Dialogue & Discourse 10(2) 56-78 doi: 10.5087/dad.2019.203

Wait Signals Predict Sarcasm in Online Debates

J. Trevor D'Arcey JDARCEY@UCSC.EDU
Department of Psychology
University of California, Santa Cruz

Shereen Oraby SORABY@UCSC.EDU
Department of Computer Science
University of California, Santa Cruz

Jean E. Fox Tree FOXTREE@UCSC.EDU
Department of Psychology
University of California, Santa Cruz

Editor: Patrick Healey

Submitted 06/2018; Accepted 11/2019; Published online 12/2019

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.

Keywords: sarcasm, irony, fillers, ellipses, online debate, spontaneous communication, written communication, quotations, wait signals

1 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.

1.1 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 if 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.

1.1.1 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.

1.1.2 Machine Sarcasm Identification

On the other hand, computational methods to 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).

1.2 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).

1.2.1 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).

1.2.2 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).

1.2.3 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 *ums*; 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.

1.3 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-signalling 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 meeting 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.

1.4 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.

1.5 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 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.

2.1 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. It has close to 200 hours of transcribed audio recorded at the University of Michigan in Ann Arbor (Simpson, Briggs, Ovens, & Swales, 2002). Fillers in MICASE generally appear to 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.

2.2 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.

2.3 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 one 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 *uh*s than *um*s. The exception was American conversational corpora where there were more *um*s than *uhs.* In Study 2, we turn to the test of whether fillers in writing indicate sarcasm.

3 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).

3.1 Method

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

3.1.1 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.

3.1.2 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.

3.1.3 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.

3.2 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, χ^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, χ^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, χ^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.

3.3 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.

4 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%.

4.1 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.

4.1.1 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.

4.1.2 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).

4.1.3 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.

4.2 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, $\chi2(1) = 363.7$, 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, $\chi2\ (1) = 309.8$, 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, $\chi2\ (1) = 122.6$, $p < .001$, $\Phi = .19$, 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, $\chi2\ (1) = 195.2$, $p < .001$, $\Phi = .24$, 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.

4.3 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.

5 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 this, or elongations like thiiiiis. 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 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.

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.

Data cited herein have been extracted from the British National Corpus Online service, managed by Oxford University Computing Services on behalf of the BNC Consortium. All rights in the texts cited are reserved.

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 Pragmatic*s, *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