



Wait Signals Predict Sarcasm in Online Debates



J. Trevor D'Arcey, Shereen Oraby, and Jean E. Fox Tree
University of California, Santa Cruz

Sarcasm in Writing

- Written sarcasm cannot take advantage of auditory, facial, and gestural cues that can help in identifying spoken sarcasm.
- Many textual markers of sarcasm (e.g., quotations and emoticons) do not overlap with spoken cues to sarcasm.
- Why do specific terms like *let's all*, *really*, and *you mean* contribute to sarcastic perceptions?
- One potential explanation is that these cues call attention to incongruity.
- Other incongruous elements in writing include *um* and *uh*, which are generally used in speaking, not writing.

Wait Signals

- Wait signals are tools used by communicators to slow down consumption of information.
- Although they are less common, *ums*, *uhs* and other wait signals still occur in writing.
- Hearing *um* at the beginning of a turn leads listeners to infer a number of things, such as that the speaker is having production trouble, is uncomfortable with the topic, or is preparing a dishonest answer.
- We hypothesize that reading *um* suggests that writers are intending something different from what they've literally written, such as that they are being sarcastic.
- **Hypothesis: When asked about sarcasm, people will interpret posts with wait signals as more sarcastic.**

Corpus Analysis

Frequency of *Um* and *Uh* (and British equivalents *er* and *erm*) per million words

Spoken				
Word	Michigan Corpus of Academic Spoken English	Corpus of Contemporary American English	BNC	ARTWALK
er	13	4	8542	123
erm	0	0	6029	0
uh	9043	13	N/A	9174
um	9644	6	N/A	11377

Written				
Word	Internet Argument Corpus	SUBTLEX	Corpus of Contemporary American English	BNC
er	17	38	11	11
erm	2	0	0	2
uh	20	717	14	N/A
um	19	87	7	N/A

We determined an approximate magnitude of the difference between spoken and written *ums* and *uhs* by averaging the frequencies of *er* and *erm* for the BNC corpora and the frequencies of *um* and *uh* for all other corpora ($M = 61.39$).

We estimate that *um* and *uh* are an average of about 60 times more common in spoken communications than in written communications, and suggest that this ratio is probably conservative: transcription in the COCAE may have neglected to include filler words.

Overall, this analysis supports the intuition that *um* and *uh* are more common in spoken communication than in written communication, supporting the argument that they are incongruous when written.

Turn-Initial Wait Signals

Materials

We collected 1000 post-response pairs including one of six patterns from the Internet Argument Corpus:

- Starts with *Um*
- Starts with *Uh*
- Includes *obviously*
- Includes *surely*
- Includes *no doubt*
- Includes *clearly*

Examples:

What on earth do people need to tow that can't be managed by a normal family car with a 1.5 litre engine? most minibuses and transit vans get better than 10mpg (a fully loaded minibus gets about 15mpg).

Uh, Jo. Do you plan on towing a 12,000lb 5th wheel trailer, and about a 3,000lb four wheeler trailer (so total about 7.5 tons), with a van?

Examples being? Sorry I'm not the best in history!

Um... I have a question. If we're boycotting French stuff does that mean we boycott the statue of liberty? Cuz I believe it was a gift from France.

Methods: Each post-response pair was rated by 5 unique Mechanical Turk workers as either "includes sarcasm" or "does not include sarcasm"

Results:

Textual Pattern	Sarcasm rate
Starts with "Um"	29.5%
Starts with "Uh"	24.7%
Includes "obviously"	23.4%
Includes "surely"	21%
Includes "no doubt"	18.5%
Includes "clearly"	15.1%
Baseline for the corpus (Walker, Fox Tree, Anand, Abbott, & King, 2012)	12%

Within-Turn Wait Signals

Materials

We collected 720 post-response pairs in four categories from the Internet Argument Corpus:

- Includes *uh*
- Includes *um*
- Includes ellipses
- Includes a quoted word (e.g., "democracy").

Methods were identical to Experiment 1.

Results:

Textual Pattern	Sarcasm rate
Includes "um"	64.1%
Includes "uh"	57.8%
Includes ellipses	40.8%
Includes quoted words	42.1%
Baseline for the corpus (Walker, Fox Tree, Anand, Abbott, & King, 2012)	12%

Incongruity and Sarcasm

People were more likely to rate posts as sarcastic when they included *um*, *uh*, ellipses, and quoted words compared to the corpus in general.

We propose that signaling delay in writing invites readers to consider non-literal interpretations.

We believe that ratings of sarcasm may come from the content's incongruity -- things that don't belong in the medium are perceived as cues. For instance, asking your addressee to wait using an *um* or an *uh* doesn't make as much sense in asynchronous contexts.

Acknowledgements

This research was supported by NSF grant # 1302668.

We also wish to thank our research assistants Valerie Muñoz, Joaquin Canizalez, and Justin Siegel.

Presented at the Psychonomic Society 59th Annual Meeting

Nov 15-18, 2018, New Orleans, Louisiana, United States

Special thanks: Spontaneous Communication Lab