Irony Man: Augmenting a Social Robot with the Ability to Use Irony in Multimodal Communication with Humans

Socially Interactive Agents Track

Hannes Ritschel, Ilhan Aslan, David Sedlbauer and Elisabeth André
Human-Centered Multimedia, Augsburg University
Augsburg, Germany
{ritschel,aslan,andre}@hcm-lab.de

ABSTRACT

Interpersonal communication is often full of irony and irony related humor, which can shape the quality of a conversation and how conversation partners perceive each other. If social robots were able to integrate irony in their communication style, their human conversation partners might perceive them as more natural, credible, and ultimately more attractive and acceptable. In order to explore this assumption, we first describe an approach to transform non-ironic inputs on-the-fly into multimodal ironic utterances. Irony markers are used to adapt language, prosody and facial expression. We argue that doing this allows to dynamically enrich a robot's spoken language with an expression of socially intelligent behavior. We then demonstrate the feasibility of our approach by reporting on a user study, which compares an ironic version of a robot with a non-ironic version of the same robot in a small talk dialog scenario. Results show that participants are indeed able to correctly identify a robot's use of irony and that a better user experience is associated with an ironic robot version. This is an important step for dynamically shaping a robot's personality and humor, and to increase perceived social intelligence.

ACM Reference Format:
Hannes Ritschel, Ilhan Aslan, David Sedlbauer and Elisabeth André. 2019. Irony Man: Augmenting a Social Robot with the Ability to Use Irony in Multimodal Communication with Humans. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 9 pages.

1 INTRODUCTION

Humor is something peculiar that we as humans seem to experience as part of our daily activities and conversations. Some people are known for their great sense of humor and ability to use words in funny but also smart ways. A paradigmatic example is Albert Einstein, who is known for his quirky quotes, such as "It's not that I'm smart, it's just that I stay with problems longer". Without doubt Einstein seems a likable and credible character and it is likely that his ability to use humor contributes to why people like him. It seems obvious that humor is an important quality of human conversation. People use it for the purpose of entertainment, but also to regulate conversations, to ease communication problems or to cope with critique or even stress. While canned jokes are the first type of humor which comes to mind, *conversational* humor [9] often emerges out of the moment, depending on the context and dialog. It can be found in different forms, including irony, which can but does not necessarily be amusing. Humans use such stylistic devices deliberately and also appreciate the unique communication style of our conversation partners' character. However, the use of conversational humor in the context of human-computer and human-robot interaction is by no means the norm.

While humor increases interpersonal attraction and trust in interpersonal communication, conversational agents, including social robots, can benefit from it in the form of more natural and enjoyable interactions as well as increased credibility and acceptance [27]. For example, findings confirm that jokes are perceived funnier when told by a robot compared to their text-only equivalents Sjöbergh and Araki [34] and also that there is an interaction effect between user personality and preferred type of humor Mirnig et al. [23]. Several research experiments explore robots as entertainers in the context of stand-up comedy and joke telling, which also adapt the performance based on an audience's feedback. The actual contents are scripted and prepared in advance.

In order to enrich human-robot dialog with conversational humor, scripted contents might not always be sufficient. Keeping the diversity of interaction contexts, tasks and human preferences in mind, social robots should not only express humor, but also be able to generate it dynamically. Apart from the dialog context, which has to be taken into account to generate suitable language, the robot's non-verbal behaviors also have to be adjusted in an appropriate manner. As outlined by Mirnig et al. [24], adding humorous elements to non-humorous robot behavior alone does not automatically result in increased perceived funniness. Appropriate prosody, facial expressions and gestures are a crucial element in any generation process.

One of the contextual challenges for generating and using dynamically generated humor is appropriateness. There can be situations where the use of irony is inappropriate and not funny. It will yield misunderstanding when humor is used in the wrong situation [27]. A rather uncomplicated conversation situation with accepted conventions but also blurry limits to what is allowed and what isn't is small talk dialog. A small talk's main aim seems entertaining each other without necessarily talking about controversial content. An additional goal in small talk is often to break down barriers. Thus it seems to be a good scenario to study the use of irony and irony induced humor in conversations between a human and a robot.

In order to explore the effect of irony in a robot's verbal and non-verbal behaviors, we have implemented an approach to transform

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

non-ironic sentences into ironic utterances, supplemented by multimodal irony markers from the literature. Apart from the linguistic markers in the text, the robot employs tailored prosody and facial expressions to make it easier to identify its expressed behaviors as irony. It can be embedded into dialog scenarios to augment the robot's language with conversational humor in order to shape the robot's character, personality and to increase perceived social intelligence. The results of an evaluation in a small talk dialog scenario demonstrate that the robot's dynamically generated multimodal expression of irony was successfully identified by the participants. Beyond that, the ironic robot was assigned better user experience and was generally preferred in comparison to a non-ironic robot.

2 RELATED WORK

The desire to create original contents, as well as to augment social agents with the ability to include humor in their expressed behaviors, has motivated researchers to examine techniques for computational generation of verbal humor for several decades now. Experiments for generating humor in the form of text include for example the "Light Bulb Joke Generator" [30], "JAPE" and "STANDUP" for punning riddles [4, 38], "HAHACRONYM" [36] for humorous acronyms or generation of lyrics parodies [13].

With respect to conversational humor, Dybala et al. [8] embed a joke generator into a non-task oriented conversational system. Results of their studies show that the system was evaluated as higher/better when it used humor, compared to when it didn't. They also mention the issue of appropriateness and that one aspect for improving timing is to take the human's emotional state into account. Nijholt [27] uses misunderstandings in the context of a chatbot to generate humor by employing erroneous anaphora resolution. While he takes short two-part discourse scripts as a basis, contrast or opposition between the interpretation of both parts is required to result in humor. In general, Morkes et al. [25] have shown that the presence of humor can enhance task-oriented interactions independently on whether the conversation partner is a human or a computer.

Valitutti and Veale [37] use a Twitter bot to produce verbal irony. The content itself is generated by creating contrasting phrases, combined with quotes and the hashtag *#irony*. Their experiments show that quotes and contrast in the polarity of the phrases help communicate the presence of irony while the hashtag is less suitable to strengthen the effect of irony.

Several works investigate how irony is communicated in text and (non-)verbal communication (see Section 3.3). For example, Carvalho et al. [6] and Frenda [11] explore and model linguistic cues in text corpora of user-generated web contents. Recent irony detection tasks cover English [17] and Italian [7] Twitter messages. Attardo et al. [1, 2] and Williams et al. [40] look at facial expression, prosody and gestures observed in human-human interactions. By transferring these markers to the robot, these insights provide a good basis for the synthesis of multimodal, ironic behaviors.

2.1 Robots Presenting Humor

In the context of Human-Robot Interaction (HRI), several experiments investigate the effect of humor as well as how to present it. First of all, for the purpose of entertainment, the goal is to explicitly make the user laugh. Stand-up comedy [28] is one of the use cases which addresses a larger audience. For example, research by Knight [20] uses a NAO robot to present jokes and adapt the show to the audience at the same time. While the actual contents are scripted, spectators can actively influence the performance by their laughter, applause, direct prompts (i.e., likes or dislikes) and by holding up green or red cards. Based on the current estimated audience's enjoyment level, the robot selects the next jokes dynamically. For this task, every joke is associated with several attributes.

Similarly, Katevas et al. [19] use the RoboThespian™ platform for standup comedy with special regard to non-verbal behaviors in terms of gaze and gestures. They optimize the joke delivery in real-time by reacting to the audience. The robot can explicitly address individual visitors by looking into their face or responding to them. For this purpose, every scripted comedy text includes positive or negative responses which can be presented by the robot, depending on whether laughter has been observed or not.

Hayashi et al. [16] present a Japanese "Manzai" comedy dialog conversation with two robots. They utilize two Robovie robots capable of coordinating their communication with each other and reacting to external stimuli from the audience. The authors also estimate the audience's amusement reactions in terms of clapping and laughing to adjust speech and motion timing of the performance.

In contrast to these scenarios with larger audiences, Weber et al. [39] learn joke preferences for a single person, including scripted jokes as well as combinations with sounds and grimaces. The robot's performance is optimized based on human social signals. Instead of relying on explicit feedback, they use laughter and smile exclusively.

A positive effect of humor in HRI is confirmed by Niculescu et al. [26], which explore the relationship between voice characteristics, language cues (including empathy and humor) and the perceived quality of interaction. Their results show that the robot's use of humor improves the perceived task enjoyment and that the voice pitch impacts the user's perceived overall interaction quality and overall enjoyment. Furthermore, research by Mirnig et al. [23] comes to the conclusion that positively attributed forms of humor (i.e., self-irony) are rated significantly higher than negative ones (i.e., schadenfreude) when it comes to robot likability.

2.2 Robot Embodiment and Multimodal Cues

The robot's embodiment is a crucial point as it opens up the possibility to deliver multimodal humorous contents. Sjöbergh and Araki [34] evaluate the difference of perceived funniness of jokes which are either presented in text form or by a robot. Their results show that the presentation method has a significant impact: participants rated jokes significantly funnier when they were presented by the robot, compared to their text-only equivalents.

Mirnig et al. [24] confirm that the modality of how humor is presented plays an important role. Adding unimodal verbal or non-verbal, humorous elements to non-humorous robot behavior does not automatically result in increased perceived funniness. The authors point out that humor is multilayered and that several modalities have to be combined to create humorous elements. Especially in the context of irony, multimodal cues, such as facial expression and prosody, play an important role for helping the listener to identify spoken words as irony (see Table 1 and Section 3.3).

Table 1: Multimodal irony markers

Modality	Markers
Language	Exaggeration and understatement [1], positive and negative interjections, onomatopoeic expressions for laughter [6, 11], quotation and heavy punctuation marks [1, 6, 11], dots [1], hashtags [7, 17, 37], emojis [7, 17]
Facial expr.	Gaze aversion (saccade) [40], wink [1, 2], rolling eyes, wide open eyes and smiling [2]
Prosody	Intonation and nasalization [1, 2], stress patterns [1], speech rate, extra-long pauses and exaggerated intonational patterns [2]
Gestures	Nudges [1]

2.3 Contribution

When used sparingly and carefully, humor can help social agents to solve communication problems and to increase acceptance [27], such as in the context of natural language interfaces [3]. Thus, the generation of conversational humor for social robots is of particular interest. In contrast to the applications from the entertainment domain outlined in Section 2.1, canned texts might probably be not sufficient in the context of conversational humor. Recent works on emotional conversation generation [42] and automatic dialog generation with expressed emotions [18] in the context of Natural Language Generation (NLG) use artificial neural networks to generate emotionally consistent responses. Our presented approach addresses a different problem, i.e. the transformation of a given response into a multimodal (in this case ironic) version based on rule-based Natural Language Processing (NLP) and NLG techniques. It augments the robot with the ability to create humorous contents on-the-fly, including linguistic, as well as prosodic markers for the Text-To-Speech (TTS) system and typical facial expressions.

3 MULTIMODAL GENERATION AND EXPRESSION OF IRONY

This work focuses on *verbal* irony, which has been traditionally seen as "a figure of speech which communicates the opposite of what was literally said" [41]. As Attardo [1] points out, irony is made up of both irony *factors* and *markers*. The former are essential since they affect the actual meaning of the utterance when saying one thing while meaning another. Removing this factor destroys the irony. The latter emphasize and support the effect of the applied factor and help the recipient to identify the resulting utterance as irony, e.g. by modifying the prosody. Table 1 summarizes common multimodal markers for irony from the literature in terms of language, facial expression, prosody and gestures.

The presented transformation process for creating an ironic version of a non-ironic utterance involves three steps (see Figure 1): (1) NLP checks whether it is possible to apply the transformation at all (2) NLG generates the irony factor and inserts linguistic markers (3) addition of prosodic cues which indicate irony as well as generation of accompanying non-verbal behaviors like gestures and/or facial expressions. While steps one and two are applicable for any robot since they only manipulate the text itself, multimodal cues depend on and are restricted by the robot's soft- and hardware. Not every TTS system supports all proposed manipulation tags, facial expressions and gestures require corresponding hardware actuators or a virtual representation on a screen.

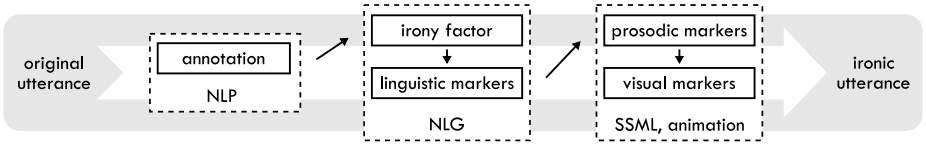
3.1 Natural Language Processing

The generator receives an arbitrary sentence in the form of text as input. First of all, NLP techniques are applied to make sure that the input actually can be transformed into an ironic utterance with the proposed approach. The input needs to have a notion of polarity so that the actual meaning can be inverted. CoreNLP [21] is used to identify adjectives, nouns and verbs with polarity based on sentiment analysis [35]. For example, in the sentence "I hate my worst enemy." the words "hate", "worst" and "enemy" indicate a negative polarity. These words are marked as candidates for creation of the irony factor in the following step, which makes up the incongruity in the resulting sentence's meaning.

3.2 Natural Language Generation

3.2.1 Irony factor. Based on the concept of *ideational reversal irony*, where "the intended meaning arises as a result of negation of a chosen element of the literally expressed meaning or the pragmatic import of the entire utterance" [10], the transformation creates an irony factor by replacing a polarized word with its antonym. For each candidate identified in the first step, antonyms are looked up in the WordNet database [22]. If a suitable antonym is found, the original word is replaced. However, only one single word is replaced in order to not nullify the negation. This could easily result from multiple replacements at the same time (e.g. "I love my best friend.") and would prevent the emergence of irony. Replacement is prioritized as follows: (1) verbs (2) adjectives (3) nouns. The flexion of words is realized with SimpleNLG [12] while preserving the conjugation of verbs, comparative and superlative of adjectives and number of nouns. Since the meaning of the original sentence can also be inverted by negating the main verb with "not", this strategy is applied if no suitable antonym can be found. Additionally, if the original sentence already contains a negated verb, "not" is removed to create the irony factor. In this case, there is no need to search for antonyms. The replacement step transforms the former example into the sentence "I *love* my worst enemy."

3.2.2 Linguistic markers. The replacement of single words does not automatically make the ironic intention recognizable without further ado. Several linguistic irony markers were identified by Attardo [1] and Carvalho et al. [6], including exaggeration (e.g. "really", "utterly"), understatement (e.g. "barely", "almost"), as well as positive and negative interjections (e.g. "Great!", "Super!", "Damn it!", etc.). Interjections occur in subjective texts to express the author's emotions, feelings and attitudes [6]. For example, the former negatively polarized example can be prefixed with a positive interjection to generate "*Great!* I love my worst enemy." Exaggerations and understatements can be realized by valence shifting [14]. Single words are modified by adding, removing or replacing of adjectives and adverbs to strengthen or weaken the sentence's meaning, e.g. resulting in "I *absolutely* love my worst enemy." This increases or decreases the intensity of the irony factor.




```
if is_polarized(input_utterance):
  factor = determine_irony_factor(input_utterance)
  utterance = replace_with_antonym(input_utterance, factor)
  utterance = insert_linguistic_markers(utterance)
  utterance_with_ssml = create_prosodic_marker(utterance)
  facial_expression = add_random_animation([SMILE,
        ROLL_EYES, WINK, EYES_WIDE_OPEN, SACCADE])
  start_animation(facial_expression)
  speak(utterance_with_ssml)
  stop_animation()
```

Figure 1: Overview of and simplified pseudo-code for the transformation approach.

Furthermore, onomatopoeic expressions for laughter (e.g. "haha"), acronyms like "lol" or "rofl", emoticons (e.g. ";-)"), quotation and heavy punctuation marks (e.g. "!!!") as well as dots ("…") are used in written language. Keeping in mind that the output medium for the generated ironic text is the robot's TTS system, these text-only markers cannot be applied. Instead, the robot uses prosody, facial expression and gestures depending on the available actuators and TTS capabilities, which supersedes the imitation in form of text-only markers. In contrast to the irony factor, multiple markers can be applied at the same time to emphasize the use of irony, e.g. "*Great!* I *absolutely* love my worst enemy."

3.3 Multimodal cues

While written text may use direct, typographic or morpho-syntactic markers to help the reader identify ironic content, paralinguistic and visual clues are of special interest to support and complement the linguistic irony factor. Depending on the robot's hardware and TTS software, intonation and visual clues in terms of prosody and facial expression are of special interest.

3.3.1 Prosody. Two acoustic parameter modulations from Attardo et al. [2] allow to generate typical prosody for ironic utterances: *compressed pitch pattern* and *pronounced pitch accents*. These atypical speaking behaviors contrast normal speech modulations in terms of pitch, rhythm, speech rate and accents and can be imitated with Speech Synthesis Markup Language (SSML)[1]. The compressed pitch pattern is characterized by a "flat" intonation, causing very little pitch movement while pronouncing the utterance. As shown in Listing 1, the *x-soft* prosody variant is used to prevent emphasis and emotion in the resulting audio. In contrast, pronounced pitch accents exaggerate the intonation by accentuating words throughout the whole sentence, certain words or multiple syllables of the same word. Often, they are combined with elongation and stilted pauses. Listing 2 illustrates the generated markup, which reduces the overall speech rate and utilizes both *emphasis* tags for the main verb, all adjectives, adverbs and nouns as well as *break* tags to accentuate these even more. For the example above, this results in putting emphasis on "I *absolutely love* my *worst enemy*." Interjections are emphasized for both patterns independently of the prosodic marker to highlight their emotional intensity.

3.3.2 Facial expression. In addition to prosodic markers, the physical embodiment of a robot allows to employ facial expressions for underlining the ironic intention. Typical cues include raised or lowered eyebrows, wide open eyes, squinting or rolling, winking, smiling or a so-called *blank face*, which is perceived as "expressionless", "emotionless" and "motionless" [2]. In addition, gaze aversion is a typical cue accompanying sarcastic statements [40]. Depending on the robot's face and actuators, these markers can be applied by animating the whole head, mouth, eyes, eye lids and eyebrows. Since the Reeti robot does not have eyebrows, the corresponding markers cannot be applied. Figure 2 illustrates the markers used during evaluation (see Section 4). These are shown for the duration of the ironic text, the robot returns to its neutral pose afterwards.

[1] https://www.w3.org/TR/speech-synthesis/

Listing 1: SSML for compressed pitch pattern

```
<emphasis level="strong">Great!</emphasis>
<prosody volume="x-soft">
  <emphasis level="none">I absolutely love my worst enemy.</emphasis>
</prosody>
```

Listing 2: SSML for pronounced pitch accents

```
<prosody rate="x-slow">
  <emphasis level="strong">Great!</emphasis> I <emphasis level="strong">
    absolutely<break strength="medium"/> love <break strength="medium"/>
  </emphasis> my <emphasis level="strong">
    worst <break strength="medium"/> enemy<break strength="medium"/>
  </emphasis>.
</prosody>
```

Figure 2: Facial markers

3.4 Further Markers and Restrictions

Realization of prosodic and visual markers heavily depends on the robot's hard- and software. Prosodic features encoded with SSML require a compatible TTS system while the strength of the audible effect may vary depending on the internal implementation and voice. Not all markers can be realized with SSML, e.g. *nasalization* [2]. Similarly, facial expressions and gestures (e.g. *nudge*) are restricted by the available motors and degrees of freedom. The *blank face* may be used if the robot's face is permanently animated. As long as the robot's face is primarily static, the marker does not show a significant difference to the non-animated face. For prosody, the *strong within-statement contrast* marker has not been implemented since the division into high-pitch and low-pitch part is non-trivial.

Ambiguity is a challenge for both checking the polarity of words as well as antonym lookup when generating the irony factor. For example, instead of marking it as positive polarized word, "great" might be classified as neutral since it is associated with more than one synset, including "big". Similarly, "great job" could be transformed into "small job" instead of "bad job". Knowledge about the

word's context could improve the antonym lookup with respect to semantic relatedness [29].

4 EVALUATION

In order to evaluate the performance and the effects of the previously described irony generation approach in a multimodal human communication setting we conducted an empirical study with users in the lab. Moreover, the study aimed to explore (1) whether participants would be able to identify the generated behavior as ironic and (2) how the generated behavior would influence participants' user experience. Furthermore, our hypothesis was that an ironic robot would be considered more "fun" to communicate with and "entertaining" to interact with, and thus, an ironic robot would receive higher user ratings associated with hedonic qualities than a non-ironic robot. Since both (i) the context and topic of the conversation play an important role for the appropriateness of irony usage and (ii) related research has shown that a shared sense of humor has a powerful effect on interpersonal attraction when people meet each other for the first time [5], we have decided to focus on a small talk scenario for the evaluation. Furthermore, we believed that through small talk barriers can be broke down, and thus the use of irony would be acceptable.

4.1 Participants, Apparatus, and Procedure

Twelve participants (6 male, 6 female) aged from 19 to 32 (avg. 25) were recruited from a university campus. The study design was a within-subject design: each participant interacted with all versions of the robot (i.e., a baseline robot and an ironic robot) in a counter-balanced order. In order to minimize potential carry over effects we further chose to use two Reeti robots instead of reusing the same physical robot. We did this to clarify that participants were interacting with two different robots (or personalities). Our decision to use two robots was also informed by previous work [39] reporting on a carry over effect in a within-subject study, which utilized the same physical instance of a social robot.

The study procedure consisted of four subsequent parts. First, participants were welcomed and provided with an introduction, which included a collection of demographic data, a short description of the study and additional information about the further study procedure. Participants were told that they would be asked to conduct dialogs for approximately ten minutes with two robots of the same type, which were differently programmed and afterwards we would ask them user experience questions about how they perceived the robots. We did not tell them that one of the robots uses irony or that the study was about irony or humor.

The information we presented to the participants pointed out that they would be able to simply speak with the robot, but we wouldn't use an automatic speech recognizer, instead the experimenter would enter their statements via keyboard and that therefore short delays in the robot's answer could occur. Furthermore, participants were asked to both reply to the robot's questions and to ask questions to the robot themselves whenever they wanted to or felt it was appropriate to do so.

At the beginning of both sessions, the participant was placed in front of one of the two Reeti robots. Both robots used the Cerevoice[2]

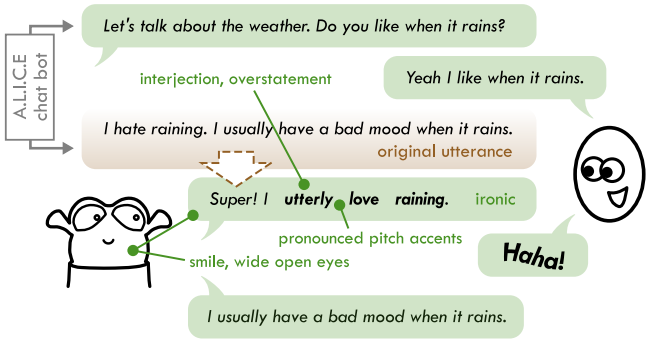


Figure 3: Sample of the robot's produced irony during the evaluation

Figure 4: Basic emotions (neutral, happy, sad) used independently on the irony condition

TTS system with the "William" voice and identical animations. The robot's answers were generated by the open-source A.L.I.C.E chat bot[3], which used freely available corpora for small talk. In average, the chatbot generated 19.25 answers per session. Each session started with the robot greeting the participant and asking him or her for the name. At any point in time, participants could see what the experimenter entered via keyboard on the screen next to the robot. We did this to make it transparent that the experimenter was not acting as a "wizard" and entering the robot's answers but in fact typing the participants' replies and questions.

In the sessions with the *ironic* robot condition, the robot's answers were computationally – whenever applicable – transformed into an ironic utterance with multimodal cues of irony (see Figure 3). In average, 4.33 ironic answers were produced per session. In contrast, in the *neutral* (or baseline) condition, the robot's answers were never manipulated in order to turn them ironic. In both conditions, basic emotion postures (see Figure 4) were added to the non-ironic spoken texts according to the robot answer's polarity.

At the end of each dialog session participants were asked to fill out a questionnaire measuring their experience. Half of the participants started with the ironic condition, the other half with the neutral condition. Similarly, half of the subjects started with the left robot and half with the right robot, independent of the condition. A sample dialog with both non-ironic and ironic contents is presented in Figure 8.

At the end of all sessions, participants were also asked which of the two versions of the robot they preferred, as well as whether they liked/disliked the voice and visual appearance of the robot. We did this to collect explicit robot preferences and to measure if the robots appearance and voice was in general acceptable.

[2]https://www.cereproc.com/en/products/sdk

[3]https://sourceforge.net/projects/alicebot/

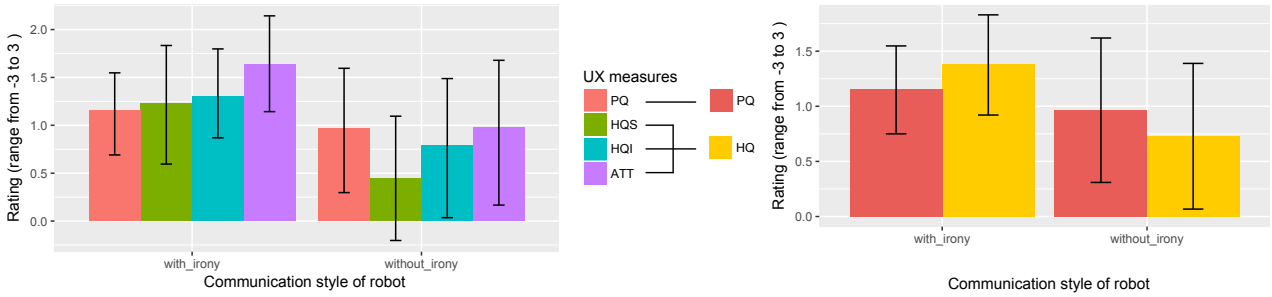


Figure 5: Mean ratings for the measured UX constructs PQ, HQS, HQI, and ATT explaining pragmatic (i.e., perceived traditional usablity) and hedonic qualities for both the ironic robot and the baseline robot considering. Error bars denote 95% confidence intervals.

4.1.1 Questionnaires. We chose to use the standardized attrakDiff [15] questionnaire for measuring UX, because it is widely used in research and industry, providing an overview of a product's or technique's perceived qualities, especially hedonic qualities beyond traditional usability, which we believed to be tightly connected with irony and the use of irony in communication. The attrakDiff questionnaire consists of 28 items, seven of those items are used to measure pragmatic quality (PQ), which is a measure for perceived traditional usability. The rest of the items measure hedonic quality (HQ), which results from a combination of HQS, HQI, and ATT. Each of these (sub)constructs of hedonic quality are measured by seven items of the attrakDiff. HQS measures the perceived ability of a product to meet a person's desire for self-improvement, HQI measures the perceived ability of a product to communicate a valuable identity to others, and ATT measures overall attractiveness.

In addition, we asked for agreement scores on a five-point Likert scale for statements, such as "the robot's output was ironic", which were specific to our user study setup and aimed to measure participants' subjective impression of the robot's output (replies and questions) in terms of content, naturalness, humor, irony, fitting facial expressions and voice, and perceived (social) intelligence (see Figure 7).

4.2 Results

In this section, we aim to provide answers to the user study's aforementioned research questions and results of testing our hypothesis (i.e., that an ironic robot based on our irony generation approach will receive higher user ratings on the hedonic dimension of UX). In order to present results in a structured manner, we start by presenting general "trends" based on graphical presentations of the collected UX data. Then, results will be interpreted based on fitting a statistical model to the data (i.e., results of statistical test will be provided). Afterwards, we describe the results of the additionally collected explicit robot preferences and agreement scores for the user study specific statement.

4.3 General Trends

The left side of Figure 5 presents the mean values for all four measured UX constructs (i.e., PQ, HQS, HQI and ATT). It seems that the ironic robot received higher mean ratings in all measured constructs.

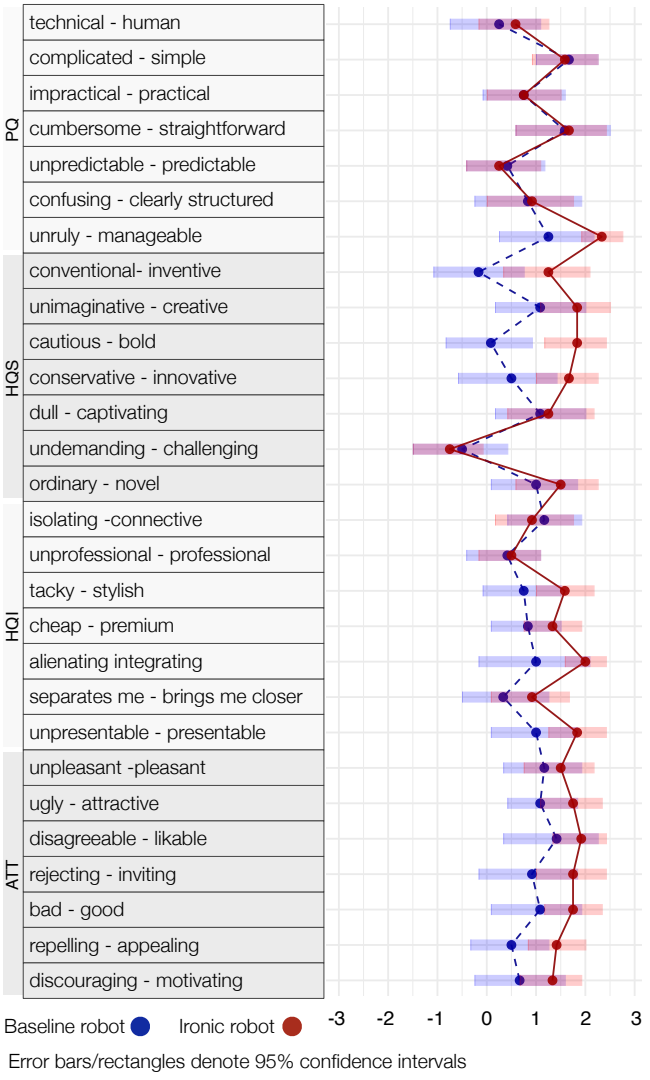


Figure 6: Details showing the mean ratings for each item of the attrakDiff questionnaire.

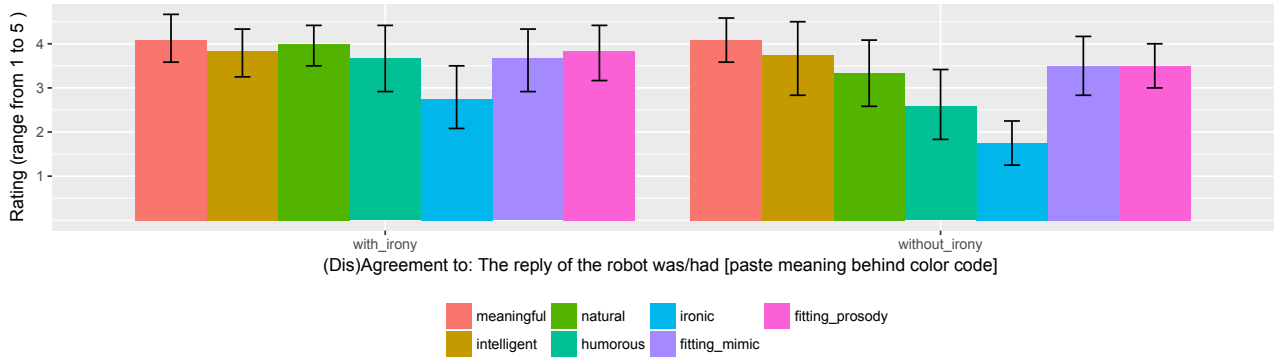


Figure 7: Mean ratings for the agreement scores to all additional statements (which were translated and slightly paraphrased to fit the plot) on a five-point Likert scale (1 = strongly disagree, 5 = strongly agree). Error bars denote 95% confidence intervals.

The differences in how the ironic robot was perceived compared to the baseline robot seem more in constructs explaining hedonic qualities, which is what we had hypothesized. The biggest differences of the means seem in ATT (overall perceived attractivness) and HQS (i.e., the hedonic quality that is associated with the perceived ability of a product/technique to meet a person's desire for self-improvement) and the smallest difference seems to exist in PQ (i.e., perceived traditional usability).

The right side of Figure 5 shows the difference in how the ironic robot is perceived different from the baseline (non-ironic) robot considering pragmatic quality (PQ) and hedonic quality (HQ) (i.e., combination/aggregation of HQS, HQI, and ATT). While we can observe a difference in HQ between both robots the difference between the baseline robot and the ironic robot considering pragmatic quality (PQ) seems very small.

Figure 6 depicts for each construct mean values for all constituting seven items, detailing how participants experienced the interaction with both robots. Considering HQS, the ironic robot was experienced, as one might have expected, as more bold, but also as more innovative. It seems that the ironic robot received higher ratings for most items explaining the modalities' hedonic quality, such as being perceived as more stylish, appealing, or likable, and thus potentially communicating a valuable identity to others (HQI) and generally being perceived as attractive (ATT). The overall difference seems systematic with the ironic robot having been perceived as consistently more "desirable" (i.e., higher ratings on the hedonic dimension of UX while there seem to be no clear differences on the pragmatic dimension of UX).

Figure 7 depicts the mean ratings explaining (dis)agreement of study participants with statements probing for example if they recognized irony and their opinion of if the mimic and prosody fitted the replies robots provided. The main differences in mean ratings seem to exist for the statements about robots being ironic, humorous, and natural, which indicates that participants were able to identify the use of irony and experienced humor and naturalness as associated causes of irony use.

4.4 Statistical Analysis

We first compare participants' ratings for PQ, HQS, HQI and ATT for both modalities, conducting paired-samples t-tests. For HQ and its constituting constructs HQS, HQI, ATT we used a one-sided test since our hypothesis was that the ironic robot would receive higher ratings on hedonic quality compared to the baseline robot. Because we did not have a hypothesis on how irony would affect PQ we use a two-sided test for testing the significance of users' self-reports considering PQ in both robot conditions.

We found a main effect of irony on overall hedonic quality HQ ($t=1.86$, $p=.044$, $r=.49$); that is, as hypothesized participants found the ironic robot as more "desirable" and the difference in self-reports on the hedonic dimension of UX was significant. When we look at the specific (sub)constructs of HQ separately, we find that the difference in HQS ($t=2.29$, $p=.021$, $r=.56$) is significant. But differences are non-significant for HQI ($t=1.27$, $p=.115$, $r=.35$), and in ATT ($t=1.44$, $p=.089$, $r=.39$). There was no significant difference in PQ ($t=-.67$, $p=.516$, $r=.19$).

We conducted similar statistical tests, considering the scores for agreement presented in Figure 7. We found significant differences for statements about the robots ironic ($t=2.34$, $p=.019$, $r=.57$) and humorous ($t=1.85$, $p=.045$, $r=.48$) behavior while all other differences were non-significant, including naturalness ($t=1.26$, $p=.116$, $r=.35$). Overall, the results demonstrate that participants were able to correctly rate the ironic robot as significantly ironic and humorous, and that user experience (especially hedonic quality) were significantly higher due to augmenting the social robot with our irony generation approach.

We had asked each participant at the end of the study which of the robots they would prefer overall and seven participants preferred the ironic robot, four the baseline (non-ironic) robot and one participant was undecided. Thus, an ironic robot seems to be preferred overall in a small talk dialog.

5 DISCUSSION

In the beginning, we have argued that if we could augment a social robot with the ability to use irony in multimodal human-robot communication we might be able to improve how their human conversation partners perceive them. We have also referred to

related work, describing for example the relation between conversational humor and irony. The integration of irony and related humor have undoubtly benefits for a relation between humans and their non-human conversation partners if robots can be humorous in appropriate contexts (e.g., in a small talk dialog). Ultimately, we hypothesized that a proper and meaningful use of irony would improve the conversation skills of social robots and consequently improve human conversation partners' conversation experience.

In order to explore this idea, we have first described the implementation of an irony generation approach in detail for the Reeti robot utilizing NLP, NLG, the A.L.I.C.E chat bot, as well as multimodal irony markers in terms of facial expression and prosody. Then we conducted a small-sized user study with twelve participants to demonstrate the effect of a robot using irony in a context-specific manner (i.e., small talk where the use of irony is appropriate) compared to a version of the same robot, which makes no use of irony.

Results have clearly shown that the irony generation approach works well and that indeed participants (i.e., human conversation partners in a human-robot communication setting) (i) were able to correctly identify their robotic conversation partner's use of irony and experienced associated humor, (ii) associated a better user experience with the ironic robot and (iii) overall, more participants preferred the ironic robot. The results have been consistent with what we expected in case our irony generation approach worked.

With respect to how speech based interfaces (including assistants developed by large companies, such as Amazon, Apple, Microsoft, and Google that are being employed in many homes) are becoming increasingly important and popular, we believe that our research and research contribution is timely. We hope that it will inspire fellow researchers and ultimately contribute to improving the quality of future conversations with non-human agents and their perceived value beyond traditional usability and functionality. But one should be aware that there are additional and related challenges for future research, which are beyond the scope of the research at hand, such as learning or identifying the right contexts to use irony.

Despite the encouraging user study results considering the performance of the reported irony approach, one should also be aware of potential limitations associated with the specific chat bot and robot utilized in our technical implementation and study setup.

Limitations

Occasionally, the A.L.I.C.E chat bot produced confusing responses which could have negatively influenced participants' user experience in both conditions. For example, the chat bot sometimes had problems with matching pronouns and extracting the relevant data from too complex responses. We have not controlled for dialog *quality* but asked participant to *rate* (see Figure 7) if the robots' responses were meaningful. In both conditions ratings were high in mean, suggesting that associated limitations were small.

Another potential limitation of our results and study is associated with the embodiment of the Reeti robot. The Reeti robot's overall appearance is already cute and potentially funny (e.g., it has very expressive and large eyes), which might have made it easier for the Reeti robot to convey irony and associated humor. Thus, it is unclear how well the irony generation approach will effect user perceptions of differently embodied agents/robots. Overall, results

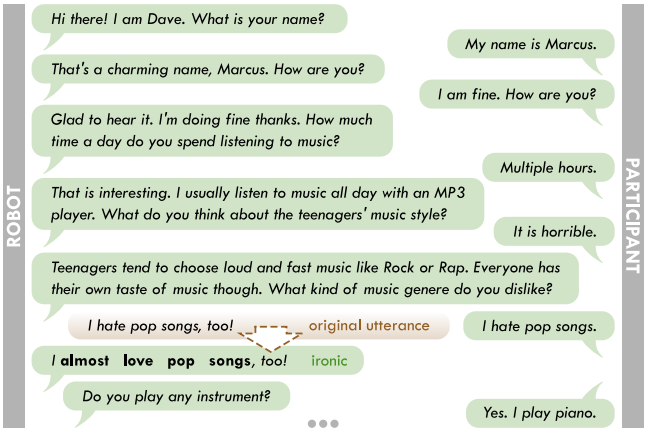


Figure 8: Beginning of a dialog from the evaluation logs

of the user study should be interpreted carefully, considering the small number of study participants.

Last but not least, we have predefined small talk as a context appropriate for irony use. There might be better or worse contexts to use irony and our approach of producing irony may need to be adapted. For example, in the study setup we have tried to use irony whenever possible, assuming within the context of small talk it would be accepted – without overly controlling its appropriateness within the dialog content. Consequently, there is room for improvement of the irony generation approach, which we aim to address in our future work to make it more robust and context-aware.

6 CONCLUSION

We have presented and evaluated a generation approach for computationally transforming linguistic content into multimodal ironic content, emphasizing spoken words generated by Natural Language Generation with matching prosody and facial expressions in real-time. The presented procedure is an important step to augment social robots with expressive conversational humor. Instead of scripting all contents in advance, it allows the robot to use verbal irony dynamically in a dialog context. Even if the generated ironic behaviors may not be perceived as humor, the procedure allows to support the expression of personality, which we believe improves the user experience of human conversation partners. To this end, we presented empirical evidence based on a user study. We were able to demonstrate that ironic robots using our multimodal procedure indeed receive better user experience ratings. Moreover, human conversation partners consider them as significantly more humorous than their non-ironic counterparts, which overall result in users preferring ironic robots over non-ironic robots in a small talk situation. For future work, we are also interested in optimizing the robot's irony and humor presentation strategy for the individual user based on human social signals [31–33, 39].

ACKNOWLEDGMENTS

This research was funded by the Bavarian State Ministry for Education, Science and the Arts (StMWFK) as part of the ForGenderCare research association.

REFERENCES

[1] Salvatore Attardo. 2000. Irony markers and functions: Towards a goal-oriented theory of irony and its processing. *Rask* 12, 1 (2000), 3–20.

[2] Salvatore Attardo, Jodi Eisterhold, Jennifer Hay, and Isabella Poggi. 2003. Multimodal markers of irony and sarcasm. *Humor* 16, 2 (2003), 243–260.

[3] Kim Binsted et al. 1995. Using humour to make natural language interfaces more friendly. In *Proceedings of the AI, ALife and Entertainment Workshop, Intern. Joint Conf. on Artificial Intelligence*.

[4] Kim Binsted and Graeme Ritchie. 1997. Computational rules for generating punning riddles. *HUMOR-International Journal of Humor Research* 10, 1 (1997), 25–76.

[5] Arnie Cann, Lawrence G Calhoun, and Janet S Banks. 1997. On the role of humor appreciation in interpersonal attraction: It's no joking matter. *Humor-International Journal of Humor Research* 10, 1 (1997), 77–90.

[6] Paula Carvalho, Luís Sarmento, Mário J. Silva, and Eugénio de Oliveira. 2009. 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 (TSA '09)*. ACM, New York, NY, USA, 53–56.

[7] Alessandra Teresa Cignarella, Simona Frenda, Valerio Basile, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2018. Overview of the EVALITA 2018 Task on Irony Detection in Italian Tweets (IronITA). In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) (CEUR Workshop Proceedings)*, Vol. 2263. CEUR-WS.org.

[8] Pawel Dybala, Michal Ptaszynski, Shinsuke Higuchi, Rafal Rzepka, and Kenji Araki. 2008. Humor Prevails! - Implementing a Joke Generator into a Conversational System. In *AI 2008: Advances in Artificial Intelligence, 21st Australasian Joint Conference on Artificial Intelligence, Auckland, New Zealand, December 1-5, 2008. Proceedings (Lecture Notes in Computer Science)*, Vol. 5360. Springer, 214–225.

[9] Marta Dynel. 2009. Beyond a Joke: Types of Conversational Humour. *Language and Linguistics Compass* 3, 5 (2009), 1284–1299.

[10] Marta Dynel. 2014. Isn't it ironic? Defining the scope of humorous irony. *Humor* 27, 4 (2014), 619–639.

[11] Simona Frenda. 2016. Computational rule-based model for Irony Detection in Italian Tweets. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016. (CEUR Workshop Proceedings)*, Vol. 1749. CEUR-WS.org.

[12] Albert Gatt and Ehud Reiter. 2009. SimpleNLG: A Realisation Engine for Practical Applications. In *ENLG 2009 - Proceedings of the 12th European Workshop on Natural Language Generation, March 30-31, 2009, Athens, Greece*. The Association for Computer Linguistics, 90–93.

[13] Lorenzo Gatti, Gözde Özbal, Oliviero Stock, and Carlo Strapparava. 2017. Automatic Generation of Lyrics Parodies. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017*. ACM, 485–491.

[14] Marco Guerini, Carlo Strapparava, and Oliviero Stock. 2008. Valentino: A Tool for Valence Shifting of Natural Language Texts. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*. European Language Resources Association.

[15] Marc Hassenzahl, Franz Koller, and Michael Burmester. 2008. Der User Experience (UX) auf der Spur: Zum Einsatz von www.attrakdiff.de. In *Berichtband des sechsten Workshops des German Chapters der Usability Professionals Association e.V., Usability Professionals 2008, Lübeck, Germany, September 7-10, 2008*. Fraunhofer Verlag, 78–82.

[16] Kotaro Hayashi, Takayuki Kanda, Takahiro Miyashita, Hiroshi Ishiguro, and Norihiro Hagita. 2008. Robot *Manzai*: Robot Conversation as a Passive-Social Medium. *I. J. Humanoid Robotics* 5, 1 (2008), 67–86.

[17] Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 Task 3: Irony Detection in English Tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT, New Orleans, Louisiana, June 5-6, 2018*. Association for Computational Linguistics, 39–50.

[18] Chenyang Huang, Osmar R. Zaïane, Amine Trabelsi, and Nouha Dziri. 2018. Automatic Dialogue Generation with Expressed Emotions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, Volume 2 (Short Papers)*. Association for Computational Linguistics, 49–54.

[19] Kleomenis Katevas, Patrick GT Healey, and Matthew Tobias Harris. 2015. Robot Comedy Lab: experimenting with the social dynamics of live performance. *Frontiers in psychology* 6 (2015).

[20] Heather Knight. 2011. Eight Lessons Learned about Non-verbal Interactions through Robot Theater. In *Social Robotics - Third International Conference, ICSR 2011, Amsterdam, The Netherlands, November 24-25, 2011. Proceedings (Lecture Notes in Computer Science)*, Vol. 7072. Springer, 42–51.

[21] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, System Demonstrations*. The Association for Computer Linguistics, 55–60.

[22] George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[23] Nicole Mirnig, Susanne Stadler, Gerald Stollnberger, Manuel Giuliani, and Manfred Tscheligi. 2016. Robot humor: How self-irony and Schadenfreude influence people's rating of robot likability. In *25th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2016, New York, NY, USA, August 26-31, 2016*. IEEE, 166–171.

[24] Nicole Mirnig, Gerald Stollnberger, Manuel Giuliani, and Manfred Tscheligi. 2017. Elements of Humor: How Humans Perceive Verbal and Non-verbal Aspects of Humorous Robot Behavior. In *Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, HRI 2017*. ACM, 211–212.

[25] John Morkes, Hadyn K. Kernal, and Clifford Nass. 1999. Effects of Humor in Task-Oriented Human-Computer Interaction and Computer-Mediated Communication: A Direct Test of SRCT Theory. *Human-Computer Interaction* 14, 4 (1999), 395–435.

[26] Andreea Niculescu, Betsy van Dijk, Anton Nijholt, Haizhou Li, and See Swee Lan. 2013. Making Social Robots More Attractive: The Effects of Voice Pitch, Humor and Empathy. *I. J. Social Robotics* 5, 2 (2013), 171–191.

[27] Anton Nijholt. 2007. *Conversational Agents and the Construction of Humorous Acts*. Wiley-Blackwell, Chapter 2, 19–47.

[28] Anton Nijholt. 2018. Robotic Stand-Up Comedy: State-of-the-Art. In *Distributed, Ambient and Pervasive Interactions: Understanding Humans - 6th International Conference, DAPI 2018, Proceedings, Part I (Lecture Notes in Computer Science)*, Vol. 10921. Springer, 391–410.

[29] Ted Pedersen and Varada Kolhatkar. 2009. WordNet: : SenseRelate: : AllWords - A Broad Coverage Word Sense Tagger that Maximizes Semantic Relatedness. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, Boulder, Colorado, USA, Demos*. The Association for Computational Linguistics, 17–20.

[30] Victor Raskin and Salvatore Attardo. 1994. Non-literalness and non-bona-fide in language: An approach to formal and computational treatments of humor. *Pragmatics & Cognition* 2, 1 (1994), 31–69.

[31] Hannes Ritschel. 2018. Socially-Aware Reinforcement Learning for Personalized Human-Robot Interaction. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2018, Stockholm, Sweden, July 10-15, 2018*. International Foundation for Autonomous Agents and Multiagent Systems Richland, SC, USA / ACM, 1775–1777.

[32] Hannes Ritschel and Elisabeth André. 2018. Shaping a Social Robot's Humor with Natural Language Generation and Socially-Aware Reinforcement Learning. In *Workshop on Natural Language Generation for Human–Robot Interaction at INLG 2018*. Tilburg, The Nederlands.

[33] Hannes Ritschel, Tobias Baur, and Elisabeth André. 2017. Adapting a Robot's Linguistic Style Based on Socially-Aware Reinforcement Learning. In *26th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2017, Lisbon, Portugal, August 28 - Sept. 1, 2017*. IEEE, 378–384.

[34] Jonas Sjöbergh and Kenji Araki. 2008. Robots Make Things Funnier. In *New Frontiers in Artificial Intelligence, JSAI 2008 Conference and Workshops, Asahikawa, Japan, June 11-13, 2008, Revised Selected Papers (Lecture Notes in Computer Science)*, Vol. 5447. Springer, 306–313.

[35] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013*. ACL, 1631–1642.

[36] Oliviero Stock and Carlo Strapparava. 2002. HAHAcronym: Humorous agents for humorous acronyms. *Stock, Oliviero, Carlo Strapparava, and Anton Nijholt. Eds* (2002), 125–135.

[37] Alessandro Valitutti and Tony Veale. 2015. Inducing an ironic effect in automated tweets. In *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015, September 21-24, 2015*. IEEE Computer Society, 153–159.

[38] Annalu Waller, Rolf Black, David A. O'Mara, Helen Pain, Graeme Ritchie, and Ruli Manurung. 2009. Evaluating the STANDUP Pun Generating Software with Children with Cerebral Palsy. *TACCESS* 1, 3 (2009), 16:1–16:27.

[39] Klaus Weber, Hannes Ritschel, Ilhan Aslan, Florian Lingenfelser, and Elisabeth André. 2018. How to Shape the Humor of a Robot - Social Behavior Adaptation Based on Reinforcement Learning. In *Proceedings of the 2018 on International Conference on Multimodal Interaction, ICMI 2018, Boulder, CO, USA, October 16-20, 2018*. ACM, 154–162.

[40] Jason A Williams, Erin L Burns, and Elizabeth A Harmon. 2009. Insincere utterances and gaze: eye contact during sarcastic statements. *Perceptual and motor skills* 108, 2 (2009), 565–572.

[41] Deirdre Wilson and Dan Sperber. 1992. On verbal irony. *Lingua* 87, 1 (1992), 53 – 76.

[42] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18)*. AAAI Press, 730–739.