association for Computational Linguistics.

Ivanko, S. L., Pexman, P. M., & Olineck, K. M. (2004). How sarcastic are you? Individual differences and verbal irony. *Journal of Language and Social Psychology*, 23(3), 244-271. https://doi.org/10.1177/0261927X04266809

Hancock, J. T. (2004). Verbal irony use in face-to-face and computer-mediated conversations. *Journal of Language and Social Psychology*, 23(4), 447-463. https://doi.org/10.1177/0261927X04269587

Heritage, J., & Atkinson, J. M. (1984). Structures of social action. *Studies in Conversation Analysis*.

Holtgraves, T. (2000). Preference organization and reply comprehension. *Discourse Processes*, 30(2), 87–106. https://doi.org/10.1207/S15326950DP3002_01

Johansson, V. (2009). Lexical diversity and lexical density in speech and writing: A developmental perspective. *Working Papers in Linguistics*, 53, 61-79.

Jorgensen, J. (1996). The functions of sarcastic irony in speech. *Journal of pragmatics*, 26(5), 613-634. https://doi.org/10.1016/0378-2166(95)00067-4

Jucker, A. H. (1993). The discourse marker 'well': a relevance theoretical account. *Journal of Pragmatics*, 19(5), 435–52. https://doi.org/10.1016/0378-2166(93)90004-9

Kaland, N., Møller-Nielsen, A., Callesen, K., Mortensen, E. L., Gottlieb, D., & Smith, L. (2002). A new 'advanced' test of theory of mind: evidence from children and adolescents with Asperger syndrome. *Journal of Child Psychology and Psychiatry*, 43(4), 517-528. https://doi.org/10.1111/1469-7610.00042

Koehrsen, W. Beyond Accuracy: Precision and Recall - Choosing the right metrics for classification tasks [Web log message]. Retrieved from

https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c

Kovaz, D., Kreuz, R. J., & Riordan, M. A. (2013). Distinguishing sarcasm from literal language: Evidence from books and blogging. *Discourse Processes*, 50(8), 598-615.

Kreuz, R. J., Long, D. L., & Church, M. B. (1991). On being ironic: Pragmatic and mnemonic implications. *Metaphor and symbol*, 6(3), 149-162. https://doi.org/10.1080/0163853X.2013.849525

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3), 411-433.

LaMarre, H. L., Landreville, K. D., & Beam, M. A. (2009). The irony of satire: Political ideology and the motivation to see what you want to see in The Colbert Report. *The International Journal of Press/Politics*, 14(2), 212-231. https://doi.org/10.1177/1940161208330904

Lampert, M. (2013). Say, be like, quote (unquote), and the air-quotes: interactive quotatives and their multimodal implications. *English Today*, 29(04), 45-56. https://doi.org/10.1017/S026607841300045X

Lee, D. (2006). Humor in spoken academic discourse. *NUCB Journal of Language Culture and Communication*, 8(1), 49-68.

Leech, G., & Rayson, P. (2014). Word frequencies in written and spoken English: Based on the British National Corpus. Routledge.

Liebrecht, C. C., Kunneman, F. A., & van Den Bosch, A. P. J. (2013). The perfect solution for detecting sarcasm in tweets# not. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 29-37. http://hdl.handle.net/2066/112949

Liu, K., Fox Tree, J. E., & Walker, L. (2016). Coordinating communication in the wild: The Artwalk dialogue corpus of pedestrian navigation and mobile referential communication. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (pp. 3159-3166).

Masten, A. S. (1986). Humor and competence in school-aged children. *Child development*, 461-473. http://dx.doi.org/10.2307/1130601

Nunberg, G. (2001). The Way We Talk Now: Commentaries on Language and Culture from NPR's" Fresh Air". Houghton Mifflin Harcourt.

O'Reilly, K., Peterson, C. C., & Wellman, H. M. (2014). Sarcasm and advanced theory of mind understanding in children and adults with prelingual deafness. *Developmental Psychology*, 50(7), 1862. https://doi.org/10.1037/a0036654

Oraby, S., Harrison, V., Reed, L., Hernandez, E., Riloff, E., & Walker, M. (2016, September). Creating and Characterizing a Diverse Corpus of Sarcasm in Dialogue. In *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (p. 31).

Overholser, J. C. (1992). Sense of humor when coping with life stress. *Personality and Individual Differences*, 13(7), 799-804. https://doi.org/10.1016/0191-8869(92)90053-R

Peng, C., Lakis, M., & Pan, J.W. (2015). Detecting Sarcasm in Text: An Obvious Solution to a Trivial Problem. *Foundations and trends in information retrieval*.

Peterson, C. C., Wellman, H. M., & Slaughter, V. (2012). The mind behind the message: Advancing theory-of-mind scales for typically developing children, and those with deafness, autism, or Asperger syndrome. *Child Development*, 83(2), 469-485. https://doi.org/10.1111/j.1467-8624.2011.01728.x

Poria, S., Cambria, E., Hazarika, D., & Vij, P. (2016). A deeper look into sarcastic tweets using deep convolutional neural networks. *arXiv preprint* arXiv:1610.08815.

Prichard, C., & Rucynski Jr, J. (2019). Second language learners' ability to detect satirical news and the effect of humor competency training. *TESOL Journal*, 10(1), e00366.

Reyes Pérez, A.; Rosso, P. (2014). On the difficulty of automatically detecting irony: beyond a simple case of negation. *Knowledge and Information Systems*. 40(3):595-614. doi:10.1007/s10115-013-0652-8.

Reyes, A., Rosso, P., & Veale, T. (2013). A multidimensional approach for detecting irony in Twitter. *Language Resources and Evaluation*, 47(1), 239–268. https://doi.org/10.1007/s10579-012-9196-x

Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., & Huang, R. (2013). Sarcasm as contrast between a positive sentiment and negative situation. In *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language*

Processing, Proceedings of the Conference (pp. 704-714). Association for Computational Linguistics (ACL).

Rockwell, P., & Theriot, E. M. (2001). Culture, gender, and gender mix in encoders of sarcasm: A self-assessment analysis. *Communication Research Reports*, 18(1), 44-52. https://doi.org/10.1080/08824090109384781

Simpson, R. C., Briggs, S. L., Ovens, J., & Swales, J. M. (2002). The Michigan Corpus of Academic Spoken English. Ann Arbor, MI: The Regents of the University of Michigan.

Smith, V. L., & Clark, H.H. (1993). On the course of answering questions. *Journal of Memory & Language*, 32, 25-38.

Swanson, R., Lukin, S. M., Eisenberg, L., Corcoran, T., & Walker, M. A. (2017). Getting Reliable Annotations for Sarcasm in Online Dialogues. In *LREC* (pp. 4250-4257). https://arxiv.org/abs/1709.01042

Walker, M. A., Fox Tree, J. E., Anand, P., Abbott, R., & King, J. (2012). A Corpus for Research on Deliberation and Debate. In *LREC* (pp. 812-817).

Whalen, J. M., Pexman, P. M., & Gill, A. J. (2009). "Should Be Fun—Not!": Incidence and Marking of Nonliteral Language in E-Mail. *Journal of Language and Social Psychology*. https://doi.org/10.1177/0261927X09335253

Williams, J. A., Burns, E. L., & Harmon, E. A. (2009). Insincere utterances and gaze: eye contact during sarcastic statements. *Perceptual and Motor Skills*, 108(2), 565-572.

Wit. (2019). In OxfordDictionaries.com. Retrieved from https://en.oxforddictionaries.com/definition/wit

3 Chapter 2: The Sarchasm: Sarcasm Production and Identification in Spontaneous Conversation

3.0 Pre-introduction

In this chapter, I present results from a published paper, The Sarchasm: Sarcasm Production and Identification in Spontaneous Conversation (Fox Tree et al., 2020), in which we present a novel method for eliciting large amounts of sarcasm in the laboratory, and then provide evidence that correctly recognizing sarcasm in others may be far less accurate than previously believed. Although many studies have attempted to achieve high interrater agreement on sarcasm, this study is the first to examine interrater agreement directly between the sarcasm producer (whose reports of sarcasm are taken as veridical) and their interlocutor (who must recognize the speaker's intent).

3.1 Abstract

We tested sarcasm production and identification across original communicators in a spontaneously produced conversational setting, including testing the role of synchronous movement on sarcasm production and identification. Before communicating, stranger dyads participated in either a synchronous or non-synchronous movement task. They then completed a task designed to elicit sarcasm, although no instruction to produce sarcastic content was provided. After communicating, participants immediately reviewed their conversations and identified their own and their addressees' sarcastic utterances. No definition of sarcasm was provided. We found that participants who had moved synchronously identified more

sarcasm in their own productions. They did not identify more sarcasm in their partner's productions, however. We also discovered that most identifications of sarcasm did not align across conversational participants, and neither did those of outside observers. People reported sarcasm in their addressees commensurate with the sarcasm they produced, rather than the sarcasm that their addressees self-reported. There were numerous cases of sarchasm, where producers' intended sarcasm was not identified by addressees.

3.2 Introduction

In 2019, a marketing firm asked people what phrases like, "I'll bear that in mind," mean, finding that people from the United States and the United Kingdom understood the phrases differently (Smith, 2019): Americans leaned towards, "I will probably do it," and Britons leaned towards, "I've forgotten it already." The study was reported by BBC news with the title, "YouGov survey: British sarcasm 'lost on Americans'" (BBC, 2019). The experience of misunderstanding sarcasm is common enough to have prompted the neologism *sarchasm*, "the gulf between the author of sarcastic wit and the recipient who doesn't get it" (Witte, 1998). Unlike successful sarcasm, which is understood by speakers and addressees, sarchasm reveals communicative failure. The failure can be with stock phrases like, "I'll bear that in mind," but it can also be with on-the-spot creations, such as saying, "I'll put that down in my notebook of things to remember," while meaning that the information will not be recorded (Bryant & Fox Tree, 2002, p. 114). In this report, we examine

how often sarcasm and sarchasm occur in spontaneous dialogue, and whether feelings of closeness, as created by synchronous movement, can invite increased sarcasm use.

3.2.1 What is Sarcasm?

Many researchers have proposed different ways to define sarcasm and closely related phenomena such as verbal irony (e.g., Kreuz & Glucksberg, 1989; Rockwell, 2003; Utsumi, 2000). But distinguishing sarcasm from other nonliteral language use, such as verbal irony, may be impractical. Many people cannot reliably distinguish them (Bryant & Fox Tree, 2002) and several researchers have collapsed the categories (Attardo, Eisterhold, Hay, & Poggi, 2003; Kruger, Epley, Parker, & Ng, 2005). For the purposes of this paper, we will refer to verbal irony as sarcasm, except where citing authors who refer specifically to irony. Sarcasm (or sarcastic verbal irony) is frequently understood as a type of figurative language that conveys a negative meaning that stems from a clearly incorrect literal interpretation (Kreuz & Glucksberg, 1989).

3.2.2 Cues to Sarcasm

Cues to sarcastic intent can be linguistic, behavioral, or social-contextual. An important linguistic cue to sarcasm is non-veridicality, or asserting a state of affairs that contradicts the actual state of affairs (Kreuz & Glucksberg, 1989). Another linguistic cue is hyperbole (Kreuz & Roberts, 1994). Particular words can also be cues to sarcasm, such as the phrase *yeah right* in conversations (Tepperman, David, & Narayanan, 2006), the phrases *let's all* and *I love it when* in online debates (Oraby,

Harrison, Reed, Hernandez, Riloff, & Walker, 2016), and the words *um* and *uh*, and ellipses, in written communication (D'Arcey, Oraby, & Fox Tree, 2019).

Behavioral cues include smiles, laughter, lip tightening, and slow nods (Caucci & Kreuz, 2012), as well as a lack of facial cues, such as when dead-panning (Attardo, et al., 2003). Gaze is another important cue, both towards a partner (Caucci & Kreuz, 2012) and away from a partner (Williams, Burns, & Harmon, 2009). Gaze towards a partner was observed when people produced spontaneous sarcasm and gaze away from a partner was observed when people expressed prepared sentences sarcastically or sincerely, although there may have been other differences in methods behind the opposite gaze findings. Heavy stress, nasalation, and slower rate of speech have also been linked to sarcasm (Cutler, 1976), although other researchers have found little evidence for prosodic consistency across ironic and sarcastic utterances, which shows that listeners do not rely on a specific set of acoustic cues to identify sarcastic and ironic utterances (Bryant & Fox Tree, 2005; Bryant, 2010).

Potential social-contextual cues to sarcasm include information about who is communicating and under what circumstances they are communicating. For example, watching a late-night comedy show or engaging in discussion of a topic that all parties feel cynical about may prime people to expect sarcasm. Some evidence for the use of social-contextual cues is suggested by observations that people are more likely to be sarcastic with friends (Rockwell, 2003). People may be more willing to use sarcasm with friends because they are less concerned about sarchasms and the

consequent loss of face when they say something that is not immediately understood by their addressee.

Other evidence for the importance of context is that a mismatched, contrasting context can alter judgements of sarcasm. When spontaneously produced ironic utterances from talk radio were couched in non-ironic contexts, people perceived less sarcasm than when the same items were couched in ironic contexts (Bryant & Fox Tree, 2002). Social factors also matter: The bite of sarcasm varied depending on what part of the country people came from (Dress, Kreuz, Link, & Caucci, 2008). In a cultural group that views sarcasm as negative, cues may be qualitatively different.

Cues to sarcasm are not always produced in the contexts one might expect. For example, it might seem that people would use more cues in situations where there is a higher likelihood of sarchasm or misinterpretation, such as when communicating with strangers. But people used more visual and auditory behavioral cues with friends (Caucci & Kreuz, 2012). As another example, it may seem that people would use less sarcasm when communicating using text, because it lacks potentially clarifying visual or auditory clues. But, in at least some situations, people used more irony in computer-mediated communication than in face-to-face communication (Hancock, 2004). Both the greater use of cues with friends and the greater use of sarcasm in text may result from different attitudes towards risk. More cues help friends avoid sarchasms. At the same time, fewer interpersonal risks from computer-mediated communication with anonymous strangers may enhance communicators' willingness to risk a sarchasm in the service of humor. In addition to humor, people also use

ironic language to communicate nuances of opinion and to strengthen relationship bonds (Dews & Winner, 1995).

In summary, linguistic, behavioral, and social-contextual cues all contribute to whether people are likely to interpret an utterance as sarcastic. Who one is communicating with and under which circumstances will also affect the likelihood of producing sarcasm. Friendship status in particular has an important effect on emotional expressions, as friends are more likely to share negative information with friends than with strangers (Segrin & Flora, 1998). People also felt more comfortable expressing themselves impolitely with friends than strangers, and interpreted utterances differently depending on whether they were produced by friends or strangers (Gupta, Walker, & Romano, 2007).

3.2.3 Understanding Sarcasm

Three models of sarcasm understanding are the Standard Pragmatic model, the Direct Access model, and the Parallel-Constraint-Satisfaction model. The Standard Pragmatic model holds that figurative language — including sarcasm — is always processed with the literal interpretation first. The Direct Access model takes an opposing view, that figurative language may be interpreted without first processing literal meanings (Gibbs, 2002). Similar to the Direct Access model, the Parallel-Constraint-Satisfaction model holds that people process a number of cues to ironic intent all at once (in parallel) in order to determine a speaker's meaning (Pexman, 2008). Even from an early age, children are able to select between literal and ironic meanings with equal facility (Pexman, 2008). While it is possible for literal

interpretations to arise first, contextual cues can lead people to interpret a sentence figuratively just as easily, assuming that the interpreter has a decent grasp of Theory of Mind and the executive function for non-literal sentence interpretation.

While people in general can understand nonliteral language even as children, there is some evidence that there may be individual differences in sarcasm comprehension. People with higher self-reported sarcasm were more confident in the identification of sarcastic intent, and were also faster at processing irony (Ivanko, Pexman, and Olineck, 2004). The more sarcastic a person described themselves to be, the more likely they were to select a sarcastic response from a set of four verbal responses to a vignette, when the vignette described the participant as having a conversation with a best friend (Ivanko et al., 2004). Notably for comparison to our study, however, sarcasm ratings for ironic criticisms were not affected by individual differences (Ivanko et al., 2004). The researchers also admitted to limitations in their experimental design, namely that the task was not a face-to-face, naturalistic interaction between individuals.

3.2.4 Misunderstandings in Communication

One issue with identifying miscommunication in communication is that it's unclear how often people correct misunderstandings in general. People can have misunderstandings without overt repair. For example, people are happy to answer survey questions even when their conceptualization of what is asked is different from the surveyor's, as evidenced by their willingness to change answers when provided more information about the survey question (Schober, Suessbrick, & Conrad, 2018).

People can also have misunderstandings that are not relevant to the task at hand; for example, a misunderstanding of what it means to have smoked a cigarette is irrelevant for a survey responder who has never smoked, puffed, or inhaled a cigarette, cigar, or pipe (Schober et al., 2018). Finally, people correct mistakes while moving their conversations forward without explicitly noting the need for repair; in fact, noting the repair can be viewed as leading the conversation in the wrong direction (Albert & de Ruiter, 2018).

When people do indicate misunderstanding, communication is improved. For example, when describing paths through mazes, people used feedback suggesting lack of understanding to move more quickly to more generalizable ways of conceptualizing the mazes, such as by viewing them as abstract grids rather than idiosyncratic paths (Healey, Mills, Eshghi, & Howes, 2018). Expressing lack of understanding has also been proposed to drive the presentation of alternate descriptions of abstract shapes in referential card tasks (Tolins, Zeamer, & Fox Tree, 2018), with multiple perspectives on the shapes linked to better performance at the task (Fox Tree, 1999; Fox Tree & Mayer, 2008) — a process that may also be behind the observation that increased feedback from any matcher in a multi-party referential card task resulted in better performance for all matchers (Fox Tree & Clark, 2013). Children can also take advantage of expressions of negative feedback. Children learned words better when they observed an addressee disagreeing with a partner's object label than when the addressee agreed, as long as the turns were interwoven (Tolins, Namiranian, Akhtar, & Fox Tree, 2017).

Perhaps because of these issues with recognizing miscommunication, most prior work on communication of sarcastic intent has not suggested any glaring difficulty. One study of emails suggests that sarcastic intent can be communicated effectively, although people overestimate their effectiveness (Kruger et al., 2005). People wrote sincere and sarcastic comments about 10 topics to each other. They thought they communicated accurately about 97% of the time but were actually only accurate 85% of the time — although they were more aligned when they communicated by voice (Kruger et al., 2005). Texted misunderstandings can result from linguistic or pragmatic miscommunication, as well as affective miscommunication, such as interpreting tone as angry when it wasn't or misunderstanding humor and sarcasm (Kelly & Miller-Ott, 2018). In a study of participants' self-reported misunderstanding, most affective miscommunication identified by participants was about tone, but about a fifth of the miscommunication was misinterpretation of humor and sarcasm (Kelly & Miller-Ott, 2018).

In contrast to the 85% agreement of Kruger et al. (2005), in another study of understanding sarcasm in written communication in a more naturalistic setting, researchers found very little agreement on nonliteral intent. Posts to an online fashion forum were assessed for sarcasm and humor by the posts' authors and by a variety of readers who ranged in similarity to the authors, such as by also belonging to the same forum, belonging to other forums, or being from the same demographic group but not being forum users (Kellner & Schober, 2018). Opinions about the celebrity fashions

were judged as positive or negative. Readers aligned with authors less than 10% of the time on recognizing sarcastic intent, although more similar readers aligned more.

This is similar to other research showing that friends can package messages for each other better than they can for strangers (Fussel & Krauss, 1989), suggesting some truth to people's anecdotal feelings that they understand sarcasm better when communicating with friends. But although knowing how a friend thinks seems like it would make people more likely to understand each other's sarcasm, it is also true that sharing information can create misunderstanding. In a referential communication game, directors who overlapped a lot with their addressees on object labels they were taught before the game were more likely to think their addressees knew a label that only they knew than directors who overlapped less with their addressees (Wu & Keysar, 2007).

3.2.5 Synchronous Movement

The important role of friendship status in sarcasm production and comprehension suggests that enhancing feelings of friendship may be one way to increase people's production of sarcasm and improve their ability to accurately detect it. Synchronous movement improves social bonds, so moving together may approximate feelings of friendship, causing people to use sarcasm in subsequent conversations.

Synchronous movement occurs when two or more people engage in physical action that overlaps temporally. The movements can manifest as the same action, such as when two people both wave their right hand, or as different actions, such as

when dancers engage in different movements set to the same music (Hove & Risen, 2009). Synchronous movement can affect perception of rapport; for example, the sound of footsteps occurring together and animated stick figures walking together gave perceivers the impression of heightened rapport in comparison to when they did not move together (Miles, Nind, & Macrae, 2009). Synchronous movement can also affect actual rapport. Consciously moving in synchrony facilitated rapport within dyads (Bernieri, 1988; Wheatley, Kang, Parkinson, & Looser, 2012) and produced more cooperative behaviors and feelings of being on the same team (Wiltermuth & Heath, 2009). Both large and small motor movements heightened cooperative behaviors (Wiltermuth & Heath, 2009). Synchrony also elicited compassion and altruistic behavior (Valdesolo & DeSteno, 2011). Unconscious synchrony facilitated smoother social interactions and increased regard for communicative partners (Chartrand & Baugh, 1999). Body posture mimicry also enhanced prosocial behavior towards others (van Baaren, Holland, Kawakami, & van Knippenberg, 2004).

We tested the effects of synchronous and non-synchronous movement on subsequent communication. If synchronous movement boosts general feelings of social connection, then sarcasm production might be increased and comprehension might be improved after synchronous movement compared to non-synchronous movement.

3.2.6 Current Studies

We propose that enhanced social connection from synchronous movement

prompts increased sarcasm use and more accurate sarcasm detection between conversational participants.

In Study 1, participants engaged in either synchronous or individual movement activities before participating in a sarcasm-inducing conversation about badly dressed celebrities, a task developed by Hancock (2004). Immediately after talking about the celebrities, participants went into individual booths to view a recording of their conversation, a task that is similar to one used by Amati and Brennan (2016) to investigate white lies. In private booths, each participant identified times in the recording they had used sarcasm and where they believed their partner had used sarcasm. If synchronous collaborative movement increases friendliness between strangers then people should feel more comfortable being sarcastic after they have moved together with their conversational partner. This would be observable through both individuals' identifications of their own sarcastic productions and individuals' identifications of their addressees' sarcastic productions.

In Study 2, we compared an individual's self-identified sarcastic utterances to perception of the same utterances by their interlocutor. That is, we tested to what extent dyad participants agreed on what was sarcastic in their conversation. We compared accuracy across the synchronous and non-synchronous conditions, as well as inaccuracy across conditions. Because people could note as few or as many sarcastic utterances as they wanted, it was possible for people to be both highly accurate (e.g. identifying all five sarcastic utterances their partner self-identified) as well as highly inaccurate (e.g. identifying ten sarcastic utterances that their addressee

did not identify). We predicted that participants who engaged in synchronous movement would agree more and disagree less than participants who engaged in the non-synchronous movement. More specifically, we predicted that turning strangers into friends would result in more cues to sarcasm (cf. Caucci & Kreuz, 2012), which would enhance conversational participants' abilities to correctly identify sarcasm. In addition, conversational participants who engaged in synchronous activity may be more accurate because they pay more attention to their partners (cf. Macrae et al., 2008).

In Study 3, we examined the details of participant agreement in both quantitative and qualitative ways. These details can help determine whether there are study design differences that explain the conflicts between our results and others'.

Our work differs from earlier work in numerous ways. We went beyond tests where communicators' sarcasm was produced on demand in answer to predetermined questions (e.g., Rockwell, 2003) or where production of sarcasm was assessed by selection from restricted options (e.g., Ivanko et al., 2004) by having communicators identify the sarcasm in their own and their partner's talk immediately after their conversations. We also went beyond tests that used prepared sarcastic materials such as written text (e.g., Burgers, van Mulken, & Schellens, 2012; Pexman & Olineck, 2002) or puppet dialogue (e.g., Nilsen, Glenwright, & Huyder, 2011) by inviting sarcasm production in spontaneous communication. In addition, our assessment of what was sarcastic went beyond identification based on the interpretations of people who were not part of the communication, such as trained identifiers (e.g., Burgers,

van Mulken, & Schellens, 2011; Campbell & Katz, 2012) or experimenters observing the conversations (e.g., Caucci & Kreuz, 2012): We tested how original communicators produced and identified sarcasm.

While we were hopeful that our predictions would be borne out, we foreshadow our results with a cautionary note: People overestimate how well they communicate with their friends. Prior researchers have noted that although people think they communicate better with friends (and spouses), they actually do not (Savitsky, Keysar, Epley, Carter, & Swanson, 2011). Experimental participants tried to get their addressees to select the appropriate meaning of an ambiguous sentence they produced. They produced sentences like, "What have you been up to?" which "could convey irritation that someone is late, interest in someone's well-being, suspicion over possible romantic infidelity, or playful conjecture about an imminent surprise party" (Savistky, et al., 2011, p. 271). Although they thought their friends and spouses would be much better than strangers at selecting the right intention, friends were only marginally better, and spouses were not at all better. With respect to sarcasm, we may find that creating a feeling of friendship through synchronous movement may similarly have no effect on accuracy of sarcasm identification.

3.3 Study 1: Synchronous Movement and Sarcasm

To enhance feelings of interpersonal collaboration, in the synchronous movement condition participants engaged in a brief movement activity facing each other. In the non-synchronous movement condition, participants engaged in the same movement activities, but facing away from each other. After the movement activity,

participants engaged in a conversation designed to elicit sarcasm. Immediately after this conversation, participants reviewed a videorecording of their conversation and individually noted where they produced sarcasm and where they thought their addressee had produced sarcasm.

3.3.1 Method

3.3.1.1 Participants

One hundred thirty students from the University of California, Santa Cruz participated in this study in exchange for course credit. They were grouped into 65 dyads. Due to the difficulty of getting two participants in the lab simultaneously, we ran as many dyads as possible over an 8-month period in 2017. There was no stopping rule. Sixteen dyads were excluded from analyses because research assistants who ran these participants were found to be making small talk with participants prior to their participation. Because research assistants were not blind to which condition was being run, this created a potential camaraderie confound between conditions. Three dyads were excluded from analyses because they reported that they were friends before participating in the study. Two dyads were excluded due to poor audio quality that made it difficult for participants to complete the experimental tasks. Finally, because the goal of the present study was to examine sarcasm production, the five dyads that included a person who did not identify any sarcasm were also excluded from analysis. People who did not identify sarcasm could be people who were not sensitive to sarcasm (we note here that lack of sarcasm sensitivity has been used as an indicator of communicative problems; Peterson, Wellman, & Slaughter,

2012). Another possibility is that people who did not identify sarcasm were inattentive or eager to leave the experiment quickly. Of the remaining 39 dyads, 46 participants identified as female, 31 participants identified as male, and one participant identified as non-binary. Participants' ages ranged from 18 to 37 (M = 20.04, SD = 2.59).

3.3.1.2 Procedure

Dyads were randomly assigned to the synchronous or non-synchronous condition. Each dyad engaged in either a synchronous or non-synchronous movement activity lasting approximately six to eight minutes. In the brief movement activity, participants either faced each other while engaging in synchronous movement, or faced away from each other without collaborating. In both conditions the movements were identical except for whether the participants faced towards or away from each other. Participants were not recorded during the movement activity in order to facilitate their comfort (cf. Christenfeld & Creager, 1996, where participants were asked to dance in front of a camera in order to induce anxiety).

In the synchronous condition, one participant was randomly designated the leader and interpreted the experimenter's movement instructions, and the second participant mimicked the leader's motions. First, participants performed movements as instructed by the experimenter, such as moving their left arm in circles and swaying their bodies. Then participants passed an imaginary ball of varying weights to each other. Finally, participants engaged in a mirroring exercise, where they were

instructed to make faces depicting emotions. Halfway through this synchronous movement activity, the partners switched leading and following roles.

In the non-synchronous condition, participants interpreted the same instructor-directed movements individually, then threw a ball of varying weights against the wall, and finally made the same faces while facing a wall.

After the movement activity, participants engaged in a conversation designed to elicit sarcasm. Stimuli for the conversation consisted of well-known celebrities wearing ugly outfits (cf. Hancock, 2004). One male and four female celebrities were depicted across four laminated 8.5" x11" prints. Participants were also given an envelope containing possible conversation topic prompts to use if their conversation got stuck. See Appendix A for the prompts. Of the 39 dyads included in analyses, 31 dyads used the prompts. On average, they started using them about 4 minutes 19 seconds into the conversation (*SD* = 2 min 38 sec).

The experimenter told participants that they would be engaging in a ten-minute video-recorded conversation, started recording, and then left the room to allow conversation to commence. Participants were given ten minutes for the conversation. This was the average length of time found in Hancock (2004) for face to face conversations. The participants' 10-minute conversations were video-recorded with Logitech c920 HD Pro webcams at 1280x720 resolution.

Immediately following the conversation, participants were directed into separate rooms to review the video recording of their conversation, identifying at what points in the conversation they used sarcasm and when they believed their

partner used sarcasm (cf. Amati & Brennan, 2016). The participants' responses were identified as timestamps indicating the beginning of utterances that corresponded to the video of the dyads' conversations. Finally, participants answered a 13-item survey about their typical sarcasm use, their familiarity with their partner prior to the study, their comfort with the task, and demographic information, see Appendix B.

3.2 Results

All individual reports of sarcasm were aggregated into a spreadsheet which was then analyzed using Python scripts and SPSS Statistics. In the 39 dyads there were 775 reports of sarcasm, 406 in reports of speakers' own use of sarcasm and 369 in speakers' reports of their addressees' sarcasm. Some of these uses may overlap, because a person might report themselves as being sarcastic with a particular utterance, and their addressee may also identify the same utterance as sarcastic. That is, all instances of agreement will be double counted in this 775 figure. In Study 2, we look more closely at agreement.

Sarcasm was prevalent throughout the 10-minute sessions, with an approximately equal distribution of reports throughout the conversation. See Figure 1.

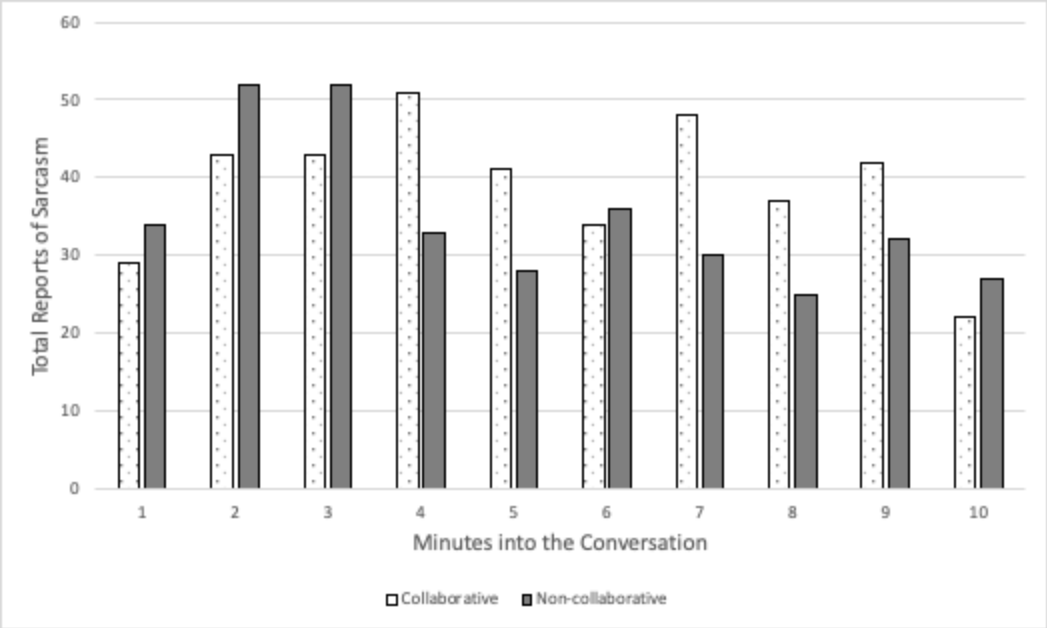

Figure 1. Frequency of sarcasm reports over all 10-minute conversations.

In the synchronous condition there were 229 instances of self-sarcasm and 182 instances of other-sarcasm. In the non-synchronous condition there were 177 instances of self-sarcasm and 187 instances of other-sarcasm. The average number of sarcastic instances per 10 minutes of dyadic conversation was 9.97 (*SD* = 7.09). The five dyads where at least one participant identified no sarcasm (11% of 44 possible dyads) were not included in this analysis.

Participants who engaged in synchronous movement reported more sarcastic utterances in their speech, *M* = 6.36, *SD* = 4.31, compared to participants who did not engage in synchronous movement, *M* = 4.21, *SD* = 3.52, *t*(76) = 2.42, *p* = .018, 95% CI for the difference, [0.38, 3.91]. See Figure 2.

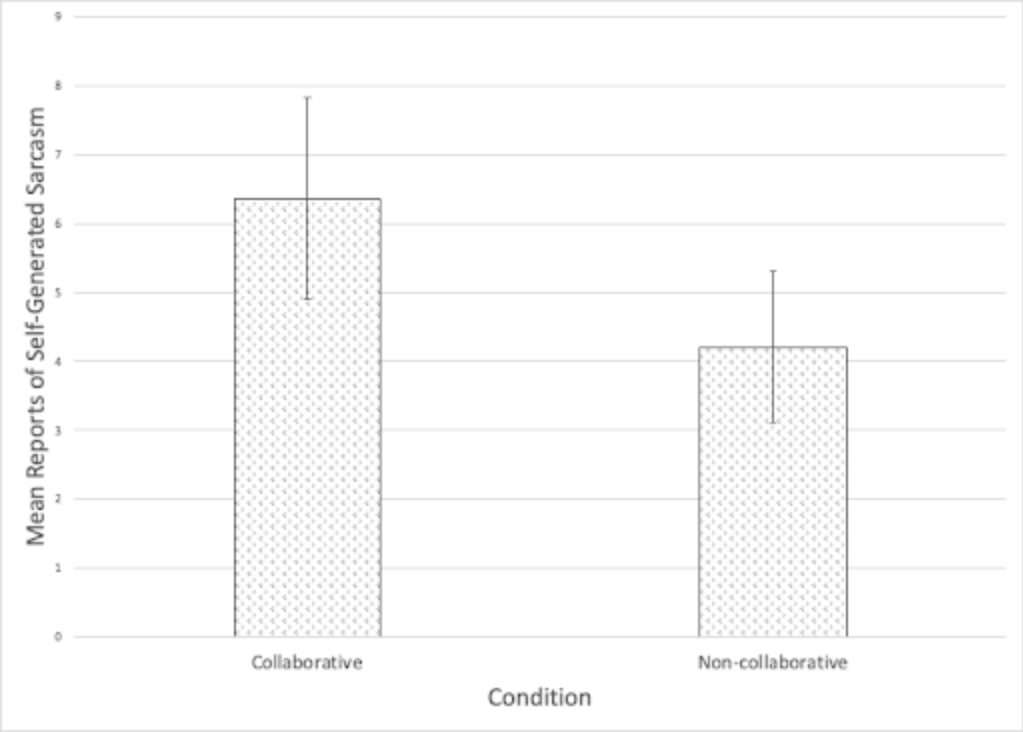

Figure 2. Mean of self-reported sarcastic utterances by condition. Error bars represent 95% confidence intervals.

We assessed the role of the research-assistant-camaraderie confound by running an analysis that included the sixteen removed dyads. Including these participants yielded similar results, with the synchronous group still reporting more instances of sarcasm (M = 6.11, SD = 4.39) than the non-synchronous group (M = 4.43, SD = 4.06), $t(108) = 2.09$, $p = .039$, 95% CI for the difference, [.09, 3.28].

Participants who engaged in synchronous movement reported similar levels of sarcastic utterances in their partners' speech, M = 5.06, SD = 4.11, compared to participants who did not engage in synchronous movement, M = 4.45, SD = 4.02, $t(76) = 0.65$, $p = .52$, 95% CI for the difference, [-1.24, 2.44]. See Figure 3.

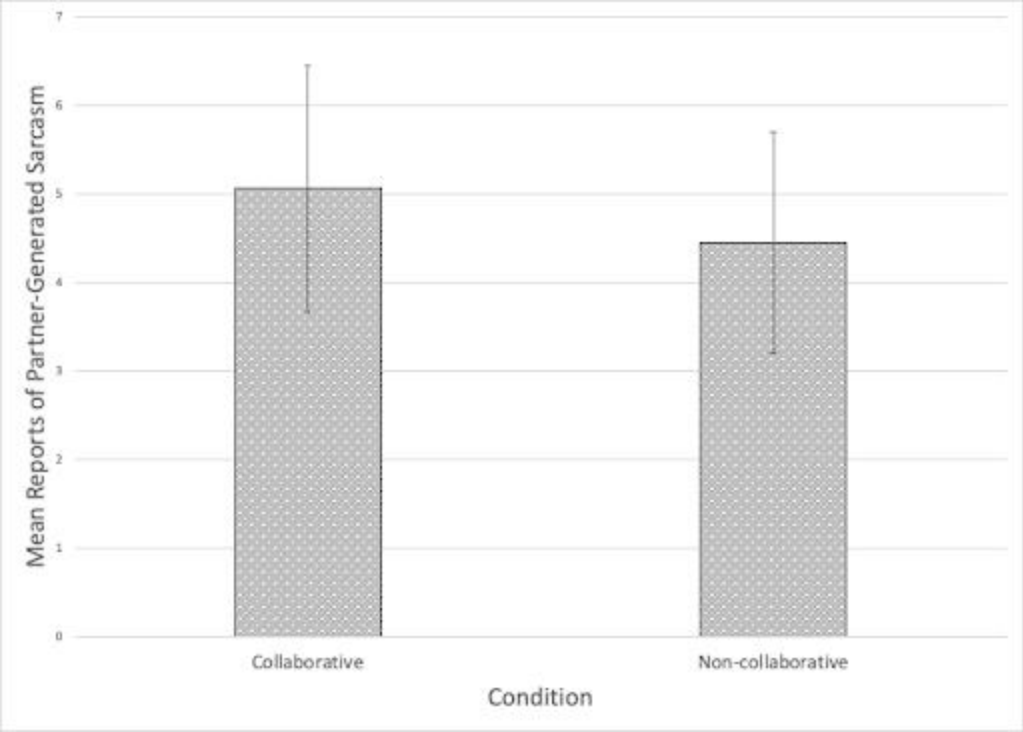

Figure 3. Mean of partner-reported sarcastic utterances by condition. Error bars represent 95% confidence intervals.

In addition to differences by condition, the more participants reported using sarcasm, the more they perceived their partners to have used sarcasm. Every sarcastic utterance a participant reported was associated with a 0.54 increase (the slope of the regression line) in that participant's report of their partner's sarcastic utterance, $r(76) = .55$ (the strength of the correlation), $p < .001$, 95% CI [.37, .69]. Self-reported sarcasm explained a significant proportion of the variance in report of their partner's sarcasm, $R^2 = .30$, $F(1,76) = 32.29$, $p < .001$. See Figure 4.

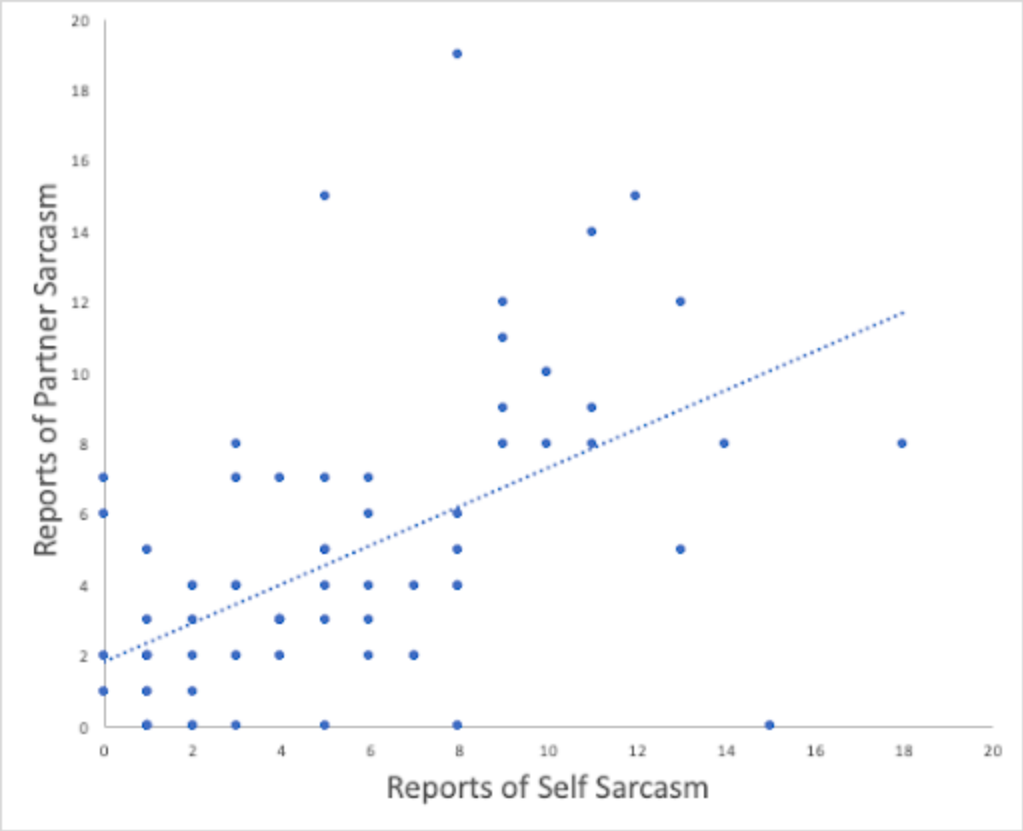

Figure 4. Relationship of self-reported sarcastic utterances to reports of partner's sarcastic utterances collapsed across synchronous and non-synchronous conditions.

Results were similar for synchronous and non-synchronous conditions. For the synchronous condition, every sarcastic utterance a participant reported was associated with a 0.45 increase in that participant's report of their partner's sarcastic utterances, $r(34) = .43$, $p = .01$, 95% CI [.11, .66]. Eighteen percent of the variance in the number of reports can be explained by this relationship, $R^2 = .181$, $F(1, 34) = 7.54$, $p = .01$. For the non-synchronous condition, every sarcastic utterance a participant reported was associated with a 0.6 increase in the participant's report of their partner's

sarcastic utterance, $r(40) = .68$, $p < .001$, 95% CI [.47, .82]. Forty-six percent of the variance in number of reports can be explained by this relationship, $R^2 = .47$, $F(1, 40) = 34.75$, $p < .001$.

There was no relationship between what one person reported for their partner and what the partner reported for themselves, however; $r(76) = .01$, $p = .93$ across both conditions (95% CI [-.21, .23]), $r(34) = .08$, $p = .645$ in the synchronous condition (95% CI [-.26, .40]), and $r(40) = -.11$, $p = .51$ in the non-synchronous condition (95% CI [-.40, .20]).

We use the term *sarcasm reciprocity illusion* to describe the difference between how much people think their addressees produce sarcasm and how much sarcasm is actually produced. Said another way, people think that their addressees produce sarcasm commensurate with their own sarcasm production, but there is actually no relation.

One possibility is that people report more sarcasm because they feel more comfortable with each other. We found some evidence supporting this. People reported that the movement activity was more effective at making them feel comfortable with their partner when facing each other. In answer to the question "Did the movement activity (throwing the imaginary ball) make you feel more or less comfortable with your partner?," participants in the synchronous condition reported that the movement activity was better at facilitating comfort on a seven-point scale from 1 *extremely uncomfortable* to 7 *extremely comfortable*, $M = 4.69$, $SD = 1.17$,

compared to the non-synchronous condition, $M = 3.88$, $SD = 0.89$, $t(76) = 3.49$, $p = .001$, 95% CI for the difference [.35, 1.28].

On the other hand, we also found evidence against a comfort-sarcasm link. When asking people about their actual level of comfort with their partner during the procedure, people's reports did not differ. In answer to the question, "How comfortable were you with your partner after the movement activity?," participants in the synchronous condition reported being similarly comfortable on a seven-point scale from 1 *extremely uncomfortable* to 7 *extremely comfortable* scale, $M = 4.66$, $SD = 1.26$, compared to the non-synchronous condition, to $M = 4.43$, $SD = 0.96$, $t(73) = 0.91$, $p = .37$, 95% CI for the difference, [-.28, .74] (three participants did not answer this question). Further, by the time they had engaged in the conversational task, people did not report feeling significantly more or less comfortable across conditions. We compared participants' responses to the question, "How comfortable did you feel with your partner after engaging in a conversation about badly dressed celebrities?" Participants in the synchronous condition ($M = 5.33$, $SD = 1.04$) reported being similarly comfortable to participants in the non-synchronous condition ($M = 5.17$, $SD = 1.39$), $t(76) = 0.59$, $p = .56$, 95% CI for the difference, [-.40, .73], on a seven-point scale from 1 *extremely uncomfortable* to 7 *extremely comfortable*. Finally, as expected, participants reported feeling more comfortable with their partner after the ten-minute conversation ($M = 5.31$, $SD = 1.22$) than after the movement activity ($M = 4.53$, $SD = 1.11$), $t(74) = 6.17$, $p < .001$, 95% CI for the difference [.52, 1.02].

To understand the relationship between comfort and sarcasm better, we correlated how comfortable a person felt with their rate of sarcasm. There was no correlation for any of the three comfort questions. Their answers to "Did the movement activity (throwing the imaginary ball) make you feel more or less comfortable with your partner?" were not correlated with their sarcasm production, $r(76) = -.04$, $p = .76$ overall (95% CI [-.26, .18]); $r(34) = -.13$, $p = .43$ in the synchronous condition (95% CI [-.44, .20]); $r(40) = -.17$, $p = .28$ in the non-synchronous condition (95% CI [-.45, .14]). Their answers to "How comfortable were you with your partner after the movement activity?" were also not correlated, $r(73) = .09$, p, $= .45$ overall (95% CI [-.14, .31]); $r(34) = -.1$, $p = .57$ in the synchronous condition (95% CI [-.41, .23]); $r(38) = .30$, $p = .058$ in the non-synchronous condition (95% CI [-.01, .56]). And their answers to "How comfortable did you feel with your partner after engaging in a conversation about badly dressed celebrities?" were not correlated, $r(76) = .02$, $p = .84$ overall (95% CI [-.20, .24]); $r(34) = .07$, $p = .69$ in the synchronous condition (95% CI [-.26, .39]); $r(40) = -.04$, $p = .79$ in the non-synchronous condition (95% CI [-.34, .27]). We conclude that comfort was not the driving factor behind differences in sarcasm reporting.

3.3.2 Discussion

Using the procedure described here, we collected a large corpus of potentially sarcastic utterances. There were about 700 different cases of sarcasm (see Study 2 for information on removing double-counted instances). For comparison, researchers

identified 395 cases of sarcasm in the wild over a two-year period, mostly from face to face and classroom communication (Eisterhold et al., 2006).

In this first test of sarcasm producers' self-identification of where they produced sarcasm in a conversation immediately after producing the conversation, we found that people produced about one sarcastic remark per minute on average. Further, across both conditions, the more sarcastic utterances a person self-reported, the more sarcastic utterances they reported for their partners. Notably, however, people were not accurate in their reporting: People perceived sarcasm in their partners at rates similar to their own production rate regardless of what their partners reported about their own behavior. This is the first demonstration of this sarcasm reciprocity illusion.

A related observation is that people who think of themselves as more sarcastic select more sarcastic messages to send to addressees (Ivanko et al., 2004), although in this earlier study researchers looked only at what producers felt about themselves and the message they selected to send to addressees rather than what producers actually produced while speaking and what they predicted about their addressees' behavior. The authors also wrote, "It is important to note that although there was evidence, in the present experiments, that individual difference measures predicted performance on various irony tasks, the amount of variance explained by these individual difference variables was small" (pp. 265-266).

Although synchronous movement had no effect on whether each partner's self-reports matched the other partner's self-reports in a dyad, synchronous

movement did increase sarcasm identification within an individual: People identified more sarcasm in themselves after moving synchronously than after moving non-synchronously. The higher identification of self-sarcasm may have resulted from the synchronous movements' creating a feeling of friendship, allowing speakers to risk potential misinterpretations that sarcasm can bring about. Sarcasm is a risky linguistic phenomenon because it can be misinterpreted a number of ways, including but not limited to: (1) different base rate expectations for sarcastic utterances, which can stem from individual, community, and cultural differences, (2) differing definitions of what is sarcastic and humorous, (3) deliberate ambiguity that allows certain addressees to understand the sarcasm while others don't, for example to hide the true communication from a child, and (4) unintentional ambiguity that arises when common ground is not properly established between speaker and addressee.

Notably, speakers did not report more sarcasm in their addressees after moving synchronously. So although they self-identified as having produced more sarcasm, they did not identify others as having produced more sarcasm. One interpretation of this is that synchronous movement influenced what a speaker was willing to risk but did not influence how likely a speaker was to feel that their addressee was also willing to risk more.

3.3.3 Assessing Sarcasm Agreement

In Study 2, we assessed to what degree dyadic participants agreed with each other's sarcasm identification. We know that the number of times people thought their addressees were using sarcasm did not match the number of times their

addressees identified sarcasm in their own talk. But we do not know to what extent sarcasm identifications overlapped. For example, one person may have self-identified ten sarcastic comments, and their partner may have ascribed five sarcastic comments to them. Those five could be subsumed under the ten, they could overlap a little, or they could be entirely different.

We also assessed to what degree three third-party observers came to similar conclusions about what was sarcastic. In studies of deliberately produced sarcasm, two coders generally achieved high agreement. For example, a second coder achieved over 93% agreement with a first coder when sorting people's written productions into *sarcastic*, *literal*, or *other* bins (Campbell & Katz, 2012). The agreement was so high that the second coder coded only 18% of the data, to confirm that data could be sorted reliably; the remainder of the coding rested on the first coder's sorting (Campbell & Katz, 2012).

The additional rater technique was also used to confirm the reliability of sarcasm sorting in another study, with the conclusion that the rater "confirmed our analysis," although no reliability statistics were provided (Eisterhold et al., 2006, p. 1246). In this study sarcasm was identified and noted in the course of everyday interactions, mostly face-to-face and classroom communication (Eisterhold et al., 2006). In another study, two coders agreed on what was ironic or not for 97% of lines in movie reviews (Burgers et al., 2011). They used a sorting technique they developed called the Verbal Irony Procedure. And in another study, this one using three coders, high agreement ($\alpha > .90$) was achieved in rating productions of deliberately sarcastic

and sincere remarks on a 1 to 11 scale from "not sarcastic at all" to "extremely sarcastic" (Huang et al., 2015, p. 165).

Because context can matter for successful sarcasm identification, our coders had access to the transcripts and videos in making their assessments. Context can make a big difference. In a study sorting ironic and non-ironic lines from movie and books reviews and satirical and real news articles, a Krippendorff's alpha of .49 was achieved when lines were presented in isolation and .72 when lines were presented with the documents they came from along with a definition of irony (Reyes & Rosso, 2014). Two evaluators were used. The researchers summarized sorting lines in isolation as "almost a random process" (Reyes & Rosso, 2014, p. 14).

There is reason to expect both that the outside observers will be better at sarcasm detection, and that they will be worse. It is possible that outside observers will be better able to spot sarcasm because they are not engaged in the conversation. Outside observers spotted evasion better than people engaged in conversations (Bly, 1993). Because they were not engaged in building up their own and their addressees' contributions into coherent dialogues, as conversational participants were, outside observers were better able to identify when one participant was avoiding answering a question. Similarly, people engaged in a conversation could not detect each other's little white lies (Amati & Brennan, 2016), although it remains an open question whether outside observers can.

Alternatively, outside observers may be worse at sarcasm identification. Because they could not ground their communication with direction-givers,

overhearers were worse than direct addressees at understanding in a referential communication task (Schober & Clark, 1989). In this task, people identified items from a set and put them in the order the direction-giver specified. Sarcasm detection may follow the pattern observed in the referential communication task rather than the little white lies or evasion tasks because sarcasm is meant to be understood, whereas little white lies and evasion are meant to go undetected.

3.5 Study 2: Sarcasm Identification Accuracy

To assess the extent to which people accurately identified sarcasm in their addressees' communication, we compared the timestamps at which each person indicated their partner had expressed sarcasm to where the expresser had indicated that they had expressed sarcasm.

3.5.1 Method

The data from Study 1 were analyzed, with one dyad excluded due to a misunderstanding of the instructions – the participant did not write timestamps, but instead transcribed the speech perceived to be sarcastic. Therefore, although this data was interpretable for the analyses in Study 1, it was not interpretable for Study 2. This left 758 of the original 775 sarcastic identifications for analysis in Study 2. Measures of sarcasm were gathered from five sources: the speaker's own rating, their addressee's rating, and three third-party ratings. The third parties were research assistants who separately viewed the conversations after they had occurred.

The participants' indication of sarcastic content was derived from their conversation, as indicated by the timestamps on their response sheets from Study 1.

Outsider observations of sarcasm were coded for sarcastic content by three research assistants who worked independently. The research assistants used both the transcript and the video of the conversation to make a holistic evaluation of sarcastic content from information provided by the participants' body language, eye movement, gaze, and linguistic meaning. The research assistants were instructed to use their best judgement to identify cases of sarcasm from the perspective of the people who used it, as opposed to coding according to a theoretical definition of sarcasm. This method of using "folk definition[s]" of sarcasm, as opposed to academic definitions, has been used by others (Eisterhold et al., 2006, p. 1246; Campbell & Katz, 2012; Reyes & Rosso, 2014). Three coders exceeds the number others have successfully used to assess sarcasm, two (Burgers et al., 2011; Campbell & Katz, 2012; Eisterhold et al., 2006; Reyes & Rosso, 2014). Participants and research assistants marking timestamps that displayed a difference of less than three seconds and referred to the same speaker were marked as an agreement.

There are good reasons to use three seconds as a cutoff: Researchers who looked across a variety of corpora assessed the mean rate of speech in storytelling to be 3.43 (*SD* = .43) syllables per second, and in interviews to be 4.31 (*SD* = .10) syllables per second (Kowal, Wiese, & O'Connell, 1983, p. 389), which averages to 3.87 syllables per second. With an average of 1.4 syllables per word assessed in another study (Andrews & Ingham, 1971, p. 129), this is about 2.76 words per second. So, a three second cutoff means looking at units that are about 8.28 words long. The average length of phrases identified as sarcastic by our coders was 7.32

words. So, a rolling window of three seconds should be good enough to capture most identifications of the same sarcastic utterance without including utterances that refer to different conversational contributions. A manual analysis of the participants' timestamps revealed that after three seconds it became increasingly challenging to disambiguate the utterance to which the report referred. Due to the relative scarcity of sarcasm reports within the corpus, we treated the three-second rule as a rolling window to keep chains of temporally close sarcastic reports together (e.g., reports at 1:00, 1:03, and 1:06 would be treated as a single instance of sarcasm) as opposed to attempting to separate them with judgment calls. Timestamps that were not temporally close to a timestamp identified by a partner were marked as disagreements. After running our computational analyses, we also manually coded 580 overhearer reports to determine how accurate our computational analysis was likely to be.

We also computed Krippendorff's alpha as a measure of reliability between groups of annotators (Hayes & Krippendorff, 2007). In order to compute it, we needed to treat the continuous data as a set of discrete ratings. We leveraged our three-second window to split our data into a series of overlapping three second windows, resulting in approximately 600 time windows per 10 minute conversation (e.g., 0-2 seconds, 1-3 seconds, 2-4 seconds, etc.), and then dichotomized our five annotators' data to be either 0 (no sarcasm reported within this time window) or 1 (sarcasm reported within this time window). Alpha was then computed across all 39

dyads. This helps us compare reliability between the conversational participants and the overhearers.

3.5.2 Results

We analyzed three types of agreement. In one analysis, we assessed pairs of raters' agreements by finding proportional overlap between their ratings of sarcasm. In the second, we focused on the sarcastic items themselves, assessing how often particular items were identified as sarcastic and by whom. The third analysis considered agreement as reliability and used Krippendorff's alpha as described above. Python scripts were written to assist with all three analyses.

To determine how many false positives and false negatives we could expect with our computational analysis using the three second rule, we manually examined sarcastic reports from three overhearer coders. Out of 580 total reports of sarcasm, there were five instances (around 2%) where two reports fell within three seconds of each other but did not refer to the same statement, and 18 instances (around 6%) where two reports fell outside three seconds of each other but did refer to the same statement. We take this as evidence that in almost all cases, the three second rule worked well for our data.

3.5.2.1 Pairs of raters

In the pairs-of-raters analyses, we compared one rater's reports of sarcasm with one other rater's reports of sarcasm, either across the conversational participants or across the overhearers.

3.5.2.1.1 Conversational participants

Of the 758 participant-identified cases of sarcasm, six referred to the same person, occurred within the same rolling three second window, and were coded by the same participant. These six annotations were treated as duplicates and were excluded from the analysis, leaving 752 distinct reports. There were 152 annotations from conversational participants that fell within the same rolling time window and referred to the same speaker, suggesting that there were 76 instances (152/2) of agreed-upon sarcasm (note that the analyses for Table 1 includes 3 annotations that occurred within the rolling window and so were double counted). One way to look at this is as 20.2% agreement (152/752). There were 44 agreements (22.4%) in the synchronous condition and 32 agreements (17.8%) in the non-synchronous condition. Krippendorff's alpha told a similar story between the conversational participants, α = .10, 95% CI, [.06, .14].

Sarcastic utterances that the producer and addressee agreed were sarcastic were counted and divided by the total amount of sarcasm cases for the respective dyad. Dyads who participated in the synchronous condition had similarly low levels of agreement (M = 20%, SD = 19.2%) to dyads who participated in the non-synchronous movement condition (M = 19.2%, SD = 14.8%), $t(36) = -.15$, $p = .89$, 95% CI for the difference, [-.12, .11]. That is, about 80% of sarcasm identifications were not agreed upon.

We examined each of the 76 agreements between the conversational participants to see how many of them were within three seconds of an overhearer's

report. Thirty six of the agreements, or approximately 47% had an overhearer's report within three seconds, suggesting that even when conversational participants agree, it does not necessarily mean that overhearers will also agree.

3.5.2.1.2 Overhearers

Overhearers and conversational participants didn't necessarily label the same utterances as sarcastic; that is, agreement across overhearers does not necessarily refer to the same items as agreement across conversational participants. Due to our three-second rolling window for marking agreement, each additional rater changed what counted as agreement. For instance, if rater A marked a sarcastic utterance at 37 seconds, rater B marked one at 41 seconds, and rater C marked one at 39 seconds, then all three reports would be treated as a single instance of sarcasm when including all three raters in the analysis. However, the same reports of sarcasm might show no agreement when comparing only rater A and rater B. Nonetheless, while some information is lost by congealing all ratings into a single agreement analysis, comparing more than two raters' reports at once is interesting for identifying pieces of dialogue that were likely viewed as sarcastic by many raters. This calculation is less useful when looking for differences in accuracy across rater type (e.g., participant vs. overhearer).

Overhearer pairs agreed more with each other than participants agreed with each other. Of the 580 total annotations from the three third party raters for 38 transcripts, there were 315 that only one rater marked as sarcastic (using a three-second rolling window), 182 that two raters agreed were sarcastic, and 81 that all

three raters agreed were sarcastic (two annotations were excluded because the same overhearer coded two annotations within the same three-second rolling window). We calculated rater agreement for pairs of raters by taking the percent of their ratings that matched compared to the total number of ratings those two raters provided. Rater J (226 ratings) agreed with rater P (208 ratings) 29.5% of the time and rater E (146 ratings) 29.6% of the time. Rater P agreed with rater E 29.4% of the time. The average agreement across our raters was 29.5%, or around 140% the proportional agreement of the conversational participants. However, it is important to note that each additional overhearer increases the likelihood of agreement. Krippendorff's alpha accounts for this potential bias: for our three overhearer raters, $\alpha = .18$, 95% CI [.15, .21].

About half of sarcasm identifications were not agreed upon, to a similar degree across movement conditions. There were 361 agreements (53.9%) in the synchronous condition and 311 agreements (46.6%) in the non-synchronous condition.

3.5.2.2 Sarcastic Items

In the second type of agreement we assessed, we began with the sarcastic items and counted how often different types of participants – the producers, the direct addressees, the overhearers – identified those items as sarcastic. We report the frequency of various configurations of agreements when including all raters in Table 1.

Table 1. The frequency of agreement configurations.

	0 Overhearers	1 Overhearer	2 Overhearers	3 Overhearers	Total
Self + Partner	41	18	14	6	79
Self Only	234	43	21	8	306
Partner Only	219	35	25	7	286
Overhearer Only		202	36	8	246
Total	494	298	96	29	917

It is important to note that the frequencies in Table 1 reflect sarcastic items and not instances of reporting sarcasm. For example, the 6 items that were agreed upon as sarcastic by both conversational participants and all three overhearers represent 30 reports of sarcasm.

Some sarchasms were so chasm-like, in addition to the addressee's failing to identify the sarcasm, none of the three overhearers identified the sarcasm, as with "she looks like a fancy cocktail drink" in Example 1:

(1) A: I just think honestly if the top was better it might have saved it
B: yeah
A: but like I just think that the mixture
B: or if it was all one color
A: yeah
B: the fact that it goes into different colors I think that also just makes it look- she looks like a fancy cocktail drink

Of the 917 utterances that were identified as sarcastic by somebody, over 25% were cases where only the speaker identified the sarcasm. See Table 1.

Some sarcasm was identified by the addressees, but not by the speaker or the overhearers as in "It was a weird time for fashion" in Example 2 (asterisks indicate overlap):

(2) C: It was a weird time for fashion *in these past couple of years*
D: *I know*

Of the 917 utterances that were identified as sarcastic by somebody, about 24% of them were cases where only the addressee identified the sarcasm. See Table 1.

Some sarcasm was identified by a single overhearer, as in "she's got some like sky high heels like sky high thigh highs" in Example 3:

(3) E: yeah it's like black on black everything so she's got some like sky high heels
F: mhm
E: like sky high thigh highs
F: and then all the way up she has like some fishnets

Taken together with the speaker-only and addressee-only numbers, of the 917 utterances that were identified as sarcastic, about 71% of them were cases where only one person identified the sarcasm. See Table 1. This leaves 29% of cases where two or more people identified an utterance as sarcastic.

In about 4% of the 917 cases, these two people were the conversational participants – not outside observers. See Table 1. This is how sarcasm is sometimes intended – for only direct addressee to get the joke. In Example 4 the participants identified sarcasm in "that's not really its greatest look to be honest," but none of the observers did:

(4) G: yeah I don't really- I think it's like a skirt with a top
H: yeah
G: but that- that's not really its greatest look to be honest
H: Window: yeah

Similarly, in Example 5, the participants identified sarcasm in "I would fall over the first time," but none of the observers did:

(5) G: I've never seen her wear that before
H: me neither those shoes are ridiculous I don't know how you would walk in them *[laughs]*
G: *I know they're so thick* like look how far off
H: yeah I would fall over the first time *[laughs]*

These cases could be thought of as sarchasms for the observers – although they could also be failed observer identifications.

There were also cases of the inverse, where all three overhearers thought an utterance was sarcastic but the conversational participants did not, as with "what this isn't what you wear everyday" in Example 6:

(6) I: it's pretty raunchy
J: yes
I: definitely not your everyday outfit that's probably like a-
J: what this isn't what you wear everyday

These could be falsely-identified sarcasms for the observers, or failed identifications by the participants. Fewer than 1% of the 917 cases fit this category, but of course these would become increasingly rare with increased numbers of overhearers. See Table 1.

There were also cases where a conversational participant identified sarcasm in their partner, as did at least one overhearer, but the person who produced the utterance did not identify the sarcasm, as with "did you even have fun making this" in Example 7:

(7) K: this one like even she's miserable wearing this
L: yeah she's yeah
K: like there's no escape in that
L: like did you even have fun making this [laughs]

Cases like this accounted for 7% of 917 cases. See Table 1. These can be either unwitting sarcasm production, or cases of a speaker failing to identify their sarcasm after they produced it.

Finally, there were six cases where the speaker, the direct addressee, and all three overhearers identified sarcasm, as with "c'mon why not denim sunglasses" in Example 8:

(8) Speaker M: they're blue jeans as hell
Speaker N: yeah is that a denim dress
Speaker M: all denim
Speaker N: wow t- that's *impressive look at the hat too*
Speaker M: *I think the hat is denim too [laughs]*
Speaker N: I mean c'mon why not denim sunglasses **[laughs]**
Speaker M: **I know might as well**
Speaker N: missed opportunity

Another one of these six was "I think this really brings out your eyes" in Example 9:

(9) Speaker O: if you were required to recommend this outfit to a friend what would you say [laugh]
Speaker P: damn I would say yeah I think this really brings out your eyes

These examples show that sarcasm in spontaneous speech can work as intended, with both conversational participants and outside observers successfully identifying sarcastic utterances.

3.5.3 Discussion

The synchronous movement condition did not significantly increase the rate of agreements. Participants were given as much time as they required to review conversations and identify sarcastic content, reducing the possibility that their lack of agreement is due to an inability to complete the task accurately under constrained conditions. Although dyads might have provided higher rates of agreement if they had been asked to identify sarcastic utterances together, when conversational partners

rely on their own judgement to identify sarcasm, they generally do not agree on which statements are sarcastic.

Outside observers were better at identifying sarcasm, but still disagreed more than they agreed (315/580 = 54% of ratings were identified as sarcastic by one overhearer). Our result for spontaneously produced sarcasm in natural conversation differs from earlier results where outside observers coded sarcasm with great success. This includes sarcasm written by participants in experiments (Campbell & Katz, 2012; Huang et al., 2015), and verbal irony and satire written in movie reviews or news outlets, including satirical news outlets (Burgers et al., 2011; Reyes & Rosso, 2014). In another study, the items assessed were produced in spontaneously spoken settings, but they were selected for inclusion in the corpus based on the fact that an observer identified them as sarcastic, often talking to the producers about their productions (Eisterhold et al., 2006). In our dataset, these would be the items where overhearers overlapped with the speakers, which was about 28.6% of the speaker-identified sarcasms. See Table 1.

Inter-rater reliability as assessed by Krippendorff's alpha shows that reliability was better than chance for both groups of raters, suggesting that raters generally made a good faith effort to code accurately. Nonetheless, overhearer raters were slightly more reliable between each other than the interlocutors were. Although overhearers and interlocutors received identical instructions, it is possible that they were more motivated to code carefully, as they were research assistants working in the lab,

whereas the interlocutors were participants who may have budgeted less time to complete the task.

While a procedure to detect sarcasm may have improved our observers' agreement (cf. Burgers et al., 2011), it is not necessary. In two written sarcasm studies (cf. Campbell & Katz, 2012; Huang et al., 2015) and an addressee-verified sarcasm study (Eisterhold et al., 2006), researchers did not provide coders with a procedure for identifying sarcasm, yet observers were able to agree on what was sarcastic.

3.6 Study 3

In light of the extremely low level of agreement between raters, we performed several additional analyses to determine why our agreement was so much lower than other researchers have reported. Another way to measure agreement is by examining interlocutors' responses to speakers' sarcasm. Specifically, we thought that interlocutors' responses to sarcastic comments could give us a clue as to whether they understood it. Another method that has seemed to garner successful agreement is to split verbal irony into several more specific categories (e.g., Hancock, 2004; Gibbs, 2000). We therefore examined verbal irony in the conversations by binning into four categories (sarcasm, understatement, hyperbole, and rhetorical question), hoping this would lead to higher agreement. Further, an unanswered question from Gibbs (2000) is whether subtypes of verbal irony show agreement between separate coders, as Gibbs had his annotators work together to come to agreement (which was then

followed by Gibbs serving as a super-coder who verified the agreements these pairs of coders reached).

3.6.1 Method

Subsets of the data from Study 1 were analyzed by coders trained in two different ways. For the Evidence of Understanding analysis, four dyads from the synchronous condition and four dyads from the non-synchronous condition were randomly selected for analysis. Coders were given definitions of positive evidence of understanding, negative evidence of understanding, and no evidence of understanding taken from Hancock (2004):

> Finally, responses to ironic statements were analyzed to determine the addressee's comprehension of the speaker's ironic intent. Addressees could provide three types of evidence (Clark, 1996): (a) negative evidence, indicating that the speaker's ironic intent had been misinterpreted or required additional clarification for comprehension (e.g., "You really like that hat?"), (b) positive evidence, indicating comprehension of the ironic intent, either by acknowledging the ironic intent (e.g., "good one," a laugh, etc.) or by extending the initial irony with a subsequent ironic remark, or (c) no evidence, in which the addressee did not acknowledge or respond directly to the ironic statement. When the addressee provided no evidence (e.g., by changing the subject), the addressee's comprehension of irony could not be determined. (p. 453)

They were then given access to video recordings from Study 1, along with the list of timestamps at which each interlocutor reported sarcasm. This amounted to 179 reports of sarcasm total. Coders determined for each instance of sarcasm whether the addressee exhibited positive evidence (coded as 1), negative evidence (coded as -1), or no evidence (coded as 0) of comprehending the speaker's ironic intent. We then compared the way each of our coders rated the reports and compared the ratings of self-reported sarcasm to other-reported sarcasm.

For the Subtypes of Verbal Irony analysis, two coders analyzed four randomly selected transcripts, two from the synchronous condition and two from the non-synchronous condition. Coders were given the following definitions of four types of verbal irony, taken from Hancock (2004):

> (a) sarcasm, in which the speaker intended the pragmatic opposite of what was said in an effort to convey a negative attitude (e.g., "Matt Stone is looking just ravishing in his pink dress"), (b) understatement, in which the speaker stated less than was the case (e.g., "A little too much hairy cleavage for a formal event"), (c) hyperbole, in which the speaker exaggerated the situation (e.g., "The most vile thing known to man . . . hot dogs"), and (d) rhetorical questions, in which the speaker ostensibly asked a question in order to express an attitude but did not expect an answer (e.g., "What the hell was she thinking?") (p. 453)

Coders worked separately to watch the videos of the conversations and annotate separate copies of the transcripts. They did not communicate about the process once it began. Hancock (2004) did an analysis of these four sub-types of verbal irony, but it is unclear from the paper whether there was any measure of inter-rater agreement for this analysis, or whether the kappa reported refers to the subsequent analysis of irony cues. Finally, once coders were finished with their individual annotations, we combined all of their annotations into one spreadsheet and asked the same two coders to resolve their disagreements. This part of the process is identical to Gibbs' (2000) method, except that he used students from his class and included jocularity as a type of irony.

3.6.2 Results

In the Evidence of Understanding analyses, we found that coders generally agreed. In the Subtypes of Verbal Irony analyses, we found that coders generally did not agree.

3.6.2.1 Evidence of Understanding

We computed Cohen's kappa as a measure of inter-rater agreement on all 179 ratings by both coders (κ = .38, 95% CI [.25, .51]), on the 96 self-reported sarcastic reports (κ = .46, 95% CI [.29, .63]), and on the 83 other-reported sarcastic reports (κ = .29, 95% CI [.09, .50]). These signify fair to moderate agreement between coders. Next we examined the proportion of positive, negative, and no evidence across self- and other-reported instances of sarcasm, see Table 2.

Table 2. Subset of data (N = 8 dyads) assessed for proportion of positive, negative, and no evidence across self- and other-reported instances of sarcasm.

Self-Reported Sarcasm	**Positive Evidence**	**No Evidence**	**Negative Evidence**
Coder A	65.6%	28%	6.4%
Coder B	54.2%	35.4%	10.4%
Average	59.9%	31.8%	8.3%
Other-Reported Sarcasm	**Positive Evidence**	**No Evidence**	**Negative Evidence**
Coder A	72.3%	20.5%	7.2%
Coder B	55.4%	42.2%	2.4%
Average (%)	63.9%	31.3%	4.8%

3.6.2.2 Subtypes of Verbal Irony

The two coders turned in 73 reports of verbal irony. Sixteen of these reports matched perfectly (21.9%). Another 10 reports referred to the same statement but were classified as different forms of verbal irony. The other 47 utterances were only marked as ironic by one coder.

Table 3. Subset of data (N = 4 dyads) coded for subtypes of verbal irony.

	Rhetorical Q	**Hyperbole**	**Understatement**	**Sarcasm**
Coder A	20	6	9	7
Coder B	11	11	3	6
After agreement	18	7	7	6

After resolving their disagreements, the two coders reported higher levels of verbal irony than reported by Gibbs (2000). There were on average 9.5 utterances per 10-minute video that were reported to include verbal irony. Gibbs (2000) reported 4.7 per 10-minute conversation, suggesting that the badly dressed celebrities manipulation successfully elicited verbal irony (as also shown in Hancock, 2004). Coders consolidated their 73 independent reports into 42 combined reports across 753 turns, with the 38 distributed as in Table 3. There were 4 instances in which the coders' disagreements could not be resolved.

3.6.3 Discussion

The amount of combined positive and negative evidence across both raters was slightly lower than that reported by Hancock (2004). Hancock reported that 83.4% of the face-to-face interactions included some type of evidence, whereas our data showed only 68.4% — although this is still higher than Hancock's reported 58.5% for computer-mediated communication. In both Hancock's and our study, negative evidence is low. Hancock (2004) suggested that the low proportion of negative evidence meant that misunderstandings were low, but we believe this may be driven in part by participants' attempts to avoid awkward interactions: when mutual understanding deteriorates, it can be helpful to not direct attention to the misunderstanding. This phenomenon may be especially prevalent when interlocutors believe that mutual understanding will not be important or relevant in the long term (e.g., in a one-time 10-minute conversation with a stranger). Similarly, the high rates of positive evidence do not necessarily mean that agreement on sarcasm is high. People may have nodded or smiled even though they didn't recognize the sarcasm. That is, people can provide evidence of understanding even when they do not fully understand.

In further support of this, we note that our two coders attained only modest inter-rater agreement despite reporting similar rates of positive, negative, and no evidence. Unfortunately, Hancock (2004) does not provide inter-rater agreement for this measurement, so we are unable to compare. Our results could imply that coder training was insufficient, but it is more likely that the reason for low agreement was

that the task was quite challenging. Prior to analyzing our coders' data, we asked each of our coders to report on their experience of coding, explicitly asking them to discuss how the process was easy, hard, or surprising. One coder remarked that it was unexpected how little participants' reports of sarcasm overlapped, while noting that his own understanding of sarcasm also frequently conflicted with the conversational participants' reports. This made coding difficult, because if the coder is unable to understand the sarcasm, it is harder to determine whether the addressee is providing positive, negative, or no evidence. He wrote,

> "...In these scenarios, if I'm as sure as I can be, I code literal responses to statements I consider to have any hint of humor in them as a negative understanding and literal responses to literal statements as not enough evidence for a negative or positive understanding. I did not feel confident taking an obviously literal statement and literal response as a positive or negative sarcastic understanding because it was not sarcastic… I'm finding myself doubting the participants' motivations when they report the sarcasm times because many times it seems to make no sense when they are indicating there was a sarcastic remark, but I could be wrong."

This report, combined with the fact that our three overhearer annotators also achieved a relatively low agreement, suggests that despite good-faith efforts (at least on the part of our trained coders), no one had very good agreement about what constituted sarcasm during the conversations.

Regarding subtypes of irony, we should begin by stating that we believe that rhetorical questions, hyperbole, understatement, and sarcasm are useful constructs for researchers to study. However, our trained coders had difficulty agreeing on these constructs in our corpus of spontaneous productions. In addition, the coding process

— whether coded independently or in pairs — affected the categories items were binned into.

3.7 General Discussion

Conversational participants identified their own and their partners' sarcasm use immediately after producing a conversation. We observed a sarcasm reciprocity illusion: people reported more sarcasm in their addressees when they reported more sarcasm in themselves. We also observed high levels of two related problems in conveying sarcasm: People produced sarcasm but addressees and overhearers didn't understand it as sarcasm (false negatives, from the listener's perspective), and people produced non-sarcastic statements that addressees and overhearers believed were sarcastic (false positives, from the listener's perspective). Both of these phenomena were far more common than expected. Although pairs identified sarcasm in their conversations about once per minute, most of the sarcasm produced was not identified as sarcastic by addressees.

Synchronous movement increased rates of self-reported, but not other-identified, sarcasm use. That is, synchronous movement increased participants' reports of sarcasm in themselves, but it did not increase reports of sarcasm in addressees. Synchronous movement also did not make identifying sarcasm in addressees more accurate.

Unlike many previous studies where sarcasm identification was done by people who were not in the original conversation, in the current studies sarcasm identification is defined as a match between the speaker's own identification of their

sarcastic productions and their addressee's or an overhearer's identification. This definition sidesteps the thorny debates about what is technically sarcasm as opposed to verbal irony (e.g. Attardo et al., 2003; Caucci & Kreuz, 2012). People believe they know what sarcasm is, and when they use it, they believe they are communicating effectively with addressees (Chin, 2011). But before the studies reported here, how accurately people identified sarcasm in spontaneously produced, natural conversation with strangers was unknown. We show that people generally perform poorly.

One possibility for the low performance we observed is that because people generally use cues with friends (Caucci & Kreuz, 2012), we can anticipate that our study with strangers may not have included enough cues for accurate identification. However, we could argue just the opposite: that the hard part of doing stranger studies on a college campus is that most people consider themselves to be relatively friendly with each other because they are about the same age, doing similar things (such as attending college), and living in the same town. People who are part of similar social groups are more likely to agree on sarcastic intent than people who are completely unaffiliated (Kellner & Schober, 2018). So while they were strangers, they were likely closer to being on friendly terms than other possible addressee pairings in a non-campus study.

Another possibility for the low performance is that we are measuring performance on a different task than those that have been assessed before. Our participants engaged in naturalistic conversation with no guidance other than to talk for ten minutes. Participants could talk about any aspect of the photos that they

pleased. They were not asked to take on a specific conversational role while discussing the photos (e.g., to communicate "as if they were providing commentary for a fashion show," Hancock, 2004, p. 452), or to annotate written language (Burgers et al., 2011), which may contain more verbal irony anyway (Hancock, 2004). In other words, it may be that in spontaneous, natural conversation, verbal irony takes a form that is much more difficult to reliably annotate.

It is worth noting that Gibbs (2000) also studied naturalistic conversations and then annotated for the presence of verbal irony. While the study is similar to ours at face value, it differs in important ways, most notably by defining and annotating verbal irony in a way that may have eliminated some cases that we include. Gibbs asked students in his class to record conversations between friends and roommates and reported 289 instances of irony across 62 conversations. But there are three reasons to believe that this number is a poor estimate of what interlocutors actually thought was ironic. One is that the students' transcriptions varied broadly in quality, as Gibbs noted, which would make recognition of verbal irony more variable. Another is that we don't know how many utterances were originally marked by raters as containing irony, only that there were 314 instances once two raters had come to a consensus, and 289 once Gibbs reviewed annotations. There were an unknown number of utterances marked as ironic prior to collaborating with a second coder (and if our similar analysis is any indication, only about 22% of these would match). Third, Gibbs reviewed the transcripts and the 314 utterances that students marked as ironic but did not report finding even one ironic utterance that his raters had missed,

just 25 that he did not agree with. Because recognition of verbal irony was bootstrapped from one rater to the next in this way, it is difficult to determine what agreement was actually achieved. Based on our experience asking people to code just four, or even one type of verbal irony (rather than five), we find it difficult to believe that similar students would achieve acceptable inter-rater agreement if the study was replicated without the bootstrapped annotation procedure, unless there was a specialized training process that wasn't reported.

If people are judging sarcasm based on the rules researchers have observed (see Utsumi, 2000, for a review of theories of irony), there should be more consensus among participants in a dyad. The low rates of agreement suggest that researchers' rules are different from the internal rules that people use while producing and interpreting sarcastic comments.

We can discount two alternative explanations for the discoveries reported here. The findings are not a result of synchronous movements' increasing participants' comfort with each other: Comfort ratings after the verbal activity were the same across conditions. The findings are also unlikely to be a result of synchronous movements' prompting participants to focus their attention (Valdesolo et al., 2010), allowing for greater attention to sarcastic cues: People do not actually appear to be picking up on each other's cues to sarcasm, failing to identify 80% of them. A third alternative explanation is that the findings result from rapport-building (Bernieri, 1988; Wheatley, Kang, Parkinson, & Looser, 2012); this is possible —

although positive emotions are not necessary for synchronous movement to promote cooperation (Wiltermuth & Heath, 2009).

Our results contrast with other findings that the rate of misunderstanding sarcasm is very low, where in fewer than 5% of conversational turns did participants indicate they had not understood the irony when judged by outside observers (Hancock, 2004). But negative evidence does not necessarily correspond to a low rate of misunderstanding, because people might be choosing not to signal their misunderstanding. In addition, by definition, sarchasms occur when people do not understand that they've missed something. So the rate of sarchasms is not assessable through negative evidence — people experiencing sarchasm are unable to identify that they didn't understand the irony.

3.8 Conclusions

Observing sarcasm in the wild is difficult (Eisterhold et al., 2006). Although sarcasm can be produced reliably in the lab (39 out of 40 dyads produced at least one ironic turn in Hancock, 2004), sometimes lab elicitation is less successful (18 out of 29 dyads produced at least one sarcastic utterance in Caucci & Kreuz, 2012). To test how people use spontaneously generated sarcasm, we introduced a novel technique for heightening sarcasm production in the lab: a synchronous movement sequence preceding a sarcasm-inducing task, with conversational prompts to extend the conversation if necessary. With this procedure, we increased self-reported sarcasm by 51% (on average, 4.21 times per conversation without synchronous movement compared to 6.36 times per conversation with synchronous movement). Studying

spontaneous sarcasm in a laboratory requires reliably creating spontaneous sarcasm in a laboratory, and synchronous movement can help researchers increase sarcasm production.

While we were successful at fostering sarcasm as identified by the speakers themselves, we also discovered substantial mismatches between speakers' and addressees' identifications of spontaneously spoken sarcasm, as well as the sarcasm outside observers identified. People generally did not agree on what was sarcastic. What appears to be special in conversation are those few cases where conversational participants do agree. These cases may be especially creative, humorous, or enjoyable. Perhaps working towards that goal makes all the misses worthwhile.

3.9 References

Albert, S., & De Ruiter, J. P. (2018). Repair: The interface between interaction and cognition. *Topics in Cognitive Science*, *10*(2), 279-313. https://doi.org/10.1111/tops.12339

Amati, F. & Brennan, S. E. (2016, November). Those Little White Lies: Deception and Politeness in Spontaneous Conversation. Poster presented at the annual meeting of the Psychonomic Society, Boston, MA.

Andrews, G., & Ingham, R. J. (1971). Stuttering: Considerations in the evaluation of treatment. *British Journal of Disorders of Communication*, *6*(2), 129-138. https://doi.org/10.3109/13682827109011538

Attardo, S., Eisterhold, J., Hay, J., & Poggi, I. (2003). Multimodal markers of irony and sarcasm. Humor: *International Journal of Humor Research*, *16*(2), 243–260. https://doi.org/http://dx.doi.org/10.1515/humr.2003.012

BBC (2019, January 11). YouGov survey: British sarcasm 'lost on Americans.' Retrieved from https://www.bbc.com/news/world-us-canada-46846467

Bernieri, F. J. (1988). Coordinated movement and rapport in teacher-student interactions. *Journal of Nonverbal Behavior*, *12*, 120–138. https://doi.org/10.1007/BF00986930

Brown, P. and Levinson, S. (1978). Universals in language usage: Politeness phenomena. In E. N. Goody (Ed.) *Questions and Politeness: Strategies in Social Interaction.* Cambridge: Cambridge University Press.

Bryant, G. A. (2010). Prosodic contrasts in ironic speech. *Discourse Processes*, *47*(7), 545–566. https://doi.org/10.1080/01638530903531972

Bryant, G. A. & Fox Tree, J. E. (2002). Recognizing verbal irony in spontaneous speech. *Metaphor and Symbol, 17*(2), 99-117. https://doi.org/10.1207/S15327868MS1702_2

Bryant, G. A., & Fox Tree, J. E. (2005). Is there an ironic tone of voice? *Language and Speech*, *48*(3), 257–277. https://doi.org/10.1177/00238309050480030101

Burgers, C., Van Mulken, M., & Schellens, P. J. (2011). Finding irony: An introduction of the verbal irony procedure (VIP). *Metaphor and Symbol*, *26*(3), 186-205. https://doi.org/10.1080/10926488.2011.583194

Burgers, C., van Mulken, M., & Schellens, P. J. (2012). Type of evaluation and marking of irony: The role of perceived complexity and comprehension. *Journal of Pragmatics*, *44*(3), 231-242. https://doi.org/10.1016/j.pragma.2011.11.003

Campbell, J. D. & Katz, A. N. (2012) Are there necessary conditions for inducing a sense of sarcastic irony?, *Discourse Processes*, *49*(6), 459-480. https://doi.org/10.1080/0163853X.2012.687863

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

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception–behavior link and social interaction. *Journal of Personality and Social*

Psychology: Attitudes and Social Cognition, *76*(6), 893–910. https://doi.org/http://dx.doi.org/10.1037/0022-3514.76.6.893

Chin, R. (2011, November 14). The Science of Sarcasm? Yeah, Right. *Smithsonian.com.* Retrieved from https://www.smithsonianmag.com/science-nature/the-science-of-sarcasm-yeah-right-25038/

Christenfeld, N., & Creager, B. (1996). Anxiety, alcohol, aphasia, and ums. *Journal of Personality and Social Psychology*, 70(3), 451–460. http://dx.doi.org.oca.ucsc.edu/10.1037/0022-3514.70.3.451

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

Cutler, A. (1976). Beyond parsing and lexical look-up: An enriched description of auditory sentence comprehension. In R. J. Wales & E. Walker (Eds.), *New Approaches to Language Mechanisms: A Collection of Psycholinguistic Studies* (pp. 133-149). Amsterdam: North-Holland.

D'Arcey, J. T., Oraby, S., and Fox Tree, J. E. (2019). Wait signals predict sarcasm in online debates. *Dialogue & Discourse*, *10*(2) 56-78.

Eisterhold, J., Attardo, S., & Boxer, D. (2006). Reactions to irony in discourse: Evidence for the least disruption principle. *Journal of Pragmatics*, *38*(8), 1239-1256. https://doi.org/10.1016/j.pragma.2004.12.003

Fox Tree, J. E. (1999). Listening in on monologues and dialogues. *Discourse Processes, 27,*

35-53. https://doi.org/10.1080/01638539909545049

Fox Tree, J. E. & Clark, N. B. (2013). Communicative effectiveness of written versus spoken feedback. *Discourse Processes*, *50*(5), 339-359. https://doi.org/10.1080/0163853X.2013.797241

Fox Tree, J. E. & Mayer, S. A. (2008). Overhearing single and multiple perspectives. *Discourse Processes, 45,* 160-179. https://doi.org/10.1080/01638530701792867

Fussell, S. R., & Krauss, R. M. (1989). Understanding friends and strangers: The effects of audience design on message comprehension. *European Journal of Social Psychology*, *19*(6), 509-525. https://doi.org/10.1002/ejsp.2420190603

Gibbs, R. W. (2000). Irony in talk among friends. Metaphor and symbol, 15(1-2), 5-27. https://doi.org/10.1080/10926488.2000.9678862

Gibbs, R. W. (2002). A new look at literal meaning in understanding what is said and implicated. *Journal of pragmatics, 34*(4), 457-486. https://doi.org/10.1016/S0378-2166(01)00046-7

Giles, H. (2008). Communication accommodation theory: "When in Rome …" or not! In L. Baxter & D. Braithwaite (Eds.), *Engaging Theories in Interpersonal Communication: Multiple Perspectives* (pp. 161–174). Thousand Oaks, California, United States: SAGE Publications, Inc. https://doi.org/10.4135/9781483329529.n12

Giles, H. (2009). Accommodation theory. In S. Littlejohn & K. Foss, (Eds.), *Encyclopedia of Communication Theory.* Thousand Oaks, California, United States: SAGE Publications, Inc. https://doi.org/10.4135/9781412959384.n1

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

Grice, H. P. (1978). Further notes on logic and conversation. In P. Cole (Ed.), *Syntax and Semantics: Vol. 9. Pragmatics*. New York: Academic Press.

Hancock, J. T. (2004). Verbal irony use in face-to-face and computer-mediated conversations. *Journal of Language and Social Psychology*, *23*(4), 447–463. https://doi.org/10.1177/0261927X04269587

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. Communication methods and measures, 1(1), 77-89. https://doi.org/10.1080/19312450709336664

Healey, P. G., Mills, G. J., Eshghi, A., & Howes, C. (2018). Running repairs: Coordinating meaning in dialogue. Topics in Cognitive Science, *10*(2), 367-388. https://doi.org/10.1111/tops.12336

Huang, L., Gino, F., & Galinsky, A. D. (2015). The highest form of intelligence: Sarcasm increases creativity for both expressers and recipients. *Organizational Behavior and Human Decision Processes*, *131*, 162-177. https://doi.org/10.1016/j.obhdp.2015.07.001

Hove, M. J., & Risen, J. L. (2009). It's all in the timing: Interpersonal synchrony increases affiliation. *Social Cognition*, *27*, 949–961. https://doi.org/10.1521/soco.2009.27.6.949

Hussey, K. A., & Katz, A. N. (2006). Metaphor production in online conversation: Gender and friendship status. *Discourse Processes*, *42*(1), 75–98. https://doi.org/10.1207/s15326950dp4201_3

Ivanko, S. L., Pexman, P. M., & Olineck, K. M. (2004). How sarcastic are you? Individual differences and verbal irony. *Journal of Language and Social Psychology*, *23*(3), 244-271. https://doi.org/10.1177/0261927X04266809

Kelly, L., & Miller-Ott, A. E. (2018). Perceived Miscommunication in Friends' and Romantic Partners' Texted Conversations. *Southern Communication Journal*, *83*(4), 267-280. https://doi.org/10.1080/1041794X.2018.1488271

Kellner, C., & Schober, F., (2018). Misunderstanding authors' humor and sarcasm in an online fashion forum. *59th Annual Meeting of the Psychonomic Society, New Orleans, Louisiana.*

Kowal, S., Wiese, R., & O'Connell, D. C. (1983). The use of time in storytelling. *Language and Speech*, *26*(4), 377-392. https://doi.org/10.1177%2F002383098302600405

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. https://doi.org/http://dx.doi.org/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. https://doi.org/10.1207/s15327868ms1001_3

Kruger, J., Epley, N., Parker, J., & Ng, Z. W. (2005). Egocentrism over e-mail: Can we communicate as well as we think?. *Journal of Personality and Social Psychology*, *89*(6), 925. https://psycnet.apa.org/doi/10.1037/0022-3514.89.6.925

Lee, C. J., & Katz, A. N. (1998). The differential role of ridicule in sarcasm and irony. *Metaphor and Symbol*, *13*(1), 1–15. https://doi.org/10.1207/s15327868ms1301_1

Muñoz, V., Lavega, P., Serna, J., Ocáriz, U. S. de, & March, J. (2016). Mood states when playing alone or in cooperation: two unequal motor and affective experiences. *Anales de Psicología / Annals of Psychology*, *33*(1), 196–203. https://doi.org/10.6018/analesps.33.1.233301

Macrae, C. N., Duffy, O. K., Miles, L. K., & Lawrence, J. (2008). A case of hand waving: Action synchrony and person perception. *Cognition*, *109*(1), 152–156. https://doi.org/10.1016/j.cognition.2008.07.007

Miles, L. K., Nind, L. K., & Macrae, C. N. (2009). The rhythm of rapport: Interpersonal synchrony and social perception. *Journal of Experimental Social Psychology*, *45*(3), 585-589. https://doi.org/10.1016/j.jesp.2009.02.002

Nilsen, E. S., Glenwright, M., & Huyder, V. (2011). Children and adults understand that verbal irony interpretation depends on listener knowledge. *Journal of*

Cognition and Development, *12*(3), 374-409. https://doi.org/10.1080/15248372.2010.544693

Oraby, S., Harrison, V., Reed, L., Hernandez, E., Riloff, E., & Walker, M. (2016, September). Creating and Characterizing a Diverse Corpus of Sarcasm in Dialogue. In 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue.

Peterson, C. C., Wellman, H. M., & Slaughter, V. (2012). The mind behind the message: Advancing theory-of-mind scales for typically developing children, and those with deafness, autism, or Asperger syndrome. *Child Development*, *83*(2), 469-485. https://doi.org/10.1111/j.1467-8624.2011.01728.x

Pexman, P. M., & Olineck, K. M. (2002). Does sarcasm always sting? Investigating the impact of ironic insults and ironic compliments. *Discourse Processes*, *33*(3), 199–217. https://doi.org/10.1207/S15326950DP3303_1

Pexman, P. M. (2008). It's fascinating research: The cognition of verbal irony. *Current Directions in Psychological Science, 17*(4), 286-290. https://doi.org/10.1111/j.1467-8721.2008.00591.x

Reyes Pérez, A.; Rosso, P. (2014). On the difficulty of automatically detecting irony: beyond a simple case of negation. Knowledge and Information Systems. 40(3):595-614. https://doi.org/10.1007/s10115-013-0652-8

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

Rockwell, P. (2003). Empathy and the expression and recognition of sarcasm by close relations or strangers. *Perceptual and Motor Skills*, *97*(1), 251–256. https://doi.org/10.2466/pms.2003.97.1.251

Rockwell, P., & Theriot, E. M. (2001). Culture, gender, and gender mix in encoders of sarcasm: A self-assessment analysis. *Communication Research Reports*, *18*(1), 44–52. https://doi.org/10.1080/08824090109384781

Savitsky, K., Keysar, B., Epley, N., Carter, T., & Swanson, A. (2011). The closeness-communication bias: Increased egocentrism among friends versus strangers. *Journal of Experimental Social Psychology*, *47*(1), 269–273. https://doi.org/10.1016/j.jesp.2010.09.005

Schober, M. F., Suessbrick, A. L., & Conrad, F. G. (2018). When do misunderstandings matter? evidence from survey interviews about smoking. *Topics in Cognitive Science*, *10*(2), 452-484. https://doi.org/10.1111/tops.12330

Semin, G. R. (2007). Grounding communication: Synchrony. In A. W. Kruglanski & E. T. Higgins (Eds.), *Social Psychology: Handbook of Basic Principles* (2nd Edition), pp. 630–649. New York, NY: The Guilford Press.

Segrin, C., & Flora, J. (1998). Depression and verbal behavior in conversations with friends and strangers. *Journal of Language and Social Psychology*, *17*, 492-503. https://doi.org/10.1177/0261927X980174005

Slugoski, B. R., & Turnbull, W. (1988). Cruel to be kind and kind to be cruel: Sarcasm, banter and social relations. *Journal of Language and Social Psychology*, *7*(2), 101–121. https://doi.org/10.1177/0261927X8800700202

Smith, M. (2019, January 11). British subtext: Half of Americans wouldn't be able to tell that a Briton is calling them an idiot. Retrieved from https://yougov.co.uk/topics/lifestyle/articles-reports/2019/01/11/half-americans-wouldnt-be-able-tell-british-person

Tolins, J., Namiranian, N., Akhtar, N., Fox Tree, J. E. (2017). The role of addressee backchannels and conversational grounding in vicarious word learning in four-year-olds. *First Language*, 37(6) 648-671. https://doi.org/10.1177%2F0142723717727407

Tolins, J., Zeamer, C., & Fox Tree, J. E. (2018). Overhearing dialogues and monologues: How does entrainment lead to more comprehensible referring expressions? *Discourse Processes*, *55*(7), 545-565. https://doi.org/10.1080/0163853X.2017.1279516

Wilson, D., & Sperber, D. (1999). *Relevance and Relevance Theory. In MIT Encyclopedia of the Cognitive Sciences* (pp. 719–722). MIT Press.

Wu, S., & Keysar, B. (2007). The effect of information overlap on communication effectiveness. *Cognitive Science*, *31*(1), 169-181. https://doi.org/10.1080/03640210709336989

Valdesolo, P., & DeSteno, D. (2011). Synchrony and the social tuning of compassion. *Emotion, 11*(2), 262–266. https://doi.org/10.1037/a0021302

Valdesolo, P., Ouyang, J., & DeSteno, D. (2010). The rhythm of joint action: Synchrony promotes cooperative ability. *Journal of Experimental Social Psychology, 46*(4), 693–695. https://doi.org/10.1016/j.jesp.2010.03.004

van Baaren, R. B., Holland, R. W., Kawakami, K., & van Knippenberg, A. (2004). Mimicry and prosocial behavior. *Psychological Science*, *15*(1), 71–74. https://doi.org/10.1111/j.0963-7214.2004.01501012.x

Williams, J. A., Burns, E. L., & Harmon, E. A. (2009). Insincere utterances and gaze: Eye contact during sarcastic statements. *Perceptual and Motor Skills*, *108*(2), 565–572. https://doi.org/10.2466/pms.108.2.565-572

Wiltermuth, S. S., & Heath, C. (2009). Synchrony and cooperation. *Psychological Science*, *20*(1), 1–5. https://doi.org/10.1111/j.1467-9280.2008.02253.x

Wheatley, T., Kang, O., Parkinson, C., & Looser, C. E. (2012). From mind perception to mental connection: Synchrony as a mechanism for social understanding. *Social and Personality Psychology Compass*, *6*(8), 589–606. https://doi.org/10.1111/j.1751-9004.2012.00450.x

Witte, T. (1998, August 2). Reported in *The Stool Invitational Week 281: Calculate the Odds*. Retrieved from http://www.washingtonpost.com/wp-srv/style/invitational/invit980802.htm

3.10 Appendix A: Optional Conversational Prompts

- Would you wear this outfit to a fancy restaurant?
- If you were required to recommend this outfit to a friend, what would you say?
- Would you wear this outfit?
- Would you wear this outfit for $5? $50? $500? $5000?
- Please describe the personality of someone who would wear this outfit.
- Under what circumstances would you wear this outfit?
- Could you be convinced to wear this outfit if someone paid you money? How much money would you require?
- If you could meet the designer of this outfit, what would you say to them?
- Do you know anyone who would wear this outfit? How do you feel about this person? Would this person look good in this outfit?

3.11 Appendix B: Post-Experiment Questionnaire

(Note: Percentages represent all participants, including participants who were later excluded.)

What is your gender identity?

- Male (31.4%)
- Female (67.6%)
- Non-binary (1%)

What do you predict is the gender identity of your partner?

- Male (33.3%)
- Female (61.8%)
- Non-binary (4.9%)

What is your age in years?

- Free response (M = 20.04, SD = 2.59)

Have you ever seen your partner before?

- Yes (13.7%)
- No (81.4%)
- I don't remember (4.9%)

Did you and your partner introduce yourselves to each other at any point during the experiment?

- Yes (24.8%)
- No (75.2%)

Did you consider yourself friends with your partner before you participated in this study?

- Yes (5.2%)
- Maybe (4.1%)
- No (90.7%)

Did you consider yourself friends with your partner after you participated in this study?

- Yes (6.9%)
- Maybe (49.5%)
- No (43.6%)

How comfortable were you with your partner after the movement activity?

- 1-7 scale, extremely uncomfortable (1) to extremely comfortable (7)

How comfortable did you feel with your partner after engaging in a conversation about badly dressed celebrities?

- 1-7 scale, extremely uncomfortable (1) to extremely comfortable (7)

Did the movement activity (throwing the imaginary ball) make you feel more or less comfortable with your partner?

- 1-7 scale, extremely uncomfortable (1) to extremely comfortable (7)

To what extent do you consider yourself a sarcastic person?

- 1-7 scale, extremely not sarcastic (1) to extremely sarcastic (7)

How often do you use sarcasm in a given day? (ex. 0 - 100)

- (Free response)

Do you have any comments about this experiment you would like to share?

- (Free response)

4 Chapter 3: Oh, SO Sarcastic: Diverse Strategies for Being Sarcastic

4.0 Pre-introduction

In this chapter I present results from a paper under review, Oh, SO Sarcastic (D'Arcey & Fox Tree, under review), in which we consider the tradeoffs that researchers have made in their definitions of sarcasm. We argue that highly abstract theories of sarcasm may not generalize well to a typical person's concept of sarcasm, and that the term is used broadly enough to apply in many situations where abstract theories would not. To express this belief with data, we collected undergraduates' own ideas about what constitutes sarcasm after being asked to create some of it themselves. Through this data we identify over a dozen concepts that reliably connect to sarcasm in at least one way.

4.1 Abstract

Sarcasm has been defined in a plethora of different ways, but too often the definitions hinge on researchers' own perceptions of what constitutes sarcasm or verbal irony, and not enough on the perceptions of people producing the sarcastic content. We asked people to transform internet forum posts to make them sarcastic without providing information about what sarcasm is. Participants then critically examined their creations. People identified a variety of strategies that they use to communicate sarcasm in writing. A content analysis of written productions confirmed the use of these strategies, several of which were more likely to be present alongside sarcasm. Results are useful for understanding sarcasm production, comprehension, and linguistic rapport.

4.2 Introduction

Even among a relatively homogeneous population, there is immense diversity in people's conceptions of sarcasm. People use more or less of it, and when they use it, they don't always communicate it well to each other (Fox Tree et al., 2020). In this report, we asked people to transform forum posts to make them sarcastic without providing information about what sarcasm is. Their creations, and their reports about the experience of creating sarcasm, show that there are elements that can be reliably used to indicate sarcasm, and that producers' sense of success in creating sarcastic productions is matched by, but distinct from comprehender's success in interpreting.

We begin with a review of what people might be thinking of when they think of sarcasm. Because there is so much disagreement about sarcasm, it is instructive to begin by illustrating this disagreement through more (and less) authoritative sources, like dictionary entries. The goal in doing this is not to agree on the term so much as to show that it is difficult or perhaps impossible to do so. We then describe our sarcasm-creation experiment. We asked people to transform internet forum posts to make them sarcastic without providing information about what sarcasm is. We then describe what people did when they attempted these transformations, including their own reports of strategies to create sarcasm. Finally, we examine the ability of several commonly reported strategies to predict the presence of sarcasm, both in participants' own eyes and in third party raters.

4.2.1 Dictionary Definitions of Sarcasm

Although Urban Dictionary is among the more colloquial sources from which to define a word, its community-driven nature may make it more reflective of used definitions than authoritative sources. Some highly upvoted definitions of sarcasm are "The ability to insult idiots without them realizing it" and "the bastard stepchild of irony" (Urbandictionary.com, 2020). More traditional definitions may feel more familiar or precise: "The use of irony to mock or convey contempt" (OED), but also leave substantial ambiguity in the concept. OED's definition does not speak to UrbanDictionary's suggestion that sarcasm frequently requires a lack of comprehension by the target ("without them realizing it"). And although Collinsdictionary.com also uses irony in their definition ("A taunting, sneering, cutting, or caustic remark; gibe or jeer, generally ironic"), their definition does not mandate irony in the same way that Oxford's does. Neither does Merriam-Webster's definition require irony: "A sharp and often satirical or ironic utterance designed to cut or give pain" (Merriam-Webster, 2020).

Older dictionaries provide some insight as well: Webster's 1812 dictionary defines sarcasm as, "A keen reproachful expression; a satirical remark or expression, uttered with some degree of scorn or contempt; a taunt; a gibe." This definition suggests that sarcasm is generally a witty phenomenon, as Oscar Wilde is well-known for pointing out. Literaryterms.net writes that sarcasm is "really more a tone of voice than a rhetorical device," yet it is unclear what role tone of voice plays in everyday sarcasm (Bryant & Fox Tree, 2002, 2005).

Our own field of psychology has much to say as well. Lazarus's Psychology Today blog post remarking that, "Sarcasm is actually hostility disguised as humor" notes a commonality among most of the definitions above: that sarcasm conveys contempt toward *something*, though not necessarily the addressee. Within the subdiscipline of sarcasm research itself, there are various theories about what constitutes sarcasm, possibly as a type of verbal irony. Two of the most well-known are the echoic mention theory (which suggests that sarcastic speakers are referencing some mutually understood opinion in a negative light) and pretense theory (which posits that sarcastic speakers are taking on a contrastive persona in order to mock it). Yet neither of these theories reflects the diversity and creativity of sarcasm as it is naturally used, and none remark on sarcasm's ability to draw speaker and addressee together (example taken from a corpus described by Fox Tree et al., 2020):

(1) A pair of participants looking at a photo of Brad Pitt and Britney Spears clad in denim outfits:
A: "I mean c'mon, why not denim sunglasses?"
B: "I know, right?"
A: "Missed opportunity."

Laypeople's definitions have a ring of truth that extends beyond researchers' definitions.

4.2.2 Researchers' Definitions of Sarcasm

One of the principle difficulties of studying sarcasm is defining what sarcasm is. Sarcasm is understood differently in the echoic mention theory (Sperber & Wilson, 1981) and the pretense theory of verbal irony (Clark & Gerrig, 1984). In the echoic mention theory, ironic utterances *mention* a meaning rather than *use* it, in the same

way that people mention or use a word. The distinction is subtle -- *mentioning* is saying something to call attention to the concept itself (e.g., "but what is heat, really?"), whereas *using* is saying something to convey its meaning (e.g., "I can't take this heat"). In pretense theory ironic utterances involve an act on the part of one or more of the interlocutors to portray an entity or group in a derogatory way. When defined in such distinct ways, the methods that researchers select to identify sarcasm differ enough that results will differ as well.

Another difficulty is defining what sarcasm is not. In the following constructed example, if the conversational participants are eager to get home to cook, the response to the question comes off as sarcastic:

(2) A: Oh hey, should we stop for some fast food?
B: Yes, let's do that because we got all these groceries.

But the response could be interpreted non-sarcastically if the interlocutors were tired from shopping. Recognizing this exchange (and many naturalistic exchanges) as either sarcastic or non-sarcastic requires more information than just the words.

Some researchers work towards a more scientifically testable definition by creating a definition that connects well to multiple definitions. For example, the Verbal Irony Procedure describes a logical process for determining whether a passage contains verbal irony. In each step the researchers' process connects to one or more definitions of irony, and their definition stays relatively broad: "an utterance with a literal evaluation that is implicitly contrary to its intended evaluation" (Burgers, van Mulken, & Schellens, 2005, p. 190). These approaches are important because they

attempt to bring together definitions of irony and sarcasm while also driving forward empirical testing of the combined definition.

4.2.3 Computational Identification of Sarcasm

Researchers attempting to computationally derive sarcasm from text have looked directly at what kinds of lexical patterns (symbols, words, and phrases) are likely to co-occur with sarcasm. The usual approach is to have a wide variety of raters determine whether or not words, phrases, or passages are sarcastic. Interjections like *gee* and *gosh*, for example, predict the presence of sarcastic content in novels (Kreuz & Caucci 2007), as do patterns like *wow*, *oh really*, and *I love it when* in internet arguments (Oraby et al., 2012). Writing words like *um* or *uh* and using punctuation like ellipses and quotes has also been associated with sarcasm (D'Arcey et al., 2019). From the presence of these patterns, other commonly collocated patterns can be derived from large corpora. For instance, if the word *ugh* co-occurred with *wow*, *gee*, and *I love it when*, it is possible that *ugh* would also predict the presence of sarcasm (for a typical example of this approach, see Qadir & Riloff 2014).

There are also other approaches to teaching computers to recognize sarcasm that have gained traction recently: deep learning algorithms, for one, are starting to be effective at recognizing emotions in text (Felbo et al., 2017). Deep learning methods attempt to train a virtual prediction network, usually containing hundreds or thousands of nodes in multiple layers, to classify texts into categories that match predetermined categories. These methods may change the direction of future sarcasm classifiers.

4.2.4 Current Study: Creating Sarcasm

Despite the difficulties of defining it, sarcasm is frequently produced in everyday communication. In an experimental study of sarcasm produced while describing badly dressed celebrities, pairs of communicators self-identified a sarcastic production about once a minute (Fox Tree et al., 2020). Even though sarcasm is often missed in spontaneous communication, conversational participants may risk producing it because of the joy that can be had from successful production and comprehension (Fox Tree et al., 2020).

To better understand what people are doing when they create sarcastic communication, we asked people to modify non-sarcastic utterances to make them sarcastic and then critically examine their creations. We did not tell participants in advance what we meant by the word "sarcasm," a method adopted by many prior researchers in light of definitional difficulties (e.g. Attardo 2003; Fox Tree et al. 2020; Kreuz & Caucci 2007; Rockwell 2000).

To assess sarcasm creativity, we developed a novel task where participants rewrote forum posts from internet arguments to make them sarcastic. The posts were selected because they were rated as not being sarcastic. By asking people to make non-sarcastic writing sarcastic (and not supplying a definition of sarcasm), we accomplished two goals: First, we got a set of sarcastic rewrites that can be contrasted with their original, non-sarcastic versions. Second, we primed our participants to have recent experience using and thinking about sarcasm: We had them create it, and then asked them about that process.

Because understanding sarcasm (and probably creating it) is an interactive process between the textual content and the reader's schemata (Boon, 2005), it was important to give our raters and participants some context for the original versions of the internet posts. For this reason, we limited our stimuli to only posts that directly quoted other posts. In this way, we gathered a set of pairs of posts, one of which was responding to the other, in which it was always clear that the latter post was a direct response to the former post. For example:

(3) **Post**: I have to ask you, did you do bbq ribs in your restaurant? And were they dry southern ribs?
Response: Yes, I did ribs but I like 'um juicy. It's all about the sauce baby.

Hereafter we refer to these items as post-response pairs.

We asked participants to rewrite only the responses, not the original posts, because we only got sarcasm ratings on the responses (not the posts). After rewriting each response, participants were asked to rate the difficulty of making the response sarcastic and their perceived success at doing so. We wanted to examine the difficulty of rewriting text to make it sarcastic. In most conditions, people use sarcasm in a carefree, easy way. But in our procedure, participants were tasked with interpreting another writer's meaning and then modifying their statements to change that sentiment. For this reason, we hypothesized that participants would find our task to be more challenging than they anticipated.

We also wanted to understand people's beliefs about sarcasm, so we asked them about their experiences creating it. We used these introspective self-reports to examine people's conceptions of sarcasm more closely, first by looking for similar

ideas between participants and then testing to see whether those ideas could be made into reliable concepts by training raters to recognize them.

4.3 Method

Participants rewrote responses in post-response pairs to make them sarcastic. They also assessed the quality of their creations.

4.3.1 Participants

Participants were 82 undergraduate psychology students at the University of California, Santa Cruz, who received course credit for their participation.

4.3.2 Materials

Twenty-four internet posts from the Internet Argument Corpus (Walker et al., 2012) and their responses were selected from a set of responses that were judged by at least 4 out of 5 MTurk raters to not be sarcastic. Although the post-response pairs varied considerably in their topic and content, the responses were all between 14 and 51 words long. We used PsychoPy (Peirce 2007) to create a stimulus-response system where users could type in their rewritings while reading each post-response pair.

4.3.3 Procedure

A research assistant asked each participant to sit at a computer running the PsychoPy experiment. The experiment software told the participant that they would be reading pairs of internet posts and their responses, and rewriting the responses to be more sarcastic, while attempting to keep the meaning the same (first half) or make the meaning opposite (second half). Before doing the task, participants were first asked how difficult they believed the task would be on a scale of 1 (not at all

difficult) - 7 (extremely difficult). They also rated difficulty for each item after rewriting on the same 1 - 7 scale, in addition to how successful they felt they were after rewriting on a scale from 1 (not at all successful) - 7 (extremely successful).

In order to engage and motivate participants, the experiment was posed as a performance task: They were told that the computer would judge the quality of their rewrites and, if the rewrites were of high quality, they would complete their participation more quickly. This was quantified by displaying a timer in the top right corner of the screen, which signified the remaining time to complete the experiment. Each time an answer was submitted, the participant received one of the two following feedback messages, dependent on the quality of their response: (a) "That was okay, but try for a better response next time." or (b) "Well done! This response was analyzed to be of high quality. One minute has been removed from your remaining time." When high-quality feedback was received, one minute was removed from the experiment timer to motivate participants to submit higher quality answers. "High quality," though not explicitly defined for participants, was calculated by the PsychoPy experiment as any response longer than ten typed characters (regardless of what those characters were) and more than five seconds spent typing. Participants rewrote responses, rating the perceived difficulty and their perceived success on each rewrite until the timer ran out, at which point they were asked debriefing questions about the strategies they used and what cues they use to determine when someone is being sincere.

4.3.4 Coding

Research assistants coded the rewrites in four ways: (1) Full context: how successful the participant was at making the rewrite sarcastic compared with both the original post and original response, (2) Partial context: sarcasm level of rewrite with only the original post (not the original response), (3) polarity of response (positive, negative, or neutral/ambiguous), and (4) strategies used to create sarcasm. These are described in more detail below:

4.3.4.1 Sarcasm level of post-response-rewrite

Two research assistants read the quote, the original response, and the rewrite. They then rated how successful the participant was at modifying the original response to be sarcastic on a scale of 1 - 7. These coders therefore compared the sentiment of the original response with that of the modified rewrite, closely mirroring the rating process for the participants themselves.

4.3.4.2 Sarcasm level of post-rewrite

Not only did we want to compare participants' success ratings to third party raters' success ratings, we also wanted to find out how sarcastic the rewrites were without comparison to the original responses. To address this, two different research assistants read the original quote but only saw the rewrite -- not the original response. These two coders were asked how sarcastic they felt the rewrite was. Although this coding process is less comparable to the participants' self-reports of their success, it is likely a better measure of whether our participants were actually able to create sarcasm that is recognizable by others.

4.3.4.3 Polarity of rewrites

The process of making something sarcastic may be different for information already imbued with a clear sentiment, so we assessed the degree to which positivity and negativity played a role in the creation of sarcasm. To investigate sentiment polarity, the four research assistants who rated post-response-rewrites and post-rewrites rated each of the 24 post-response pairs as positive, negative, or neutral/ambiguous. Then, taking the three most positive post-response pairs and the three most negative, the four research assistants rated whether each rewrite of those six post-response pairs was positive, negative, or neutral/ambiguous.

Four research assistants rated each of the 24 stimuli internet posts as categorically positive, negative, or neutral/ambiguous. After creating separate ratings, the raters engaged in a communal discussion about stimuli where disagreements existed. Initial ratings and ratings after discussion were recorded. The three most unambiguously positive stimuli and the three most unambiguously negative stimuli were identified by examining initial ratings and using post-communal discussion ratings as a tiebreaker. The four research assistants then coded all rewrites of these six post-response pairs in the same way -- either positive (1), negative (-1), or neutral (0).

4.3.4.4 Strategies

Finally, we examined each participant's response to the post-experiment question, "What strategies did you use to make these items sarcastic?" Four research assistants who were not involved in rating rewrites' sarcasm level or polarity collaborated with the authors to identify patterns in participants' self-reported

strategies for creating sarcasm. Once strategies were identified, we created a coding system to determine how often those strategies had been used in rewrites and tested it.

4.4 Results

Participants generated a total of 628 rewrites. Of these, 22 were dropped because of an experimental error. Seven more were dropped because the participant did not enter any responses. This left 599 rewrites for analysis.

4.4.1 Difficulty of creating sarcasm

We compared each participant's prior estimate of the difficulty of the task with their average rating of the trials they completed. Overall, their averaged post-rewrite ratings on the seven-point scale suggested a slightly higher perceived difficulty than they had anticipated, $M = 4.09$, $SD = 1.43$ for anticipated difficulty and $M = 4.78$, $SD = 0.92$ for experienced difficulty, $t(81) = 4.21$, $p < .001$, 95% CI [.37, 1.03]. Free responses in the post-experiment questionnaire supported this result, with participants remarking that it was "extremely difficult," "tricky," and "harder than I imagined." Similarly, participants rated their success only to be moderate at rewriting the forum responses, $M = 3.34$, $SD = 1.46$, on the 1 (not at all successful) - 7 (extremely successful) scale.

4.4.2 Sarcasm Level of Post-Response-Rewrite

Research assistants were asked to rate participants' success at modifying the original response into a sarcastic rewrite, on a scale of 1 (not at all successful) to 7 (extremely successful). Rater 1 ($M = 3.19$, $SD = 2.55$) coded higher than rater 2 ($M = 2.37$, $SD = 1.63$), $t(597) = 7.65$, $p < .001$. Their ratings covaried slightly, $r(597) =$

.278, $p < .001$. These findings suggest that participants were at least modestly successful at creating sarcastic meaning. We also found a modest relationship between the average of research assistants' ratings with participants' own ratings of their success, $r(597) = .173$, $p < .001$. We also computed correlations for each of our coders individually to see if one coder understood sarcasm in a very different way from our participants. Research assistant raters individually showed similar levels of covariance with participants' ratings, $r(597) = .113$, $p = .006$ and $r(596) = .156$, $p < .001$, respectively.

4.4.3 Sarcasm Level of Post-Rewrite

Separate research assistants were asked to rate the sarcasm of the rewrite, without access to the original response, on a scale of 1 (not at all successful) to 7 (extremely successful). Rater 3 ($M = 3.64$, $SD = 2.16$) coded higher than rater 4 ($M = 3.08$, $SD = 1.68$), $t(597) = 6.53$, $p < .001$. The two research assistant sarcasm raters showed moderate agreement for the level of sarcasm present in the rewrites, $r(597) = .407$, $p < .001$. There was a small but significant correlation between the participants' success rating and the average amount of sarcasm perceived by our coders, $r(597) = .160$, $p < .001$. Correlations between participants' ratings and individual research assistants' ratings were similar, $r(597) = .176$, $p < .001$ and $r(597) = .081$, $p = .048$, respectively.

4.4.4 Polarity of Rewrites

Overall, coders rated the post-response pair stimuli more negatively than positively, $t(149) = -3.033$, $p = .003$, but the most negative stimuli were rewritten

more negatively than the most positive stimuli were: Rewrites of the three most negative stimuli showed more negativity than positivity, $t(70) = -4.42$, $p < .001$, but rewrites of the three most positive stimuli did not differ on positive or negative sentiment, $t(78) = -0.19$, $p = .849$. Further, coders rated rewrites of the positive stimuli more towards the middle of the scale ($M = -.01$, SD = .59) than rewrites of the negative stimuli, which were more negative ($M = -.30$, SD = .56), $t(148) = -3.00$, $p = .003$, 95% CI for the difference, [-0.1, -.47]. See Figure 1.

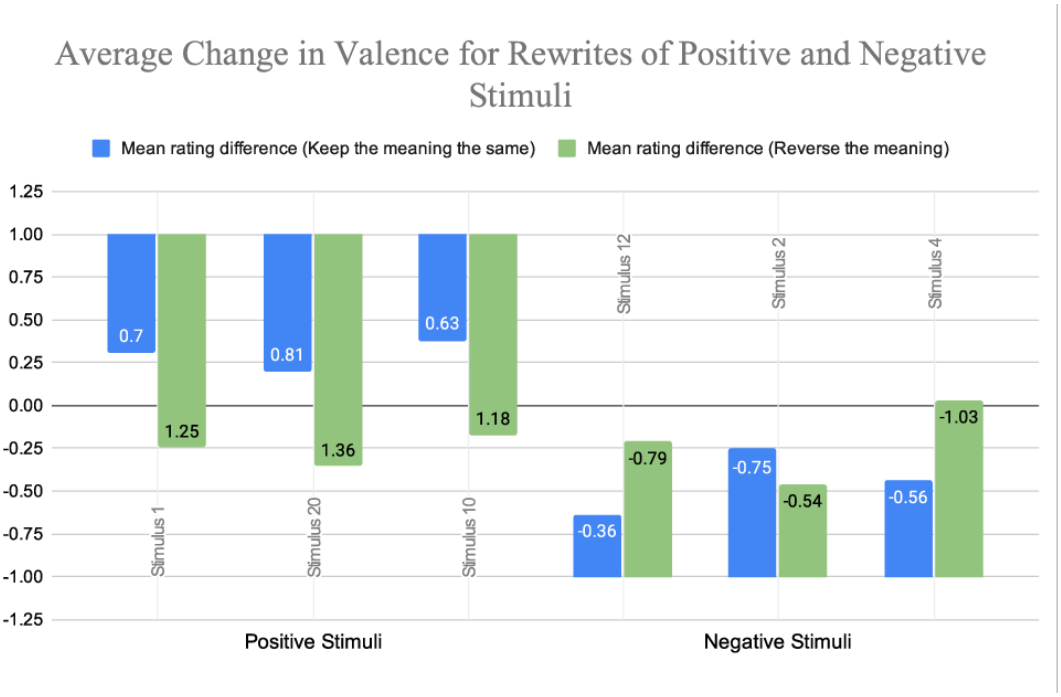

Figure 1. Initially positive stimuli saw large drops in positivity in their rewritings, (especially when participants were asked to reverse the meaning of the forum responses) becoming essentially neutral overall. Initially negative stimuli saw a moderate rise in positivity in their rewritings, coming closer to neutral, but still decidedly negative overall.

4.4.5 Strategies

Four research assistants collaborated with the authors to critically examine participants' self-reported strategies for creating sarcasm. Fifteen distinct strategies were identified and organized into five supercategories: (1) mental, to describe strategies where participants reported attempting to predict their addressees' responses to their productions, (2) structural, to describe strategies to organize their rewrite in a particular way, (3) emphasis, to describe strategies that call attention to specific elements of the response, (4) tone, to describe adjustments to the emotional content of the message, and (5) content, to describe adjusting the semantic meaning of the response in certain ways. Participants also noted particular words that suggest sarcasm, which we examined as well.

4.4.5.1 Mental

Mental strategies included *simulations* and *prior knowledge*. Fifteen participants reported attempting to predict the impact of their rewrites (simulation) or base it on sarcasm they had encountered in the past (prior knowledge). Simulation was reported by 15 of our participants, and prior knowledge by 7. Because these mental strategies do not have a clear representation in the text of the rewritten responses, we do not examine them further.

4.4.5.2 Structural

Structural strategies included *adding questions* and *brevity*. Participants reported adding questions to their rewrites to make them more sarcastic, fitting with "rhetorical questions" as a category of irony (Hancock, 2004; Fox Tree et al., 2020).

Three participants reported adding questions. Five participants reported using brevity. They suggested that shortening the length of the response would increase its sarcastic effect.

4.4.5.3 Emphasis

Emphasis strategies included *word emphasis* and *idea emphasis*. Word emphasis included capitalization (of words), elongation (of vowels), and adding punctuation of various types. Idea emphasis included exaggerating a key concept within the text, making a statement absolute (true or false in all cases), or adding discourse markers (e.g. *well*, *oh*, *like*).

4.4.5.4 Tone

Tone adjustments included *jocularity* (four participants) and *condescension* (four participants). These participants reported changing the emotional tone of the response to be more amusing in a friendly or joking way, or more nasty and insulting toward the original post.

4.4.5.5 Content

Content adjustments included *opposite sentiment* (12 participants) and *affirmations* (10 participants). These reports involved changing the meaning of the response to contrast with the perceived sentiment, or using explicit confirmations to highlight an inconsistent sentiment, e.g., “Oh yeah that totally makes sense”.

4.4.5.6 Word choices

In addition to the higher-level, heuristic strategies reported, participants also reported using specific words and phrases to create sarcasm. There were 17 words

and phrases reported: *u(h)m (8)*, *yeah (64)*, *so (84)*, *well (24)*, *like (78)*, *totally (39)*, *of course (41)*, *oh yeah (9)*, *as if (7)*, *literally (4)*, *always (36)*, *for sure (3)*, *surely that's right (0)*, *obviously (13)*, *absolutely (2)*, *really (20)*, and *clearly (10)*. We counted the frequency of these words in the 599 rewrites to determine how much people's reports matched their behavior. Because *um*, *yeah*, *so*, *well*, *like*, and *always* occurred in the original responses, they were also likely to occur in rewrites far more frequently, as participants were asked to rewrite those responses. Therefore, we did not examine them here.

4.4.6 Strategies used in rewrites

We analyzed the 599 rewrites to look for evidence of all but two of the self-reported strategies. The two we did not explore were the mental categories simulation and prior knowledge. We did not assess simulation because we had no way of knowing when participants used this strategy of imagining the effect of their statements. We did not assess prior knowledge because we had no way of knowing when participants were explicitly drawing on their previous experiences with sarcasm.

4.4.6.1 Structural

Rewrites were coded dichotomously for presence or absence of questions. First we identified rewrites that contained more question marks than the original response ($N = 60$). In order to ensure that we did not miss any added questions, we then examined the remaining rewrites that included question marks (N = 29) and determined whether the question was similar to one asked in the original response.

Two research assistants coded these as containing either "only similar or identical questions" or "at least one different question," achieving moderate interrater reliability (κ = .541). Counting only those questions that both research assistants agreed were novel, this analysis suggests that at least 72 posts contained a new question, or 12% of the rewrites.

Many of our participants' responses were brief. Of the 599 rewrites, 490 were shorter than their responses. This result is likely to be at least partially due to the fact that when repeating information people tend to make it more succinct (Clark & Wilkes-Gibbs, 1986), as well as suboptimal participant motivation.

4.4.6.2 Emphasis

We analyzed emphasis via capitalization and elongation. We did not analyze punctuation due to a limitation in our experiment setup that made it more challenging for participants to type punctuation. The capitalization category was straightforward: we created a Python script to perform an exhaustive search through the rewrites for sequences of at least two letters that were capitalized. Then, a manual check was done to eliminate acronyms (e.g., "USA", "TV", "LOL", etc). There were no capitalized words in the original responses, so all capitalizations within the rewrites were considered novel. Fifteen rewrites included at least one word that was all in capital letters. This is likely lower than might have been possible with other data collection methods (in our experiment, due to the same limitation making punctuation difficult to analyze, capitalization required pressing shift for each letter intended to be capitalized).

Assessing elongation required differentiating elongations from typographical errors. Elongations were defined as: (1) words that had one or more extra final letters when compared to a typical spelling (e.g., “itt.” “noww,” “soooo”) and (2) words contained *two* or more extra repeated letters within the word (e.g., “totallllly,” “bessst”). Although this assessment may contain some false positives that actually were due to participants accidentally hitting a key twice, it also may have helped weed out some typographical errors (e.g. “totallly,”, “besst”), which could be less noticeable in the middle of a word than extra errors at the end of a word. None of the original responses contained elongations, so all elongations within the rewrites were considered novel. There were 36 clear instances of elongation in the 599 rewrites.

4.4.6.3 Word choices

Participants reported using specific words and phrases to imbue their rewrites with sarcastic sentiment. Six of the words participants reported using (*um*, *yeah*, *so*, *well*, *like*, and *always*) were also present in the original responses from the post-response pair stimuli, so we do not report their frequency here. However, ten of the eleven other reported words did appear with various frequencies. See Table 1.

Word	Frequency in Rewrites
Totally	39
Of course	41
Oh yeah	9
As if	7
Literally	4
For sure	3
Surely that's right	0
Obviously	13
Absolutely	2
Really	20
Clearly	10

Table 1. Frequency of words and phrases in rewritten forum responses that participants reported using strategically to create sarcastic content.

Totally, *of course*, *really*, *clearly*, and *obviously* were used the most frequently. They share an ability to polarize meaning, fitting with our participants' reports of exaggeration and absolutes as ways to generate sarcasm.

4.4.6.3 Tone and Content

Some strategies were more abstract and required more extensive human coding to identify. We refer to these strategies as literary devices, although they do not all perfectly fit into that category. They are *exaggeration*, *absolutes*, *jocularity*, *opposite sentiment*, *condescension*, and *affirmations*. We created a coding scheme for

each of these strategies (See Appendix A) and asked four research assistants to code all 599 rewrites for the presence of each of them.

4.4.6.4 Reliability of Literary Devices

Prior to examining the frequency of literary devices in the rewrites, we wanted to see how reliably they could be coded, so we computed Krippendorf's alpha for each literary device (see Table 2). Although all alphas and kappas showed agreement above chance, no alpha value exceeded .25, suggesting that agreement overall was present but limited. Suspecting that some coders may agree more than others, we computed Cohen's kappa values for pairs of coders on all permutations. Indeed, there were several potentially interesting patterns that are noteworthy.

	Exaggeration	**Absolute**	**Jocularity**	**Opposite Sentiment**	**Condescension**	**Affirmation**
Krippendorff's α	.162	.225	.093	.378	.160	.150
Cohen's κ (a/b)	.148	.272	.281	.339	.199	.389
Cohen's κ (a/c)	.346	.232	.100	.319	.297	.283
Cohen's κ (a/d)	.350	.271	.102	.461	.249	.089
Cohen's κ (b/c)	.111	.196	.087	.407	.219	.111
Cohen's κ (b/d)	.138	.254	.109	.356	.115	.028
Cohen's κ (c/d)	.396	.340	.296	.444	.179	.206

Table 2. The top row displays Krippendorff's Alpha values for each of six literary devices reported by participants, followed by five more rows that display Cohen's Kappa values for each pair of raters. These results show that even identically trained coders have differences in their conceptualizations of these literary devices.

For exaggeration, coders A, C, and D seemed to agree with each other more than coder B, suggesting coder B thought about exaggeration differently than the other three coders. Together, those three coders' alpha value is .365. Nonetheless, it is clear that there is at least some similarity in how all four raters viewed exaggerative content. The presence of absolute statements showed slight reliability above chance levels as well, along with condescension. When coding jocularity, coders A and B agreed more strongly with each other, and coders C and D agreed more strongly with each other, though there was less overlap between other pairs, suggesting that there may be two reliable but distinct concepts that make up jocularity for our coders. It is not too surprising that opposite sentiment showed higher levels of agreement, partially because it is one of the more straightforward literary devices, and partially because we explicitly asked participants to rewrite their sarcastic interpretations in a way that utilized opposite sentiment. Finally, affirmations show an interesting pattern as well, suggesting that coders A and B (with the highest reliability) may partially share a concept with coder C, while coder C shares an entirely different concept with coder D.

4.4.6.5 Sarcasm content of literary devices

Since all six literary devices showed above-chance coding reliability, we next examined to what extent these literary devices were successful at creating sarcastic content. We used the three measurements of sarcasm previously discussed for each rewrite: The participants' success rating, the research assistants' full-context rating of

the participants' success, and the different research assistants' partial context rating of the rewrites' sarcasm, not including viewing the original responses.

	Participant rating	Mean Coder Rating Full Context	Mean Coder Rating Partial Context
Exaggeration	.168***	.380***	.438***
Absolutes	.104*	.125**	.180***
Jocularity	.126**	.368***	.445***
Opposite Sentiment	.135**	.069	.015
Condescension	.151***	.517***	.498***
Affirmation	.072	.015	.046

Table 3. Pearson correlations between the average of four coders' determinations of the presence of exaggeration, absolute statements, jocularity, opposite sentiment, condescension, and affirmations, and separate coders' determination of the presence of sarcasm. * = $p < .05$, ** = $p < .01$, *** = $p < .001$

When more coders agreed that exaggerations, absolutes, jocularity, or condescension were present in a rewrite, the more likely separate coders (and participants themselves) were to believe that sarcasm was present in that rewrite. Condescension was a particularly impressive predictor, accounting for approximately 25% of the variance in sarcasm ratings for coders with either partial or full context for their ratings. It is interesting that condescension was not as strong a predictor for participants' own determination of how sarcastic they were, perhaps suggesting that people may not realize that their own use of sarcasm comes off as condescending. This could also be an artifact of having research assistants who are finely tuned to recognize condescension due to their experience studying related topics.

The table above suggests that as we defined them, exaggeration, absolutes, jocularity, and condescension are predictive of sarcastic intent. We were surprised to find that opposite sentiment and affirmations were not predictive. In order to develop our understanding of our data, we ran exploratory analyses on the same dataset, but split by running block. Because participants were asked to craft some rewrites to keep the same sentiment as the original response, and were asked to reverse the sentiment for others, we wondered whether that was interfering with our analysis of the relationship between opposite sentiment and sarcasm presence. See Table 4.

	Participant rating		**Mean Coder Rating Full Context**		**Mean Coder Rating Partial Context**	
	Same meaning	**Reversed meaning**	**Same meaning**	**Reversed meaning**	**Same meaning**	**Reversed meaning**
Exaggeration	.228***	.137**	.393***	.372***	.423***	.448***
Absolutes	.124	.086	.149*	.112*	.195**	.175**
Jocularity	.105	.135*	.251***	.458***	.350***	.522***
Opposite Sentiment	.239***	.040	.209**	.008	.162*	-.049
Condescension	.101	.181**	.580***	.476***	.490***	.505***
Affirmation	.152*	.043	-.097	.081	-.035	.093

Table 4. Pearson correlations between the average of four coders' determinations of the presence of exaggeration, absolute statements, jocularity, opposite sentiment, condescension, and affirmations, and separate coders' determination of the presence of sarcasm, split by running block. In one block, participants were asked to rewrite the non-sarcastic responses while keeping the same meaning, and in the other block, participants were asked to rewrite the non-sarcastic responses while reversing the meaning. * = $p < .05$, ** = $p < .01$, *** = $p < .001$

By splitting the correlations by block, we show that opposite sentiment was a valuable predictor of sarcasm both as understood by participants and as understood by

different coders, but only when participants were asked to keep the meaning the same -- not when they were asked to reverse it. Affirmations showed a weak correlation with participants' sarcasm ratings, but no other significant correlation was present, suggesting that affirmations may not be a useful predictor of sarcastic content, except potentially as viewed by the creators of sarcasm themselves. Other correlations were relatively similar to the non-split data.

4.5 Discussion

Although there are many ways to define sarcasm and people do not always agree (or even often agree, cf. Fox Tree et al., 2020), when they are focused on creating sarcasm, such as by converting a sincere text to a sarcastic one, they can be at least modestly successful as assessed both by themselves and by outside observers. When rewriting content that uses a negative sentiment people tend to rewrite it as even more negative, but when working with positive sentiment people tend to rewrite using a more neutral sentiment. People identified a variety of strategies that they use to communicate sarcasm and a content analysis of written productions confirmed the use of several of these strategies, as well as related words and phrases.

The reported strategies can be grouped into at least 15 categories, organized into 5 supercategories, (1) mental, (2) structural, (3) emphasis (4) tone, and (5) content. Mental strategies include simulations and prior knowledge, which we were unable to test. We found evidence that participants used structural strategies, like adding questions and brevity. We also found that they used emphasis strategies like word emphasis (strategic use of punctuation, capitalization, and elongation) and idea

emphasis (adding absolutes and discourse markers). Finally, we found that they used tone strategies like jocularity and condescension. Content strategies were a more complicated story, as reversing the sentiment was present (but participants were explicitly asked to do this) and affirming a sentiment (which showed no relationship to others' ratings of sarcasm).

One limitation of this work is that the sarcasm generated by our participants, though diverse, probably does not account for all types of written sarcasm, even if our definition of "all" only includes sarcasm in internet forums. Although the 24 post-response pairs that we used were diverse in their topics and content, participants were in a situation where they had to attempt to ground with the original writers -- that is, they had to understand the writer's original intent in order to create a sarcastic version of the text. Because there was limited context (posts were from pre-2010s, participants didn't get to see the entire message thread and/or other posts by the authors), it is possible that participants refrained from using more subtle forms of sarcasm.

Another limitation is that our strategy categories were of relatively low frequency - our 82 participants reported a total of 116 distinct ideas that we grouped into 15 categories. Although this results in some cell sizes as low as three, we are confident that the cells with larger numbers of reports would be represented in a larger sample.

Sarcasm has been defined in a plethora of different ways, but too often the definitions hinge too much on researchers' own perceptions of what constitutes

sarcasm or verbal irony, and not enough on perceptions of people producing the sarcastic content. To remedy this problem, this work attempts to reconnect with the populations that researchers study, and to find out more about their own definitions of sarcasm, treating differences of opinion as diversity rather than error. It does so by first showing the diversity of strategies that people use to create sarcasm in writing. These results will help both researchers who bring strong definitions of sarcasm to their work, as well as researchers who do not. Those who try to disambiguate sarcasm from similar concepts like verbal irony will be able to use this work to inform their definitions in ways that make them more accessible to the general public, and those who use community-driven definitions of sarcasm can better understand what that definition actually entails. The information provided here could be useful to screenwriters who want to create a sarcastic character, as well as any other storytellers who want to imbue their characters with a sarcastic edge. Sarcasm's many facets allow for nuanced character development, as well as subtle storytelling. In a broad way, the message of this work is that we can rarely be certain about whether sarcasm is being used. Concepts like sarcasm are broad and fluid enough that there cannot be a single central definition for all. Without certainty of what sarcasm is, we often find ourselves in ambiguous moments in communication. That ambiguity itself has beauty -- if understanding was always easy, it wouldn't be nearly as rewarding.

4.6 References

Attardo, S., Eisterhold, J., Hay, J., & Poggi, I. (2003). Multimodal markers of irony and sarcasm. *Humor - International Journal of Humor Research*, *16*(2). https://doi.org/10.1515/humr.2003.012

Bryant, G. A., & Fox Tree, J. E. (2002). Recognizing verbal irony in spontaneous speech. *Metaphor and Symbol*, *17*(2), 99–119.

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

Burgers, C., van Mulken, M., & Schellens, P. J. (2011). Finding Irony: An Introduction of the Verbal Irony Procedure (VIP). *Metaphor and Symbol*, *26*(3), 186–205. https://doi.org/10.1080/10926488.2011.583194

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

Clark, H. H., & Gerrig, R. J. (1984). On the pretense theory of irony.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39. https://doi.org/10.1016/0010-0277(86)90010-7

Colston, H. L. (2017). Irony and Sarcasm. *The Routledge Handbook of Language and Humor*, 234.

Dress, M. L., Kreuz, R. J., Link, K. E., & Caucci, G. M. (2008). Regional Variation in the Use of Sarcasm. *Journal of Language and Social Psychology*, *27*(1), 71–85. https://doi.org/10.1177/0261927X07309512

Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., & Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *ArXiv Preprint ArXiv:1708.00524*.

Hancock, J. T. (2004). Verbal Irony Use in Face-To-Face and Computer-Mediated Conversations. *Journal of Language and Social Psychology*, *23*(4), 447–463. https://doi.org/10.1177/0261927X04269587

Huang, L., Gino, F., & Galinsky, A. D. (2015). The highest form of intelligence: Sarcasm increases creativity for both expressers and recipients. *Organizational Behavior and Human Decision Processes*, 131, 162–177. https://doi.org/10.1016/j.obhdp.2015.07.001

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). Association for Computational Linguistics.

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. http://dx.doi.org/10.1037/0096-3445.118.4.374

Nunberg, G. (2001). The Edge. In *The Way We Talk Now: Commentaries on Language and Culture from NPR's" Fresh Air*. Houghton Mifflin Harcourt.

Oraby, S., Harrison, V., Reed, L., Hernandez, E., Riloff, E., & Walker, M. (2017). Creating and Characterizing a Diverse Corpus of Sarcasm in Dialogue. *ArXiv:1709.05404 [Cs]*. Retrieved from http://arxiv.org/abs/1709.05404

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Peters, S., & Almor, A. (2017). Creating the Sound of Sarcasm. *Journal of Language and Social Psychology*, *36*(2), 241–250. https://doi.org/10.1177/0261927X16653640

Qadir, A., & Riloff, E. (2014, October). Learning emotion indicators from tweets: Hashtags, hashtag patterns, and phrases. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1203-1209).

Reyes, A., Rosso, P., & Veale, T. (2013). A multidimensional approach for detecting irony in Twitter. *Language Resources and Evaluation*, *47*(1), 239–268. https://doi.org/10.1007/s10579-012-9196-x

Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., & Huang, R. (2013). Sarcasm as Contrast between a Positive Sentiment and Negative Situation. In *EMNLP* (Vol. 13, pp. 704–714).

Roberts, R. M., & Kreuz, R. J. (1994). Why Do People Use Figurative Language? *Psychological Science*, *5*(3), 159–163. https://doi.org/10.1111/j.1467-9280.1994.tb00653.x

Rockwell, P. (2000). Actors', Partners', and Observers' Perceptions of Sarcasm. *Perceptual and Motor Skills*, *91*(2), 665–668. https://doi.org/10.2466/pms.2000.91.2.665

Sperber, D. (1984). Verbal irony: Pretense or echoic mention? *Journal of Experimental Psychology*, *113*(1), 130–136.

Sperber, D., & Wilson, D. (1981). Irony and the Use-Mention Distinction. In *Radical Pragmatics* (pp. 295–318). New York: Academic. Retrieved from http://discovery.ucl.ac.uk/1568495/

Walker, M. A., Fox Tree, J. E., Anand, P., Abbott, R., & King, J. (2012). A Corpus for Research on Deliberation and Debate. In *LREC* (pp. 812–817).

4.7 Appendix A: Coding Scheme for the Presence of Sarcasm Strategies

Exaggeration: If a rewrite describes something as more or less than its accepted reality. Understatement or overstatement. Another way to think about this is as hyperbole. If it seems like it's literally true, it's probably not an exaggeration.

- I am starving could be exaggeration
- I ate a thousand tacos (very likely exaggeration)
- Beyonce is a goddess (definitely an exaggeration)
- I cleaned all the dishes (probably not an exaggeration)
- It's so hot you could fry an egg on the sidewalk
- You should never put a baby in a hang glider (would not be included)
- I never sleep (almost definitely an exaggeration)
- "Oh, your tacos are so bad! Let's go eat McDonalds instead :)" (not likely to be an exaggeration)
- "Oh, your tacos are the worst! Let's go eat McDonalds instead :)" (is an exaggeration)

Absolute: If a rewrite contains a word or phrase that implies "all" or "none" of something, or the most or the least possible, without hedging. Statements that imply majorities or are nearly absolutes were not included. Examples:

- "I cleaned all the dishes" would be included
- "I cleaned most of the dishes in the universe" would not be included
- "Nobody loves you like I do" would be included
- "You are kinda the best thing ever" would not be included, because "kinda" hedges the absolute statement

Jocularity: Does the rewrite playfully poke fun or tease?
If it is likely to offend someone, code "no" for jocularity. Condescension better describes teasing that is intended to hurt someone's feelings. Jocularity is teasing that is not intended to hurt someone's feelings.

- "Oh, your tacos are so bad! Let's go eat dirt instead :)"
- "Oh, your tacos are the worst! Let's go eat dirt instead :)"
- "This is a terrible salad" (right after finishing it)
- "Duh" to a respected rocket scientist

Opposite Sentiment: Does the rewrite attempt to say the opposite of the original response at least in part? (literal meaning -- not actual or perceived meaning)
Please imagine a horizontal scale bisected by a "neutral" point in the middle (a visual example was given). Then examine each rewrite and its corresponding statement in the original response and determine whether the two statements would fall on opposite sides of the "neutral" point. If so, the rewrite should be marked as containing at least some opposite sentiment.

Condescension: Does the rewrite clearly act condescendingly or nastily toward something, or try to belittle something? On the surface they're nicer than what they mean. Patronizing courtesy or politeness.

- "You're so helpful! Thank you, sweetheart!"
- "When you wash the dishes, I need you to get all the food off of them"
- "You know what'd be great? If you showed up on time"
- "What a cute little tune" - to Beethoven

Affirmation: Does the rewrite attempt to agree with the original post (quote) in a clear way? This is meaning-specific, not word-specific. Does not need to have any specific word, just needs to affirm part of the post (not the response).

- "I totally agree with your statement"
- "That seems right to me"
- "No problem at all there!"

4.8 Appendix B: Top positive and negative stimuli, with examples of rewrites

The three most unambiguously positive stimuli were as follows:

[Stimulus 1]

POST: *pop song - good vibrations (the beach boys) next is - a day in the life (the beatles) one of these days (pink floyd) echoes (pink floyd) cool world (mondo rock) eagle rock (daddy cool)*
RESPONSE: *i like anything rock and sometimes punk rock, i like thursday's song o man i forgot it..... but i like it and um maroon 5 harder to brethe is awesome!*
EXAMPLE REWRITE: *i loooove anything rock and especially punk rock. I love thursday's song and i cant get enough of maroon 5 harder to breathe. its like totally awesomeee*

[Stimulus 10]

POST: *cunx, i have to ask you, did you do bbq ribs in your restaurant? and were they dry southern ribs?*
RESPONSE: *yes, i did ribs but i like 'um juicy. its all about the sauce baby.*
EXAMPLE REWRITE: *Yeah, I totally made ribs. With my nonexistent grill. I made 'um juicy, is all about the sauce when you fake barbecue*

[Stimulus 20]

POST: *there days my favorite is flashpoint, its really very interesting show.*
RESPONSE: *lot of tv shows are there as like: lost dexter heroes jersey shore hung big love born to death monk this shows are outstanding.must watch it.....!*
EXAMPLE REWRITE: *I was going to recommend you a list of other really great shows, but it sounds like you're already a TV expert*

All four coders rated the first and third stimuli as positive. Three of the four coders agreed that the second stimulus was positive and came to unanimous agreement in the communal discussion.

The three most unambiguously negative stimuli were:

[Stimulus 2]

POST: *then in that case they wouldn't be good parents now would they? RESPONSE: they treat their kids like that because they are conservative thinkers. and you wouldnt claim that all conservative thinkers are bad parents. well...maybe i would.*
EXAMPLE REWRITE: *oh yes bevause liberal thinkers make the bessssst parents*

[Stimulus 4]

POST: *the same applies to obama. he gives a great speech...so does al pacino. any good actor can give a good speech. obama has demonstrated zero ability to govern. he can't even vote regularly for or against things...hence all of his present votes in the illinois government. there's no hypocrisy. she's been a mayor and is presently a great governor. obama is nothing but a senator from the illinois political machine. he has done little or nothing. he hasn't sponsored any significant bills. he hasn't even been in the senate much in the last few months...he's been busy running for president. ...don't forget...obama has been a senator for three years...he hasn't been doing what senators normally do for three years though...he's been running for president and writing his memoirs*
RESPONSE: *frankly, obama's limited experience is towards the bottom of my list of concerns because all experience seems to do is make it so you can learn the art of true "politicing" . i mainly care that his ideas suck.*
EXAMPLE REWRITE: *oh because Obama's limited experience is totally towards the bottom of my list of concerns because all experience does is make it so you can learn the art of true politicing" right. I totally care that he has "sucky" ideas"*

[Stimulus 12]

POST: *it certainly beats the alternative of even more deaths from illegal and unregulated abortions. or perhaps you'd just want to see more babies brought to term and abandoned in dumpsters. but if you really just want to trade non sequitors instead of discussing the issue in all of its complexity, maybe i should just say: abortion sucks, but the alternative is worse.*
RESPONSE: *that is the biggest spin the liberal left every has put out.....i was alive back then and you find more babies thrown away in garbage can now then you did then.....barely a week goes by that you don't see that hapening now.....*
EXAMPLE REWRITE: *The liberal left's view is clearly so insightful and correct based on how many babies we see thrown away on the daily*

All four coders rated all three of these stimuli as negative.

5 Overall Discussion

Sarcasm has been studied from many different angles. Here I presented three directions from which to develop a more complete understanding of a phenomenon that is complex, diverse, and elusive. Each of these directions is useful in its own way, but together they begin to show that sarcasm is many different things to different people in different contexts. While it is tempting to look at the analyses above and proclaim victory with the strongest findings (e.g. *um* and *uh* strongly predict sarcasm (D'Arcey et al., 2019), condescension strongly predicts sarcasm (D'Arcey, Fox Tree, under review), it is important to remember that the many more moderate findings are just as important for understanding how people think about sarcasm (e.g. ellipses and single quoted words predict sarcasm (D'Arcey et al., 2019), exaggeration predicts sarcasm (D'Arcey, Fox Tree, under review)).

Across three research papers that are published or under review, I approached sarcasm from a computational and psychological perspective, and through both production and comprehension lenses. In the computational approach I showed evidence that sarcasm is commonly co-present in writing with signals asking the reader to wait, potentially drawing on the contextual inappropriateness of asking an interlocutor to wait using an asynchronous medium. In the experimental methods, I showed evidence that agreement on sarcasm can be quite low, whether between interlocutors themselves or even between trained research assistants. Finally, in the qualitative approach, I showed that the diversity of ways that people think about

sarcasm is far broader than the ways that researchers generally portray it, and that many of those ideas reliably predict the presence of sarcasm.

The emerging story seems to suggest that sarcasm is a multidimensional entity -- one that does not rely on any one mechanism like "saying the opposite of what you mean" or "being contextually inappropriate", but one that draws on many strategies and attributes, both linguistic and contextual. Future work that would follow this direction could stem off in a variety of different ways. For example, it would be tremendously beneficial to draw more heavily on public knowledge in a more demographically representative way. One way to do so would be to collect statements that are potentially sarcastic and ask people to help disambiguate their sarcastic content in a more qualitative way. Questions like "what do you need to know about the writer to figure out if they were being sarcastic?" could bring valuable insight to understanding the sources of sarcasm.

Likewise, in today's world, video and audio recording is omnipresent, making it easy for people to hold others responsible for their past words and actions. As a result, it has become strategic for politicians to claim, "I was joking" or "I was being sarcastic". This leads to a situation where it becomes difficult to know which story is more representative. Presumably it would be useful to be able to ascertain whether or not sarcasm was present, but we currently have few tools to do so, and have little knowledge about this strategy's effect on public opinion.

On a similar note, sarcasm, when used as a form of humor, may be used and understood differently by people in positions of power than it is by people who are

not. Since expressions of humor recognition (e.g., laughter) differ based on status (Oveis et al., 2016), it is plausible that expressions of humor themselves differ as well. And from a more general perspective, power differences may be only the tip of the iceberg.

There is already evidence that cultural differences in ideas of sarcasm vary widely (Dress et al., 2008). Taking the strategies described in Oh SO Sarcastic above and testing them against representative populations could lead to several important results: First, it could unearth new strategies that haven't yet been examined. Second, it can give us an idea of whether there is regional or cultural variation in the connection of each strategy to the idea of sarcasm. For instance, maybe university students in California tend to strongly associate sarcasm with condescension, but for adults in the Midwest, sarcasm is much more about affirmation. Third, it could tell us whether some strategies are more widely associated with sarcasm than others. Perhaps jocularity will turn out to be associated with sarcasm at low levels for most people.

In this complex problem space, it is no wonder that researchers have studied sarcasm and come to very different conclusions about its nature. One is reminded of the tale of the blind men and the elephant (e.g., Daigneault, 2013), in which the blind protagonists all touch a different part of the animal and draw conclusions about the nature of the elephant as a whole, resulting in six completely different (but all accurate) definitions of an elephant. It is my hope that by thinking about sarcasm

from a multitude of different viewpoints (especially those of non-scientists), we will open our eyes to a broader view of how sarcasm functions in our world.

6 Introduction and Discussion References

D'Arcey, J. T., & Fox Tree, J. E. (under review). Oh, SO Sarcastic: Diverse Strategies for Being Sarcastic.

Daigneault, P.-M. (2013). The Blind Men and the Elephant: A Metaphor to Illuminate the Role of Researchers and Reviewers in Social Science. *Methodological Innovations Online*, 8(2), 82–89. https://doi.org/10.4256/mio.2013.015

Dress, M. L., Kreuz, R. J., Link, K. E., & Caucci, G. M. (2008). Regional Variation in the Use of Sarcasm. Journal of Language and Social Psychology, 27(1), 71–85. https://doi.org/10.1177/0261927X07309512

Duffy, K. A., & Chartrand, T. L. (2015). The Extravert Advantage: How and When Extraverts Build Rapport With Other People. *Psychological Science*, 26(11), 1795–1802. https://doi.org/10.1177/0956797615600890

Fox Tree, J. E., D'Arcey, J. T., Hammond, A. A., & Larson, A. S. (2020). The Sarchasm: Sarcasm Production and Identification in Spontaneous Conversation. *Discourse Processes*, 57(5–6), 507–533. https://doi.org/10.1080/0163853X.2020.1759016

Gallois, C., Ogay, T., & Giles, H. (2005). Communication Accommodation Theory: A look back and a look ahead. In *Theorizing about intercultural communication* (pp. 121–148).

Oveis, C., Spectre, A., Smith, P. K., Liu, M. Y., & Keltner, D. (2016). Laughter conveys status. *Journal of Experimental Social Psychology*, 65, 109–115. https://doi.org/10.1016/j.jesp.2016.04.005

Parkinson, C., Kleinbaum, A. M., & Wheatley, T. (2018). Similar neural responses predict friendship. *Nature Communications*, 9(1), 332. https://doi.org/10.1038/s41467-017-02722-7