Title
Intended and Perceived Sarcasm Between Close Friends: What Triggers Sarcasm and What Gets Conveyed?

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 45(45)

Authors
Jang, Hyewon
Braun, Bettina
Frassinelli, Diego

Publication Date
2023

Peer reviewed

Intended and Perceived Sarcasm Between Close Friends: What Triggers Sarcasm and What Gets Conveyed?

Hyewon Jang, Bettina Braun and Diego Frassinelli
Department of Linguistics, University of Konstanz
{hye-won.jang, bettina.zinn, diego.frassinelli} @uni-konstanz.de

Abstract

We conducted two experiments to investigate what triggers sarcasm between close friends and whether the factors prompting sarcastic comments in production are also shared with an external observer. In Experiment 1, participants freely reacted to different types of situations in written form and rated their perception of the given contexts, the level of sarcasm of their responses, and the intentions behind their responses. Results showed that the intentions to say clever things or to mock the addressee in a hilarious or friendly manner triggered a higher number of sarcastic answers. In contrast, the intentions to be direct or to be nice to the addressee triggered less sarcastic answers. In Experiment 2, a new group of participants rated the responses collected in Experiment 1 on the same dimensions. Overall, we observed similar patterns in both experiments. However, the intentions to criticize the addressee softly and to say clever things were stronger predictors of sarcasm for the observers than for the producer of the statement.

Keywords: sarcasm; verbal irony; humor; figurative language; non-literal language; experimental findings

Introduction

Sarcasm or verbal irony is a widely studied form of figurative language (Colston, 1997; Sperber, 1984). There is a large body of work investigating sarcasm perception (Bryant, 2010; Woodland & Voyer, 2011) and communicative functions of sarcasm, i.e., verbal aggression (Toplak & Katz, 2000) or mocking (Gibbs, 2000). But there is not much research on what prompts or predicts sarcasm (Zhu & Wang, 2020). It is a question worth exploring as a comprehensive theory of sarcasm should not only identify its communicative functions but also the conditions under which sarcasm is used or preferred (Dews, Kaplan, & Winner, 1995).

The question of what prompts sarcasm is connected to what intentions are behind a sarcastic statement, which is a topic with contradicting findings in the literature. Some studies have argued that sarcasm is used to ridicule or criticize the interlocutor (Colston, 1997; Toplak & Katz, 2000). Other studies have argued that sarcasm implies more positive communicative intentions such as the softening of criticisms (Dews & Winner, 1995) or the enhancement of bonds between the speakers (Gibbs, 2000).

We expect the social context to influence the patterns of sarcasm use as past research has found that solidarity relationship affects verbal irony comprehension (Pexman & Zvaigzne, 2004). We also expect the interlocutor role (e.g. speaker vs. listener) to prompt different perceptions about sarcasm. Bowes and Katz (2011) and Toplak and Katz (2000) have found that identifying the intentions behind sarcasm can be influenced by the interlocutor role.

Thus, in this work, we test what factors trigger sarcasm in a certain social situation (i.e. between close friends) and whether observers can perceive those factors as well. We also test how positively interlocutors view sarcasm as this is a topic with diverging findings in the literature (Dews & Winner, 1995; Toplak & Katz, 2000).

Related Work

Sarcasm and irony

The Merriam-Webster dictionary defines sarcasm as "a sharp and often satirical or ironic utterance designed to cut or give pain". Often intertwined with irony, sarcasm has been defined as an expression of verbal irony (Colston, 2000; Gibbs, 2000), or a figurative language with an element of ridicule that verbal irony does not have (Kreuz & Glucksberg, 1989; Lee & Katz, 1998). However, recent work has been using sarcasm and verbal irony interchangeably due to the difficulty and impracticality of teasing the two concepts apart (Attardo, Eisterhold, Hay, & Poggi, 2003; Fox Tree, D'Arcey, Hammond, & Larson, 2020; Ghosh, Fabbri, & Muresan, 2018; Joshi, Bhattacharyya, & Carman, 2017). In this work, we also treat sarcasm as synonymous with verbal irony. We use *sarcasm* as the default terminology except when the term *(verbal) irony* was explicitly used in previous work. Definitions and descriptions of sarcasm from past research include the following:

- sarcasm is uttering the opposite of what the speaker meant in order to convey a negative and critical comment (Keenan & Quigley, 1999; Kreuz & Glucksberg, 1989);
- sarcasm enables speakers to bond together (Gibbs, 2000);
- sarcasm is non-literal language whereby a complex set of social and communicative goals such as being humorous is achieved (Leggitt & Gibbs, 2000);
- "sarcasm is a sophisticated way of wrapping truth, message, or even mockery within a hilarious manner" (Das & Kolya, 2021).

We take these definitions of sarcasm and use them in our two experiments as guiding elements.

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

Sarcasm for the speaker: Why be sarcastic?

Dews et al. (1995) reported that speakers choose to use verbal irony to be funny, to mute criticisms, to feel in control of their emotions, or to maintain the existing amicable relationships. In contrast, Toplak and Katz (2000) concluded that verbal irony is meant as a means of verbal aggression. Gibbs (2000) and Pexman and Olineck (2002) found that irony is used to mock the addressee in a friendly way. Dews and Winner (1995) and Matthews, Hancock, and Dunham (2006) reported that irony is sometimes used for humor purposes.

Sarcasm for the listener: How do they interpret it?

Researchers have been debating whether an ironic insult (sarcasm) enhances or attenuates the negative emotional load triggered in the addressee. Dews and Winner (1995) and Dews et al. (1995) found that ironic insults were perceived to be less harsh than their literal counterparts (Muting the Meaning Hypothesis). Pexman and Olineck (2002) found that ironic insults were perceived to be more mocking but also more polite than direct insults. Bowes and Katz (2011) found that, when shown aggressive arguments that were either sarcastic or not, participants viewed the sarcastic ones as more aggressive when taking the perspective of the listener, but viewed them as more humorous when taking the perspective of the speaker.

Comparison between intended and perceived sarcasm

Some work has investigated intended and perceived sarcasm together. Fox Tree et al. (2020) designed experiments to compare intended and perceived sarcasm. In their study, pairs of participants engaged in either synchronous movements (one participant mimicking the other participant's movements) or non-synchronous movements (both participants acting their own interpretations of a movement instruction) and had a conversation designed to elicit sarcasm. After the conversation was over, participants watched the recording of their conversation and indicated when they had used sarcasm and when they thought the addressee had used sarcasm. The results showed that participants in the synchronous condition reported more sarcasm in their own utterances but not in the utterances by the addressee.

Oprea and Magdy (2019, 2020) also compared intended and perceived sarcasm. They asked participants to report some of their own tweets that they thought were sarcastic or not sarcastic. They then had those tweets labeled by external expert annotators as sarcastic or non-sarcastic. They reported that 30% of intended sarcastic tweets were missed by external annotators and that 45% of perceived sarcastic tweets were not intended to be sarcastic by the original authors.

Our work is different from the previous work in the following details. First, we use identical materials across all participants to identify 'what prompts sarcasm in a close relationship'. Second, we collect sarcasm ratings from the producer of the comments and multiple external observers. We believe that sarcasm is inherently a subjective phenomenon, as is also stated in Oprea and Magdy (2020). For this reason, we have several external observers evaluate the same responses in terms of the level of sarcasm in an effort to address the subjectivity of the topic. Last, we collect participant responses to questions addressing why sarcasm may have occurred in a situation in order to understand what underlying factors influence the use and perception of sarcasm.

Current Study

In this study, we investigate the following research questions:

1. Which factors trigger sarcastic responses between close friends?
2. What are the common grounds between intended sarcasm and perceived sarcasm?
3. Do speakers and listeners view sarcasm as a positive or negative tool when it is used in a friendly context?

To answer these questions, we devise two experiments: 1) a generation experiment and 2) a perception experiment.

Experiment 1: Generation Experiment

In Experiment 1, we elicit sarcastic responses from participants without explicit instructions to do so. Afterwards, we ask them to evaluate their own responses in terms of their intended level of sarcasm and other related factors.

Materials Thirty-two contextual prompts were created drawing on a qualitative analysis result from the MUStARD dataset (Castro et al., 2019). The analysis suggested that many sarcastic comments in close relationships appear in certain types of contexts *i.e. when a friend is being silly or annoying*. We hypothesized that these situational cues could potentially elicit sarcastic responses. Some of the dialogues from the dataset were used as a starting point to create new situation descriptions that could happen between any two close friends. In those situations, an imaginary best friend was behaving or talking in certain ways. Sixteen situations in which 'a friend was acting in a silly or annoying manner' were categorized as *non-neutral* stimuli. 16 plain and neutral situations - unlikely to trigger sarcastic answers in particular - were categorized as *neutral* stimuli. Taken together, they formed stimuli of two different *context types* (N = 32). Below are two example stimuli used in our experiment.

> ***non-neutral:*** *You and Steve have long been planning to go to a new bar in town. But, he has canceled on you three times without telling you why. And just now, he calls you and says, "I'm so sorry, but I'm gonna have to bail again. Next time?"*
>
> ***neutral:*** *Steve bought a really expensive pair of shoes as a treat to himself for having finished a big project at work. The shoes go very well with his outfit today.*

The interlocutor had the common male name "Steve", relying on previous findings that sarcasm happens more often either among male speakers or when directed at a male speaker

(Colston & Lee, 2004; Gibbs, 2000; Zhu & Wang, 2020). However, it should be noted that the gender of the directed speaker is not a separate condition in our experiment; it was merely used as a potential tool to increase the number of sarcastic responses since we expected that the proportion of sarcastic responses would be disproportionately low in general. We also expected that giving a name to the imaginary character would help participants become more immersed in the situations and respond more naturally.

Participants 60 participants (30 women and 30 men) were recruited on Prolific[1]. Any participant whose first language was English was eligible to participate. Participants received 8 GBP per hour as compensation. The experiment lasted 45 minutes on average.

Table 1: Intention options provided to the participants.

Intentions			
1	To criticize Steve in a harsher way	**2**	To criticize Steve in a softer way
3	To mock Steve in a hilarious way	**4**	To mock Steve in a friendly way
5	To say something natural	**6**	To be direct with Steve
7	To be nice to Steve	**8**	To say something clever

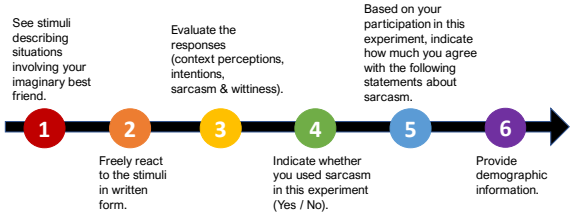

Figure 1: Experiment road map of Experiment 1. Experiment 2 excluded Step 2 and had stimuli format slightly modified in Steps 1 and 4.

Procedure The experiment following the road map in Figure 1 was conducted online using FindingFive[2]. The experiment was divided into four blocks. In the first block (Steps 1 and 2 in Figure 1), participants were asked to imagine being in a situation in which they are best friends with the imaginary character Steve. Participants saw 32 situations and reacted (in written form) to the situation descriptions or the last words of *Steve* contained in the description. In the second block (Step 3 in Figure 1), participants saw the same 32 situations with their responses and evaluated on a 6-point Likert scale (*not at all, mostly not, not so much, somewhat, mostly, completely*): 1) whether they found *Steve* silly or annoying *(context perception)*[3], 2) whether they thought their response was sarcastic *(sarc)*, and 3) whether they thought their response was witty *(witty)*. They also indicated their intentions as a multiple-choice response (multiple selections allowed; See Table 1). In the third block (4 and 5 in Figure 1), participants were asked to think back to the situations from the experiment in which they gave sarcastic responses and rate how much they agree with the given statements about sarcasm (See Table 2). They could skip this part if they thought they did not provide any sarcastic response throughout the whole experiment because we aimed to test those statements based on our experimental settings only. In the last block (6 in Figure 1), we collected demographic data including *gender* (*male, female, non-binary, prefer not to say*), *education level* (*high school diploma, undergraduate degree, graduate degree, PhD+*), and *general sarcasm use* in everyday life (6-point Likert scale). We collected gender information as previous research had suggested its effects on sarcasm use (Colston & Lee, 2004). We also assumed that the sophisticated pragmatic nature of sarcasm might be associated with the education level of the speaker. Lastly, we expected that the speaker's general sarcasm use would influence their self-evaluation criteria about sarcasm.

Table 2: Statements about sarcasm provided to the participants in the question form. In Experiment 2, the phrasing was slightly modified (e.g., "Do you think the speaker said the opposite of . . . ").

No.	Statements
1	Did you say the opposite of what you actually wanted to say?
2	Did you communicate your message in a more sophisticated way by being sarcastic?
3	Would your sarcastic reaction strengthen your bond with Steve?
4	Would Steve be offended by your sarcasm?
5	Would Steve think your response is humorous?

Data Processing The intentions were split into eight different variables and binary coded (0 / 1). The other ratings collected on a Likert scale were turned into 1 (*not at all*) through 6 (*completely*). The ratings for *sarc*(asm), *witty*, and *context perception* were z-transformed across items for each participant in order to control for the variability of each participant's sensitivity to providing ratings.

Results Initial data analysis using a linear mixed-effects model showed that *context perception* (lower or higher silliness/annoyance) was strongly associated with *context type* (neutral vs. non-neutral; $\beta = 1.00, p < 0.001$), which indicate that the experimental manipulation was successful. Due to this strong association, we only included context perception in our main analysis, which we deemed as more relevant to the sarcasm ratings than the original *context type*. We also excluded *witty* from our model because it was highly correlated with *sarc* ($r = 0.76$, $p < 0.001$), the focal point of our analysis.

We fitted a linear mixed effects regression model (Pinheiro & Bates, 2000) on the collected data. The model had the z-

[1]https://www.prolific.co/
[2]https://eu.findingfive.com/
[3]The predictor names used in subsequent analyses are in italics.

scored sarcasm ratings as dependent variable. As predictors, the model had z-scored *context perception* ratings (silly or annoying) interacting with 8 binary-coded *intentions* and the *order of stimuli presentation*, and main effects for the control variables *gender*, *education level*, and *general sarcasm use*.

A by-item random intercept was included for all the predictors that vary within items. A by-participant random intercept was not included as the continuous predictors were already z-scored by participant (grand mean = 0). As for random slopes, we excluded random slopes stepwise until it resulted in a converging model (Barr, Levy, Scheepers, & Tily, 2013). Our final model included a by-item random intercept and a by-participant random slope for *context perception*.

The model showed that 39% of the variance was explained by both the fixed and random factors (conditional $R^2 = 0.39$). We assessed the collinearity among the predictor variables by calculating the Variance inflation factors (VIFs; Zuur, Ieno, and Elphick (2010)). The VIFs for all variables were smaller than 4.68, indicating a moderate correlation among variables.

Table 3 reports the estimates and significance scores from the analyses. *Context perception* was a significant predictor for sarcasm ratings, where a higher perceived non-neutrality of a contextual prompt led to a higher self-reported sarcasm rating. The intentions to *mock hilariously*, to *mock in a friendly manner*, and to *be clever* also showed positive and statistically significant main effects on the sarcasm ratings. On the other hand, intentions to *be direct*, and to *be nice* as well as *general sarcasm use* had significant negative effects on the sarcasm ratings. No main effects of intentions to *criticize harsher*, to *criticize softer*, or to *be natural* were observed. The order in which the stimuli were presented did not affect sarcasm ratings, suggesting no transfer effect from using the same interlocutor name *Steve* across trials. Moreover, *gender* and *education level* did not show any statistical significance as reliable predictors of sarcasm. There were significant interactions between *context perception* and the intentions to *criticize softer*, to *mock hilariously*, and to *be direct*, which indicates that these intentions reduced the strong effects of *context perception* on the intended sarcasm ratings.

Experiment 2: Perception Experiment

In Experiment 2, we asked a different group of participants to evaluate the responses collected in Experiment 1 in terms of their perceived level of sarcasm and the presumed underlying intentions of the speakers. In order to account for the subjective nature of sarcasm (Oprea & Magdy, 2020), we assigned 6 evaluators to each participant in Experiment 1.

Instead of having participants take over the role of the target interlocutor, we asked them to be in the role of independent observers. In Dews and Winner (1995), the same patterns for irony perception were obtained regardless of whether the target of the remark was the addressee or a third person. We assumed that it would be more natural for our participants to be observers of the conversation than be the interlocutors in a conversation in which they never participated before.

Materials Thirty-two stimuli consisting of the context descriptions shown to the 60 participants in Experiment 1 and their responses were used as the materials for Experiment 2. The responses provided by the participants of Experiment 1 were spell-checked. The subject *you* in the stimuli from Experiment 1 was modified to *John* so that the participants would be reading conversations between *Steve* and *John*.

Participants A new group of 360 native English-speaking participants (180 women and 180 men) was recruited. Participants received 8 GBP per hour as compensation. The experiment lasted 30 minutes on average.

Procedure In the first block, participants saw conversations between *Steve* and *John* (Steps 1 and 3 in Figure 1). The 32 responses by each participant in Experiment 1 were shown to 6 participants in Experiment 2. Participants answered the same 4 questions from Experiment 1 about *John*'s responses: 1) whether they thought that John would have found *Steve* silly or annoying *(context perception)*, 2) sarcasm ratings of John's responses *(sarc)*, 3) wittiness ratings of John's responses *(witty)*, and 4) the assumed intentions behind John's remarks *(intentions)*. The following blocks had the same format as in Experiment 1 (Steps 4 - 6 in Figure 1).

Data Processing The ratings and the intentions were processed in the same way as in Experiment 1. The ratings provided by 6 observers on the same stimuli were z-transformed by observer and averaged for further analyses.

Results We fitted a linear mixed effects regression model on the obtained data. The model had the z-scored sarcasm ratings as dependent variable. As predictors, the model had z-scored *context perception* interacting with the 8 binary-coded *intentions*, and main effects for the control variables *gender*, *education level*, and *general sarcasm use* (See Experiment 1 for the full details of the categorical variables). Given that the data originated from a nested design, we included a random intercept by item, nested within each stimulus set assigned to each group of 6 observers. A by-participant random intercept was not included as continuous variables were already z-transformed by participant (grand mean = 0). We included a by-participant random slope for *context perception*. We did not include *witty* and the original *context type* in the model for the same reason described in Experiment 1. The model showed that 39% of the variance was explained by both the fixed and random factors (conditional $R^2 = 0.39$). Again, we observed no collinearity problems among the predictor variables as the VIFs for all variables were smaller than 2.92.

Similar to Experiment 1, *Context perception* was a significant predictor for sarcasm ratings, where a higher perceived non-neutrality of a context led observers to rate the response as more sarcastic. The intentions to *criticize softly*, *mock hilariously*, *mock in a friendly manner*, and *be clever* also showed significant positive main effects on the sarcasm ratings. Intentions to *be natural*, to *be direct*, and to *be nice* had significant negative effects on the sarcasm ratings. No main effects of *gender, education level, general sarcasm use*, or the intention to *criticize harsher* were observed. There were

Table 3: Lmer analysis results from Experiment 1 and Experiment 2. Dependent variable: sarcasm ratings z-scored per participant.†

	Exp 1		Exp 2	
Predictors	β		β	
(Intercept)	0.108		-0.036	
general sarcasm	-0.030	*	-0.006	
context perception (z)	0.239	***	0.343	***
be clever	0.227	***	0.348	***
be direct	-0.179	***	-0.254	***
be natural	-0.059		-0.152	***
be nice	-0.224	***	-0.122	***
criticize harsher	0.076		0.061	
criticize softer	0.052		0.138	***
mock friendly	0.760	***	0.526	***
mock hilariously	0.791	***	0.528	***
context:clever	-0.083		-0.017	
context:crit.harsher	-0.131		-0.051	
context:crit.softer	-0.129	*	-0.011	
context:direct	-0.127	**	-0.114	***
context:mock.friendly	-0.065		-0.109	***
context:mock.hilarious	-0.181	**	-0.142	***
context:natural	-0.001		-0.002	
context:nice	-0.070		0.013	
Conditional R^2	0.390		0.391	

*: p <0.05, **: p <0.01, ***: p <0.001

† Control variables that showed no significant effects (gender and education level) are not reported here.

significant interactions between *context perception* and intentions to *mock hilariously*, to *mock in a friendly way*, and to *be direct*, indicating that these intentions penalized the strong effects of *context perception* on perceived sarcasm ratings.

To test how much different raters agreed with one another, we inspected the inter-rater variability on the sarcasm ratings. We grouped the raters that saw the same stimuli and calculated the two-way intra-class correlation coefficient (ICC; Bartko (1966)) on their raw sarcasm ratings (as opposed to the z-transformed ratings). The average ICC score across all 360 observers was 0.75, indicating a 'moderate' to 'good' agreement (Koo & Li, 2016).

Comparison between Experiment 1 and 2

To directly compare the speaker's self-ratings with those of an external observer, we combined the data from the two experiments and ran another linear mixed effects model. The data was organized to include both the sarcasm ratings reported by the speakers (z-scored) and the average ratings by the observers (z-scored and averaged) in each row. Self-ratings and ratings by others were binary-coded as *generation* or *perception* for the predictor *source experiment*.

The linear mixed effects model had the z-scored sarcasm ratings as dependent variable and the z-scored *context perception* and binary-coded *intentions* as predictors, all interacting with *source experiment*. A random intercept by item nested within *source experiment* was added to the model. A by-participant random intercept was also added. Demographic variables such as *gender, general sarcasm use* or *education level* were not included because the ratings by the observers were averaged, losing their meaning in the combined dataset. The inclusion of *source experiment* in interaction with the 8 binarized *intentions* led to high collinearity (VIF = 12.8). We tested the potentially negative effect of collinearity by comparing the direction of the effect of each predictor and the corresponding p-value – elements traditionally affected by collinearity – against those from a simpler model without this interaction (VIF = 4.8). Overall, we observed that including the interaction between *source experiment* and the 8 *intentions* did not affect either the sign or the p-value of the other predictors. For this reason, we kept the interaction in our final model. The model showed that 44% of the variance was explained by both the fixed and random factors. As reported in Figure 2, we observed interactions between *source experiment* and intentions to criticize softer ($\beta = 0.19, p < 0.05$), be direct ($\beta = -0.25, p < 0.001$), and to be clever ($\beta = 0.48, p < 0.001$).

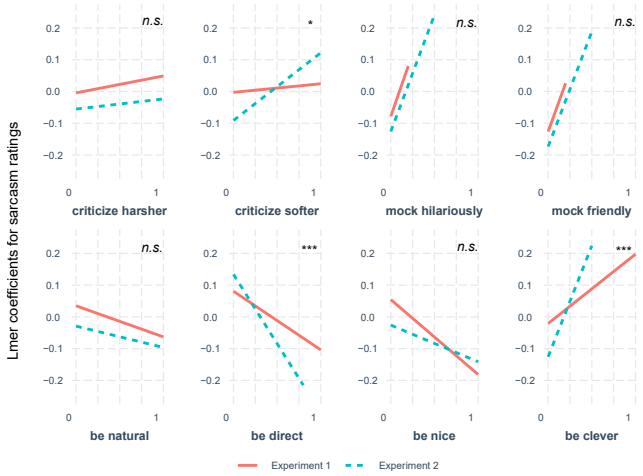

Figure 2: Lmer coefficients for the interaction between the source experiment and the eight intentions. Coefficients of the intentions from Experiment 1 are shown in red solid lines and from Experiment 2 in blue dashed lines. Significant interactions between the intentions and the source experiment are indicated by asterisks.

Results on sarcasm statements Results on participants' agreement with various statements about sarcasm suggest that both speakers and observers perceived sarcasm used in the given contexts to be a tool to be friendly and to enhance bonds between the interlocutors. Figure 3 summarizes the percentage of participants that agreed with the asked statements. The statements that most participants agreed with are aligned with Gibbs (2000) and Dews et al. (1995), who argued for the positive effects of sarcasm, rather than with Toplak and Katz

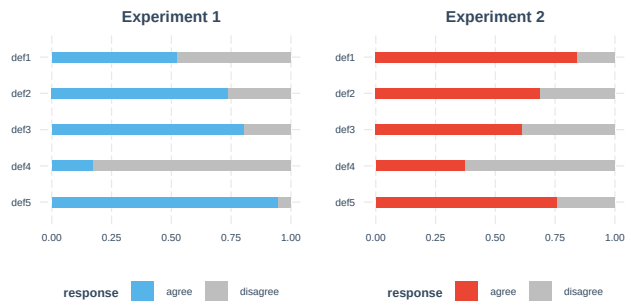

Figure 3: Participants' agreement to 5 statements about sarcasm described in Table 2, binary coded from the 6-point Likert scale (1 - 3: disagree, 4 - 6: agree). Left: from Experiment 1. Right: from Experiment 2. Shorthands for each definition are def1: opposite of intended meaning; def2: sophisticated; def3: strengthen bond; def4: offend Steve; def5: humorous. See Table 2 for the full definitions.

(2000) and Colston (1997), who argued for the aggressiveness of sarcastic comments. However, note that in our work, we limited the social context to friendly situations. It would be worth finding out whether people's perception of sarcasm changes in different social contexts.

General Discussion

In this study, we investigated which factors affect sarcasm production and whether independent observers can also perceive sarcasm and decode the factors that may have triggered it. From Experiment 1, we identifed three factors that trigger sarcasm (RQ1): perception of the addressee being silly or annoying, intention to mock the addresse, and to give clever remarks. We found that when the speakers find the addressee to be silly or annoying, it triggers more sarcastic responses. Results also showed that intentions to mock the addressee in a hilarious or friendly manner, or to give a clever remark increase the likelihood of sarcasm production. In contrast, the sarcasm level was judged to be lower when the speaker intended to be nice, direct, and natural toward the addressee. A higher sarcasm use in speaker's daily life predicted a lower sarcasm level for their responses, possibly because the higher use of sarcasm might cause a higher threshold for judging their response as sarcastic. Gender and education level of an individual did not affect the degree of sarcasm production.

Results from Experiment 2 showed that almost the same factors play a role in predicting sarcasm ratings by the observers (e.g., the addressee being silly or annoying, intentions to mock or to speak cleverly). For both the speaker and the observer alike, when the addressee was perceived to be silly or annoying, the presence of the intentions to mock the addressee hilariously or to be direct to them slightly decreased the strong effect of context perception on sarcasm ratings.

Some differences between the speakers and the observers were found. The intention to criticize the addressee softly was a strong predictor of sarcasm ratings by the observers but not by the speakers (See Figure 2). The intention to speak cleverly was a stronger predictor for sarcasm by observers than by speakers, though it was a reliable predictor for sarcasm in both experiments. No such interaction was found between the interlocutor role and the intention to mock the addressee, a communicative intent that is communicated more easily and naturally in close relationships. We construe that people are more likely to overinterpret the intentions behind sarcasm when they are on the perceiving end than on the producing end (e.g. "The speaker is probably using sarcasm to criticize in a softer way" or "The speaker is probably using sarcasm to speak cleverly"). This is especially true when the intentions behind sarcasm move toward a more criticizing end, where the decoding of the intent behind sarcasm differs by the role of the interlocutor (speaker vs. observer). These results are partially aligned with the findings from Bowes and Katz (2011) and Toplak and Katz (2000), that is, when people assume the role of a victim of sarcastic criticisms, they tend to feel more criticized than they do at direct criticisms.

Nevertheless, both speakers and observers viewed sarcasm used in the given contexts as positive (e.g., sophisticated, bond-enhancing, humorous) in general (RQ3, see Figure 3). Our finding that sarcasm is bond-enhancing is in line with Gibbs (2000), who found that verbal irony is used to enhance bonds between close friends. Furthermore, our finding that sarcasm conveys messages in a more sophisticated manner matches Jorgensen (1996), who identified a face-saving effect of sarcasm between friends, which is that one may use sarcasm in situations where direct criticisms would make them look inconsiderate and rude. Lastly, our finding that sarcasm conveys humor is in line with Dews et al. (1995), who argued for the face-saving and humor functions of irony. On the other hand, arguments by Colston (1997) or Toplak and Katz (2000) that sarcastic comments are more aggressive than their literal counterparts are not supported by our findings.

In general, we do not expect that such positive evaluations of sarcasm will always hold true; the definitions of sarcasm in people's minds may vary depending on the nature of the relationship between the speakers or the occasions in which sarcasm is used. One reason behind the positive attributions could be the assumption that it is more acceptable to mock and be mocked by a close friend. People may have perceived the word "criticize" as a bit too much in the context involving two best friends and replaced it with "mock", which may not be the case in other situations. We, therefore, argue that there is a definite need to address sarcasm occurring in different social contexts in future research for a more thorough understanding of sarcasm.

Acknowledgments

We thank Colin Davis, Henrike Bayer, Massimiliano Canzi, Andrea Ferreira, Melina Schneckenbühl, Qi Yu, and Mark Matthias-Zymla, for proofreading the stimuli and giving feedback to the pilot study that helped improve the design of our main experiments.

References

Attardo, S., Eisterhold, J., Hay, J., & Poggi, I. (2003). Multimodal markers of irony and sarcasm. *Humor - International Journal of Humor Research*, *16*(2).

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255–278.

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological reports*, *19*(1), 3–11.

Bowes, A., & Katz, A. (2011). When Sarcasm Stings. *Discourse Processes*, *48*(4), 215–236.

Bryant, G. A. (2010). Prosodic Contrasts in Ironic Speech. *Discourse Processes*, *47*(7), 545–566.

Castro, S., Hazarika, D., Pérez-Rosas, V., Zimmermann, R., Mihalcea, R., & Poria, S. (2019). Towards Multimodal Sarcasm Detection (An _obviously_ Perfect Paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4619–4629). Florence, Italy: Association for Computational Linguistics.

Colston, H. L. (1997). "I've Never Seen Anything Like It": Overstatement, Understatement, and Irony. *Metaphor and Symbol*, *12*(1), 43–58.

Colston, H. L. (2000). On necessary conditions for verbal irony comprehension. *Pragmatics & Cognition*, *8*(2), 277–324.

Colston, H. L., & Lee, S. Y. (2004). Gender Differences in Verbal Irony Use. *Metaphor and Symbol*, *19*(4), 289–306.

Das, S., & Kolya, A. K. (2021). Parallel Deep Learning-Driven Sarcasm Detection from Pop Culture Text and English Humor Literature. In I. Pan, A. Mukherjee, & V. Piuri (Eds.), *Proceedings of Research and Applications in Artificial Intelligence* (Vol. 1355, pp. 63–73). Singapore: Springer Singapore.

Dews, S., Kaplan, J., & Winner, E. (1995). Why not say it directly? The social functions of irony. *Discourse Processes*, *19*(3), 347–367.

Dews, S., & Winner, E. (1995). Muting the Meaning A Social Function of Irony. *Metaphor and Symbolic Activity*, *10*(1), 3–19.

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.

Ghosh, D., Fabbri, A. R., & Muresan, S. (2018). Sarcasm Analysis Using Conversation Context. *Computational Linguistics*, *44*(4), 755–792.

Gibbs, R. W. (2000). Irony in Talk Among Friends. *Metaphor and Symbol*, *15*(1-2), 5–27.

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

Joshi, A., Bhattacharyya, P., & Carman, M. J. (2017). Automatic Sarcasm Detection: A Survey. *ACM Computing Surveys*, *50*(5), 1–22.

Keenan, T. R., & Quigley, K. (1999). Do young children use echoic information in their comprehension of sarcastic speech? a test of echoic mention theory. *British Journal of Developmental Psychology*, *17*(1), 83–96.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155-163.

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.

Lee, C. J., & Katz, A. N. (1998). The Differential Role of Ridicule in Sarcasm and Irony. *Metaphor and Symbol*, *13*(1), 1–15.

Leggitt, J. S., & Gibbs, R. W. (2000). Emotional Reactions to Verbal Irony. *Discourse Processes*, *29*(1), 1–24.

Matthews, J. K., Hancock, J. T., & Dunham, P. J. (2006). The Roles of Politeness and Humor in the Asymmetry of Affect in Verbal Irony. *Discourse Processes*, *41*(1), 3–24.

Oprea, S., & Magdy, W. (2019). Exploring Author Context for Detecting Intended vs Perceived Sarcasm. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 2854–2859). Florence, Italy: Association for Computational Linguistics.

Oprea, S., & Magdy, W. (2020). iSarcasm: A Dataset of Intended Sarcasm. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Online: Association for Computational Linguistics.

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.

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

Pinheiro, J. C., & Bates, D. M. (2000). Linear mixed-effects models: basic concepts and examples. *Mixed-effects models in S and S-Plus*, 3–56.

Sperber, D. (1984). Verbal Irony: Pretense or Echoic Mention? *Journal of Experimental Psychology*, *113*, 7.

Toplak, M., & Katz, A. N. (2000). On the uses of sarcastic irony. *Journal of Pragmatics*, *32*(10), 1467–1488.

Woodland, J., & Voyer, D. (2011). Context and Intonation in the Perception of Sarcasm. *Metaphor and Symbol*, *26*(3), 227–239.

Zhu, N., & Wang, Z. (2020). The paradox of sarcasm: Theory of mind and sarcasm use in adults. *Personality and Individual Differences*, *163*, 110035.

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in ecology and evolution*, *1*(1), 3–14.