Journal on Multimodal User Interfaces
https://doi.org/10.1007/s12193-024-00429-w

ORIGINAL PAPER

# Towards the development of an automated robotic storyteller: comparing approaches for emotional story annotation for non-verbal expression via body language

Sophia C. Steinhaeusser[1] · Albin Zehe[2] · Peggy Schnetter[1] · Andreas Hotho[2] · Birgit Lugrin[1]

Received: 15 May 2023 / Accepted: 8 March 2024

**Abstract**
Storytelling is a long-established tradition and listening to stories is still a popular leisure activity. Caused by technization, storytelling media expands, e.g., to social robots acting as multi-modal storytellers, using different multimodal behaviours such as facial expressions or body postures. With the overarching goal to automate robotic storytelling, we have been annotating stories with emotion labels which the robot can use to automatically adapt its behavior. With it, three different approaches are compared in two studies in this paper: 1) manual labels by human annotators (MA), 2) software-based word-sensitive annotation using the Linguistic Inquiry and Word Count program (LIWC), and 3) a machine learning based approach (ML). In an online study showing videos of a storytelling robot, the annotations were validated, with LIWC and MA achieving the best, and ML the worst results. In a laboratory user study, the three versions of the story were compared regarding transportation and cognitive absorption, revealing no significant differences but a positive trend towards MA. On this empirical basis, the *Automated Robotic Storyteller* was implemented using manual annotations. Future iterations should include other robots and modalities, fewer emotion labels and their probabilities.

**Keywords** Social robotics · Storytelling · Annotations · Emotions · Machine learning

## 1 Introduction

Emotions are a key factor of communication and conveying emotions itself is a form of communication [41]. This is not only true for human communication, but also in the field of human-robot interaction (HRI). Emotion expressing behavior shown by robots can facilitate HRI by turning robots into a natural user interface humans instinctively know how to interact with [31, 39]. This can be achieved by manipulating speech parameters or using emotional body language, thus applying facial expressions, gestures and postures [60] to the robot. Especially when it comes to storytelling, conveying emotions is a critical factor for the robotic storytelling experience [86], determining the recipients' transportation, their immersion into the story told [38]. "Ideally, the robot would be able to process automatically a given tale or short story, and to play it for [its] audience" [2, p. 1] by automatically applying emotional body language matching the spoken text. This match is crucial since humans need congruence to understand emotional communication [96], and a mismatch resulting in harder processing of the narrative would impede transportation [92]. To achieve this automation, the story first has to be annotated. This can be manually done by human annotators. As an alternative, a system able to recognize emotions from natural language in text form could be used [33].

Our overall aim is the development of an *Automated Robotic Storyteller* which is able to automatically express emotions using body language while telling a story. In

✉ Sophia C. Steinhaeusser
sophia.steinhaeusser@uni-wuerzburg.de

Albin Zehe
albin.zehe@uni-wuerzburg.de

Peggy Schnetter
peggy.schnetter@stud-mail.uni-wuerzburg.de

Andreas Hotho
hotho@informatik.uni-wuerzburg.de

Birgit Lugrin
birgit.lugrin@uni-wuerzburg.de

[1] Human-Computer Interaction, University of Würzburg, Am Hubland, 97074 Würzburg, Bavaria, Germany

[2] Data Science, University of Würzburg, Am Hubland, 97074 Würzburg, Bavaria, Germany

Published online: 04 April 2024

this paper, three different approaches of annotating texts for robotic storytelling are compared in order to identify the most suitable way of annotation for the automated robotic storytelling process. Human annotators preparing a story, the word-sensitive text analysis program *Linguistic Inquiry and Word Count* version 2015 [66, LIWC], and a machine learning approach are validated and evaluated concerning their fit and the resulting storytelling experience. Based on the results, the *Automated Robotic Storyteller* framework is implemented. First, the respective theoretical background is presented. Afterwards, emotional body language for the robotic storyteller is determined. For the annotation approaches' comparison, an online validation and an evaluating user study are presented. Last, the results from both studies are discussed and the framework for the Automated Robotic Storyteller is build upon our empirical results.

## 2 Related work

Based on our scope, we consider related research on social robots as storytellers depicting emotions, approaches to prepare stories for emotional robotic storytelling including options how an automation of the later can be achieved.

### 2.1 Emotional storytelling in agents

Agents, including virtual agents and robotic agents, can perform storytelling as both actors in a story but also as the teller of the story [10]. In this work, we focus on the second perspective. Being physically embodied agents that behave in a socially interactive and understandable way [25, 45, 59], social robots are able to convey emotions which facilitate the human-robot interaction. This is especially relevant to the context of storytelling, where emotions are crucial for both story comprehension and storytelling experience [86, 98]. Thus it is not surprising that when comparing virtual agents and physically embodied robots [21] reported that robots are preferable over virtual agents for emotional storytelling.

Emotions as inner states [69] are elicited by events and their respective significance [28, 34]. In more detail, they are not only feeling states, but "complex chain[s] of loosely connected events that begins with a stimulus and includes feelings, psychological changes, impulses to action and specific, goal-directed behavior" [69, p. 345]. There are two types of emotion classification systems, dimensional and basic emotion models. While the former arrange emotions on axes such as pleasantness [28], valence or arousal [73], basic emotion models consider emotions as discrete elements that can be combined to build more complex emotions [28]. Plutchik [68, 69] integrated eight basic emotions in his *Wheel of Emotions*, conceptualized analogously to a color wheel where complementary colors are arranged vis-à-vis. The complementary basic emotions he used are joy and sadness, anger and fear, trust and disgust, and surprise and anticipation [69]. As with colors, more complex emotions can be formed by combining juxtaposed emotions. In addition, Plutchik included thee levels of intensity for each emotion, forming the wheel into a cone [69]. Expressing emotions, e.g. using facial expression, is an automatic reaction, universal and culturally different at the same time, depending in the context being private respectively social and subjected to culture rules of expression management [26].

#### 2.1.1 Emotion expression in robots

As with humans, there are several ways robots can express emotions. Most important are voice modulation, facial expressions, and body language [60]. Prosodic features within speech are the voice's pitch, intensity, volume, and tempo as well as pauses between sections. For example, a high volume and pitch combined with fast speaking indicate anger, whereas speaking slowly with low intensity but high pitch indicates sadness [94]. Further regarding non-verbal behavior, emotional facial expressions can be generated for robots. For example, the *Facial Action Codings System* [27, FACS] can be utilized to develop robotic emotional facial expressions [44]. However, to allow for emotional facial expressions a robot's capacity of motors in the facial region is a crucial factor. While social robots such as *Reeti* [72] are capable of showing emotions using their face, most common robots such as *Pepper* [79] and *Nao* [78] are not able to deploy facial expressions. Using these or similar robots, one can rely on emotional body language, such as gestures or postures, because humans can accurately interpret body language without accompanying facial expressions or vocal cues shown by virtual and embodied agents [11, 13]. Gestures illustrate speech and provide additional information by adding spatial, referential, and iconic thus semantic details [17, 74]. Furthermore, full-body postures can improve emotion recognition [103]. This way, gestures and postures improve comprehension and believability of a robot's behavior as well as the robot's evaluation [74].

Striepe et al. [86] compared three versions of the same story told by an emotional or neutral robot or via audio book using an audio track recorded by a professional human storyteller in all three conditions. Results showed that the recipients' transportation, "the extent that individuals are absorbed into a story or transported into a narrative world" [38, p. 701], was increased in the audio book and emotional robot conditions compared to the neutral robot condition. The "neutral robot decreased participants' ability to mentally involve in the narrative." [86, p. 133] and was rated worst overall indicating that conveying emotions is crucial for the robotic storytelling experience. Further, Xu et al. [98] compared robotic storytellers that showed congruent respectively incongruent emotions by using co-verbal gestures or did not use emotional gestures. Their findings indicate that using congruent emotional gestures facilitates grasping the story's emotions and thus comprehension. Alike, emotional body language improves the evaluation of and satisfaction with the robot as well as its persuasiveness [42, 98, 99] as do congruent emotional facial expressions [7].

### 2.1.2 Emotional expression generation

Because of the high importance of conveying emotions in the robotic storytelling, a robot should be able to automatically generate emotional expressions such as body language when given a story [2]. Current approaches overcome finite state machines by involving, e.g., behavior trees [46] or machine learning algorithms [70]. Nevertheless, traditional approaches such as the *Greta* platform [64] and the *Behavior Expression Animation Toolkit* [18, BEAT], developed for virtual agents, indicate which steps are needed in order to allow for an automation of the process [65]. First, XML-formatted text is converted in a parse tree to manage language tagging, behavior planning respectively generation and realization of the behavior. In the process, the system collects contextual and linguistic information from the text to derive matching gestures and voice modulation for the virtual agents. Another approach was realized by Zabala et al. [102], who developed a first prototype for automatic gesture generation for a storytelling Pepper robot. In a rule-based approach a parser extracts keywords from a story's text which are automatically combined with emblematic, deictic, iconic, and metaphoric gestures, while emotions are identified using *TextBlob*,[1] another word-based approach, to retrieve a sentence's emotional valence. To integrate the identified valence into the storytelling the robot's movement is adjusted: for positive valence straighter hips, more upright head, and faster movements are used and vice versa. Last, a *Generative Adversarial Network* trained on motion capture data of human speakers generates beat gestures. An evaluation revealed a positive perception of these rule-based approaches, but did not include the storytelling experience. However, only a limited amount of emotional features is integrated in this approach. Even though, the existing frameworks still have several issues to solve, e.g. the context-based speech creation or the high cost of developing behavior based on footage of human actors [18], they indicate starting points for the development of an *Automated Robotic Storyteller* that is capable of automatically generating emotional storytelling based on plain text. First, the text has to be prepared for the robot by adding information on the relevant emotions in the form of annotations. Second, a set of emotion expressions, e.g. full-body postures, should be linked to the annotated emotions.

In order to apply emotional expressions which improve the storytelling scenario, the underlying emotions need to be recognized by the recipients. Past research tried to imitate human body language expressing emotions by generating a corpus of gestures and postures copied from humans to robots [31, 43, 64] or virtual agents [4]. However, the pure imitation of humans is not suitable for robots' emotion expression, because of their kinematic constraints, i.e. fewer joints and fewer degrees of freedom, their smaller range of motion, and the different mass distribution compared to the human body. Similar, the robots' usually smaller size leads to different effects [31]. Instead, points of reference could be the dimensions an emotion can be described with. For example, the robot's head position can be adjusted, indicating decreased arousal, negative valence and avoiding stance by moving it down, respectively showing off a high level of energy, positive valence and an approaching stance by moving the head up [12, 13]. Furthermore, the motion's speed can be adjusted along these dimensions.

In general, the recognition of emotions expressed by robots was mixed in past research. While children in a study conducted by Beck et al. [13] were able to recognize emotions from Nao's postures more often than they could have done by chance, participants had problems recognizing emotions from Nao's gait in a study carried out by Izui et al. [47], especially fear, sadness, and joy were hard to recognize. A supplementing approach is to expand the modalities used to express emotions by adding colors and sounds [43]. Many social robots, e.g., Reeti [72], Nao [78], and Pepper [79], offer color-changing LEDs. Commonly understood colors are red depicting anger, violet for sadness, yellow indicating joy, and green resembling fear [89]. However, these associations are culture-dependent. Most of these colors are used in Plutchik's classification system of emotions, the *Wheel of Emotions* [68], too. Nonetheless, only a small set of studies has been carried out on adding eye colors to facilitate emotion recognition yet [80], reporting mixed results [43, 81].

## 2.2 Annotations and emotions

Classifying emotions in text is a part of sentiment analysis [62] which uses natural language processing to analyze texts. For example, dialogues in movies [63] but also stories in preparation for robotic storytelling [8] can be labeled with emotions to allow for quantitative analysis and subsequent processing. Next to emotions, also the story's structure, characters, speech turns, world states, and further meta information can be annotated [22, 24, 57]. However, concerning the use case of robotic storytelling, emotion labels are of highest interest due to the high importance of emotion expression in the storytelling scenario described above.

There are several approaches on annotating story emotions, not necessarily focusing on the storytelling context.

[1] https://github.com/sloria/TextBlob

The most common way of annotation is the manual annotation where human annotators label categories to the text's individual tokens. For example, the *EmoTales* database comprises manually annotated fairytales of which annotation is based on the emotional dimensions valence, activation, and power [33]. Further, the automation of the sentiment analysis for emotions in stories, as used in Zabala et al.'s approach [102] described above, is possible and has been studied extensively [56]. Word-sensitive analysis can be done by collecting words tied to an emotion, resulting in dictionaries like the ones used by the program *Linguistic Inquiry and Word Count* [66, LIWC] to analyze text on a quantitative basis. However, this approach does not take the words' context into account [97]. This issue can be partially solved by combining these word lists with semantic knowledge. Thus, the *EmoLogus* project [30] derives emotions from text by applying semantic knowledge within a dictionary that was generated based on a corpus of texts manually annotated by human annotators. However, these traditional approaches only work well when emotions are described somewhat explicitly in the text and often struggle with negations and other kinds of modifiers. Since narrative text often has a less explicit writing style than other kinds of text, many emotions therein cannot be detected by these approaches. As a potential solution to this, machine learning methods for Sentiment Analysis have seen large improvements in the last years [101]. Most recent approaches here are based on deep learning models such as BERT [23]: A Transformer-based model [91] that is pre-trained on large text corpora to provide a general understanding of language and then fine-tuned to a specific task - for example sentiment analysis. However, these models are mostly applied to domains such as product reviews and social media texts, where large amounts of training data are available. In contrast, emotion analysis is still challenging for the domain of narrative texts. Kim et al. [49] provide a comprehensive survey of applications of emotion analysis in this domain, showing that deep learning methods are not yet prevalent there. Kim et al. [50] show that deep learning can also be successfully applied on narrative texts, introducing an annotated corpus and a classifier based on a kind of recurrent neural network, specifically a Gated Recurrent Unit [19, GRU]. We have subsequently improved on their results by replacing the GRU with a BERT-based model [105].

The annotation of emotion labels should be based on consolidated models of emotion. One of the emotion models most often used for annotating stories is the *Wheel of Emotion* [68] previously described above. For example, Kolog et al. [51] identified eight basic emotions within the wheel as suitable for emotion recognition from students' life stories based on a focus group discussion, whereas [61] also utilized the secondary emotions indicating lower intensities. Thus, both can be taken into account when annotating stories for robotic storytelling. Most often, stories for robotic storytellers are annotated manually by human annotators, e.g. [7, 35, 84, 86, 87]. Less approaches use automated annotation. Augello et al. [9] identify both target words for contextual gestures and emotional content of the stories. To identify the former, the text is segmented in sentences and words are cut down to their lemmas to find matches between words in the text and available predefined gestures. Furthermore, they used the WordNet-based Synesketch API [52] to annotate the text in regard the Ekman's basic emotions [28]. Doing so, the robotic storyteller *NarRob* is able to perform automated storytelling by automatically annotating stories and using a set of pre-defined labeled gestures [9]. Evaluations indicate that the robot's narration is appreciated [14], however, the fit between annotation labels, gestures and story content was not assessed although this step is crucial for the decision on an annotation approach.

## 3 Research goals

Our overall aim is to develop an *Automated Robotic Storyteller*, which is capable of automatically conveying emotions using body language while telling a story.

Therefor, three steps are necessary. First, emotional body language has to be produced for the robot. For this, emotion expressions will be generated for the robot Nao [78], one of the most commonly used social robots, and matched to the emotions within Plutchik's *Wheel of Emotions* [68] on an empirical basis. The results will be used in the implementation and evaluation of the next step. Second, the most suitable annotation approach to prepare text for the robotic storytelling has to be determined. Accordingly, the three approaches of manual annotation by human annotators, using the text analysis program LIWC 2015 [66], and a machine learning based approach will be compared. Again, this will happen using empirical methods, (1) by validating the annotation approaches by re-labeling emotions to the robot's respective storytelling output, and (2) by evaluating the storytelling experience in a user study. Third, the *Automated Robotic Storyteller* will be implemented dependent on the decision for an annotation approach based on the studies' findings.

## 4 Implementation of the scripted robotic storyteller

To compare the three approaches of manual labels by human annotators, software-based word-sensitive annotations using LIWC, and annotations performed by machine learning, the short story *The Secret Cave or John Lee's Adventure* by H.P. Lovecraft [58] was annotated using each of the approaches. In addition, emotional postures for the Nao robot [78] were identified. Afterwards, each version was implemented using these postures.

We chose this story due to the popularity of the horror genre [76], due to its shortness, and being public domain. While H.P. Lovecraft is a famous author, this short story is rather unpopular. Thus, study participants might recognize the writer's style but likely not the plot. The story is about John Lee and his sister Alice, who discover a secret cave in the basement of their family's house. While discovering the cave, John finds a small box he takes with him. Being startled by a sudden inrush of water, little Alice drowns. John manages to return to the house with his sister's dead body and tells his parents about the accident. After Alice's funeral, he opens the box and finds a gold ingot in it. The story is told in 34 sentences and reading it out loud takes approximately four minutes. Since the experiments were conducted in Germany, the story was translated to German.

## 4.1 Verbal behavior: annotations

All approaches of annotation used the same tokenization. As sentence-based tokenization is wide spread in studies on robotic storytelling (see e.g. [82, 84, 86, 87, 102]) but also on sentiment analysis (see e.g. [5, 33]), the story was clustered into 34 tokens by dividing it at full stops. Alike, all of the annotations used the 24 subemotions of Plutchik's *Wheel of Emotions* [68] extended by a "neutral" label. The distribution of emotions in annotations according to all settings can be found in Table 1.

**Table 1** Distributions of emotions according to different annotation settings

| Annotation approach | MA | LIWC | ML |
|---|---|---|---|
| Emotion | *n* | *n* | *n* |
| Anticipation | 5 | 0 | 17 |
| Pensiveness | 1 | 1 | 4 |
| Terror | 2 | 4 | 3 |
| Fear | 2 | 5 | 3 |
| Neutral | 7 | 17 | 2 |
| Trust | 1 | 1 | 1 |
| Interest | 4 | 0 | 1 |
| Joy | 0 | 0 | 1 |
| Ecstasy | 0 | 0 | 1 |
| Sadness | 2 | 2 | 1 |
| Apprehension | 2 | 0 | 0 |
| Distraction | 2 | 0 | 0 |
| Amazement | 2 | 1 | 0 |
| Vigilance | 1 | 0 | 0 |
| Acceptance | 1 | 2 | 0 |
| Grief | 1 | 1 | 0 |
| Surprise | 1 | 0 | 0 |

*MA* manual annotation, *ML* machine learning

### 4.1.1 Manual annotations

Eleven human annotators (age: $M = 23.00$, $SD = 2.93$) were acquired. Ten of them self-reported as female, whereas one annotator self-indicated as male. All of them were native speakers or spoke German for at least ten years. Three of them stated to be employed, whereas eight participants were students in the field of media communication or human-computer interaction. The annotators were recruited from the university's participant pool and personal contacts. Students were rewarded with credits mandatory for obtaining their program of study's degrees.

Using an online survey, they annotated the story *The Secret Cave or John Lee's Adventure* and four other stories of which annotations were used as material for the machine learning approach. One page of the survey comprised one story and the tokens were displayed in the order they appeared in the story. For each token, the annotators decided which of the 24 subemotions from Plutchik's *Wheel of Emotions* [68] they would want a storyteller to act while speaking the token. Alternative options were "utral" and "I don't know". Consensus was achieved by majority decision. For the story *The Secret Cave or John Lee's Adventure* inter-annotator agreement was between 27.27% and 72.73%, calculated Fleiss Kappa =.18 indicates little agreement [54], which is comparable to similar studies (e.g., [33]).

### 4.1.2 Semi-automatic annotations via LIWC

The *Linguistic Inquiry and Word Count* 2015 [66, LIWC] can be used to analyze text on the basis of words and punctuation. It uses an internal dictionary including target words assigned to categories as well as punctuation characters which are both counted by the program when analyzing a given text [67].

To assess the story's emotions via LIWC version 2015 [66], a custom dictionary comprising relevant emotional words contained in the story was build. First, the text was analyzed for these relevant words by generating a list of all included words using the online tool *WordList Maker* [36]. This generated word list was sorted in descending order according to the number of occurrences in the story. This procedure excludes the context in which the words occur, thus only the emotion of the word itself was analyzed. Emotional words were then assigned to one of Plutchik's emotional categories [68], e.g., the word "remember" was assigned to the category "pensiveness". If a word suited more than one category, decisions were made based on the context in the story. Also, the words "yes" and "no" were integrated into the dictionary as indicators of "acceptance" respectively "loathing". Using this custom dictionary within *LIWC2015*

[66], the story *The Secret Cave or John Lee's Adventure* was analyzed.

### 4.1.3 Automatic annotations via machine learning

For the machine learning annotations, we followed Zehe et al. [105] and fine-tuned a BERT-model to the task of emotion classification. We compared different strategies for training the model, including using additional training data from related tasks (*German Novel Dataset* [104, GND] and HeiSt [40]) and different base language models (`bert-base-german-cased`, `bert-base-multilingual-cased`). The setting used for the final experiments is as follows: As a first step, we used a pre-trained German BERT-base model and fine-tuned it to a corpus of 451 German novels (342 million tokens) to improve the model's understanding of this domain. Since we only had a very small number of annotated samples from this project (four short stories, cf. Sect. 4.1.1), we additionally pre-trained the model for binary sentiment classification (positive vs. negative) on the 270 annotated sentences of the *German Novel Dataset* [104, GND], which are extracted from German novels and manually labeled. Finally, we used the four additional stories labeled in this project[2] as well as three sets of annotations from existing studies [83–85, 87], as training data, using the label provided by the majority of human annotators as the training target. We trained the model for up to 20 epochs with early stopping on a validation dataset that was randomly split from the training data. The resulting model was used to annotate the story *The Secret Cave or John Lee's Adventure*.

All training steps were implemented using the FARM library.[3]

## 4.2 Non-verbal behavior: identification of emotional robotic expressions

Since full-body postures, "specific positioning that [a robot's] body takes during a timeframe" [13, p. 63], were already shown to allow for recognizable emotion expression [31], we decided to use this kind of body language as a basis and supplemented them with varying head positions and eye LED color to obtain multi-modal full body emotion expressions. To allow for the emotional robotic storytelling that is automatically generated based on emotion labels, one emotion expression of the robot Nao [78] matching one emotion of Plutchik's *Wheel of Emotions* [68] each was determined using an online survey. In addition to the mapping, we also recorded a measure of how well participants were able to associate the posture with the corresponding emotion.

### 4.2.1 Expression creation

Following the approach by Zhang et al. [106], who photographed postures performed by a robot and presented the photos together with emotion labels, we started by creating two to three different expressions for each of Plutchik's subemotions using *Choregraphe* 2.8.10 [3] based on the findings of related studies presented in Sect. 2.2. As recognitions rates in previous works were higher when using pictures compared to videos [103], we used photos of the emotional expressions in our study. In addition, our *Automated Robotic Storyteller* should be able to show to consecutive emotion expressions without going back to the neutral position. Therefore, videos in which the bodily expression changes from neutral to a certain emotion and back to neutral would not reflect our final product.

Every expression was photographed twice: One picture was taken with neutral (white) eye LEDs and one picture with colored eye LEDs matching the colors in the *Wheel of emotions* [68]. Both versions were treated as individual expressions in the following. In total, 112 pictures of Nao executing the non-verbal expressions were taken with constant lighting, camera setup and background.

In an online study, the expressions were validated by assigning each picture to one emotion. Entering the website hosted via *LimeSurvey* [55], participants first gave informed consent. In the survey, they were shown one of the pictures at a time without any emotional context and asked to choose which one of Plutchik's subemotions the robot's expression might represent. Alternatively, a "neutral" label or the answer "I don't know" could be chosen. After the participants chose the emotional category they think fits best for the shown expression in the picture, they proceeded to the next picture. At the end of the online study participants were asked to fill out a demographic questionnaire.

In total, 56 persons took part in the survey. Eleven of the participants self-reported as male, whereas 45 indicated themselves as female. None of the participants self-reported being diverse gender. The mean age was 21.13 years, *SD* = 2.85.

### 4.2.2 Assignment process

To determine non-verbal expressions matching Plutchik's emotions [68], (1) the frequency a picture was assigned to an emotion as well as (2) the frequency specific emotions were assigned to a picture were counted.

Using the frequencies from (1), an initial set of emotional expressions was selected including expressions that were most often assigned to an emotion. If two or more expressions were assigned to the same emotion most often, results from (2) were taken into account, deciding for the expression the emotion was assigned to with the highest frequency. Alike,

[2] Average Kappa =.08.

[3] https://github.com/deepset-ai/FARM

if an expression was assigned to two or more emotions most often, the emotion the expression was assigned to with the highest frequency was chosen and the expression was deleted from the other emotions' lists. This procedure was necessary, because some emotions, e.g. "interest", were never the most often assigned one, but an expression should be matched to these emotions, too. In some cases (1) and (2) led to the exact same frequencies and coincidences could not be solved by the procedure described. In this case, the emotion that was less dominated by others was chosen for an expression.

Using this decision making process, one non-verbal expression was matched to each of Plutchik's 24 subemotions [68]. The resulting expressions are displayed in Appendix A. The percentage that expressions' pictures were assigned to the emotion they were chosen to represent in the end ranged from 10.71% to 64.71%. The highest percentage was achieved for the subemotions "pensiveness", "ecstasy", "distraction", and "terror". The lowest percentage was found for "admiration", "vigilance", "boredom", "acceptance", and "amazement". We record these percentages as the *Recognition Rate* of the posture or emotion. They are displayed in Table 2. Expressions with a higher recognition rate presumably represent the corresponding emotion better and are therefore more easily recognized by the study's participants. Since negative emotions show descriptively better recognition rates ($M$ = 34.00%, $SD$ = 16.35%, $n$ = 13) compared to positive ($M$ = 22.89%, $SD$ = 14.15%, $n$ = 9) and emotions of unclear valence ($M$ = 16.07%, $SD$ = 7.58%, $n$ = 2; namely surprise and vigilance), this reinforces our choice of a story containing mainly negative emotions.

### 4.2.3 Resulting expression library

With this online survey, emotional non-verbal expressions for the robot Nao [78] were matched to the 24 subemotions of Plutchik's *Wheel of Emotions* [68]. From 112 expressions, the 24 most fitting, displayed in Appendix A, were identified. They were stored in a custom library in *Choregraphe* [3] to allow the robotic storyteller to apply the expressive emotional body language. Since the agreement of the participants' rating strongly differs between the emotions, the integration of expressions with low percentages of agreement, and thus low recognition rate, should be considered carefully.

## 4.3 Combining verbal and non-verbal behavior

To be used as stimulus material, the story was implemented for the robot Nao [78] using *Choregraphe* version 2.8.10 [3]. For each token, the emotional expression determined in Sect. 4.2 matching the emotion labeled by the annotation approach was applied on the start of the token. This procedure was repeated for each of the annotation approaches resulting in three versions of the story differing in the expressions carried out by the robot accompanying the respective token.

**Table 2** Recognition rates of the resulting expression library

| Sub-emotion | Recognition rate (%) |
|---|---|
| Serenity | 25.49 |
| Joy | 17.86 |
| Ecstasy | 58.82 |
| Acceptance | 16.07 |
| Trust | 23.53 |
| Admiration | 10.71 |
| Apprehension | 31.37 |
| Fear | 49.02 |
| Terror | 50.98 |
| Distraction | 56.86 |
| Surprise | 21.43 |
| Amazement | 16.07 |
| Pensiveness | 64.71 |
| Sadness | 33.33 |
| Grief | 27.45 |
| Boredom | 10.71 |
| Disgust | 25.00 |
| Loathing | 17.65 |
| Annoyance | 19.61 |
| Anger | 28.57 |
| Rage | 26.79 |
| Interest | 19.64 |
| Anticipation | 17.86 |
| Vigilance | 10.71 |

## 5 Validation

To validate the three annotation approaches, each version of the storytelling was video-recorded and re-annotated. We chose this approach because of the high importance of multimodal congruence in human-robot but also human-human communication (see e.g., [7, 88, 96]), supposing that high recognition speaks for high congruence of emotion assumed by the annotation approaches and the human re-annotators and thus for suitability of an annotation approach. By relabeling emotions to the individual video-recorded tokens and comparing them to the initial emotion labels, (1) the recognition of emotional expressions as well as (2) the fit of the respective annotation were investigated, because both factors affect the re-annotated emotions. Since related works highlight advantages of all three approaches, we postulate an undirected hypothesis:

**H1**: Re-annotation congruence differs between the three different annotation approaches.

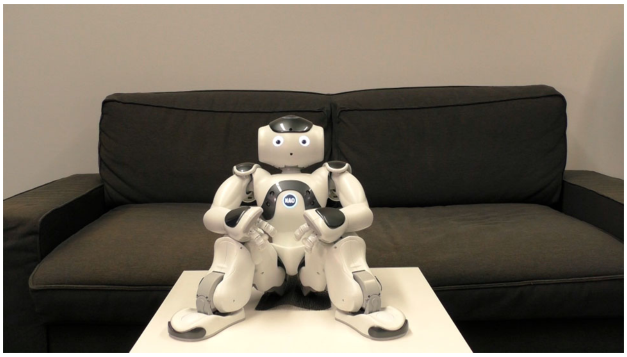

**Fig. 1** Video-taped robotic storyteller in the neutral expression

## 5.1 Method

The three annotation approaches were validated in an online validation using a between groups design (manual annotation vs. LIWC vs. machine learning) to compare the congruence between the original and re-annotated emotion labels of each annotation approach.

When entering the website hosted via *LimeSurvey* [55], participants first gave informed consent. After being randomly assigned to one of the three conditions, i.e. manual annotations, LIWC or machine learned annotations, they started the task. They watched a video of the first token told by the robot Nao [78] which applied the respective emotional expression as described in Sect. 4 Implementation for the annotated emotion. Setting and angle are displayed in Fig. 1. On the same page, they were asked to indicate which emotion may be expressed in the video. Therefore, an answer set including all the 24 subemotions of Plutchik's *Wheel of Emotions* [68] plus the label "neutral" and the alternative answer "I don't know" was provided. Also, participants were asked to comment on their decision, e.g. referring to the robots behavior and/or the story content. This procedure was repeated for each of the 34 tokens of the story in linear order.

After completing the story's re-annotation, participants provided demographic data, e.g., their age and self-reported gender. Last, a question on story details was asked to control whether the story's tokens were received attentively.

## 5.2 Participants

A total of 151 persons took part in the study. However, 39 participants had to be excluded because they answered the question concerning story details incorrectly, indicating they were not attentively listening to the presented tokens. Thus, 112 participants' (age: $M$ = 21.22, $SD$ = 1.92) records were included into the analysis. Twenty-four participants self-reported as male (age: $M$ = 21.83, $SD$ = 1.76), 87 of the test persons indicated themselves as female (age: $M$ = 21.07, $SD$ = 1.95), and one person self-reported as diverse gender (age = 20 years). Almost all participants were native speakers ($n$ = 111), only one of them reported speaking German for more than ten years. Being randomly assigned to one of the three groups, 36 participants re-annotated the videos produced using the manual annotations and the machine learning approach each, whereas 40 persons were assigned to the LIWC condition. Descriptive data per group is displayed in Table 3.

**Table 3** Participants' demographic data per group from Validation

| | ♀ | ♂ | Diverse | Age | | *n* |
|---|---|---|---|---|---|---|
| | | | | *M* | *SD* | |
| MA | 28 | 7 | 1 | 21.19 | 2.10 | 36 |
| LIWC | 29 | 11 | 0 | 21.25 | 1.69 | 40 |
| ML | 30 | 6 | 0 | 21.22 | 2.03 | 36 |

*MA* manual annotation, *ML* machine learning

## 5.3 Results

To investigate and compare the annotation approaches' influence on the congruence between initial emotion labels and labels re-annotated by the participants (**H1**), $\chi^2$-tests using 2 (matching or other labels) x 3 (manual annotation, LIWC, machine learning) tables were carried out. Descriptive results are displayed in Fig. 2. All analyses were calculated using *Microsoft Excel* 2016.

The highest congruence of 27.35% between initial and re-annotated emotion labels was achieved in the LIWC condition, while 26.39% congruence was achieved in the group that re-annotated the video based on manual annotations. Only 10.21% accordance to the initial labels was achieved in the machine learning condition. There was a significant association between annotation approach and the congruence of the re-labeled emotions to the initial annotations, $\chi^2$ (2) = 137.18, $p < .001$.

Since tokens initially labeled as "neutral" were the easiest to recognize (48.41% in the manual annotation condition, 38.53% in the LIWC condition, 58.33% in the machine learning condition), the test was repeated excluding these tokens ($n_{\text{manual annotation}} = 7$, $n_{\text{LIWC}} = 17$, $n_{\text{machine learning}} = 2$) in a follow-up analysis. Doing so, 27 tokens remained in the manual annotation, 17 tokens in the LIWC condition, and 32 tokens in the machine learning condition. Again, the congruence to the initial annotations was significantly associated to the condition ($\chi^2$ (2) = 82.61, $p < .001$), with initial manual annotations showing the highest congruence of 20.68%, followed by LIWC with 16.18%, and 7.20% congruence in the machine learning condition.

Because the overall congruence was rather low, the emotion labels were generalized to Plutchik's eight basic emotions (joy, sadness, anger, fear, trust, disgust, surprise,

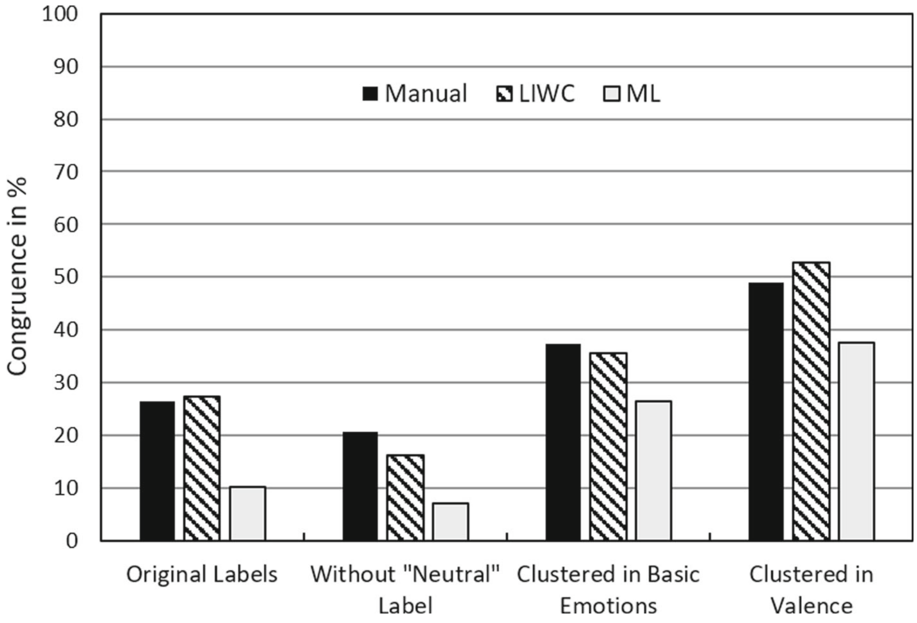

**Fig. 2** Descriptive results of the re-annotations (*Manual* manual annotations, *ML* machine learning)

anticipation) [68] extended by the "neutral" label to calculate another follow-up analysis. Once more, the $\chi^2$-test indicated a significant association between congruence of initial and re-annotated emotion labels and annotation approach used for the initial label, $\chi^2$ (2) = 37.66, $p$ < .001. Again, the highest congruence was achieved for the manual annotation condition (37.42%), followed by the LIWC condition (35.59%), and the worst results were again indicated for the machine learning condition with 26.55% congruence. The highest level of generalizability in this context is the emotions' valence. In a last follow-up analysis, the emotion labels were sorted by valence, either being positive, negative, or neutral. One more time, the calculated test showed a significant association between congruence and annotation approach used for the initial labels, $\chi^2$ (2) = 62.74, $p$ < .001. On this level of generalization the LIWC condition showed the highest congruence of 52.79%, whereas a congruence of 48.86% was achieved in the manually annotated story version, and only 37.66% congruence were determined for the machine learning condition.

Participants were asked to provide comments on their decision for an emotion label. Four categories of reasons were counted: bodily expression-related, voice-related, related to story content and related specifically to the robot's eye LEDs' color. If two or more categories were mentioned in a comment, it was counted for each of them. Most comments included the robots bodily expression ($n$ = 1417) and the content of the token, $n$ = 1053. Further, the robot's voice was mentioned as a reason for emotion decision 371 times, although the voice was not manipulated in this study. Few comments included the robot's eye color, $n$ = 41. To determine whether the comments provided by participants from the three conditions differed in their amount regarding the categories, $\chi^2$-tests were calculated using 2 (comments with or without the respective category) x 3 (manual annotation, LIWC, machine learning) tables for each category. Computed values are displayed in Table 4. The results showed a significant association between annotation approaches and comments on the tokens' content and the robot's voice or eye color. Regarding descriptive data, most comments included the tokens' content in the machine learning condition, while voice was mentioned most frequently in the LIWC condition. Eye color, again, was most often named in the machine learning condition. Mentioning the robot's posture as a reason for deciding for an emotion was not significantly associated to the annotation approach.

**Table 4** Comments per category per group from Study I - Validation

| | *n* | | | $\chi^2$ | *p* |
|---|---|---|---|---|---|
| | MA | LIWC | ML | | |
| Posture | 475 | 472 | 470 | 5.75 | .057 |
| Content | 333 | 327 | 393 | 21.12 | <.001*** |
| Voice | 113 | 161 | 97 | 11.75 | .003** |
| Eye color | 6 | 16 | 19 | 6.68 | .035* |

*MA* manual annotation, *ML* machine learning
**p* <.05, ***p* <.01, ****p* <.001

We additionally analyzed the potential influence of emotions that are more or less frequently annotated by different methods on these results. To this end, we used the recognition rate of expressions from the prestudy (cf. Sect. 4.2), which measures how well an emotion can be recognized from Nao's expression. Computing the average recognition rate of emotions according to the different annotation settings, we find scores of 29.89% for the manual annotations, 39.98%

for LIWC and 31.74% for the machine learning annotations when excluding neutral annotations (for which Nao's default sitting posture was used). If we assume a recognition rate of 1 for the neutral emotion because of the supposed easiness of its identification, we reach average scores of 44.32% (manual annotation), 69.99% (LIWC) and 35.75% (machine learning).

### 5.4 Discussion

To validate and compare the three annotation approaches, participants re-annotated the recorded storytelling per token. As can bee seen in Table 1, the investigated annotation approaches differ largely in emotion assignment, emphasizing the importance of their comparison. A significant association between annotation approach and congruence between the initial and re-annotated labels was found, hence **H1** is accepted. The highest congruence was achieved in the LIWC condition, followed by the manually annotated version. This could have been caused by the high amount of "neutral" labels in the LIWC condition which included 17 "neutral" labeled tokens compared to seven in the manual annotated version and two "neutral" tokens in the machine learning condition: Because neutral tokens are portrayed by Nao's default sitting position, we assume that they are very easy to recognize. This assumption is also supported by the results of the validation study presented above, where neutral tokens were recognized with comparatively high percentage in all conditions. Excluding neutral tokens from the analysis changed the pattern. The highest congruence was achieved in the manually annotated version, followed by the LIWC condition. The worst results were yielded for the machine learning condition. This pattern was replicated when generalizing the labels on the basic emotion and valence level, always on a significant level. Thus, the manual annotation and word-sensitive annotation performed best in the validation.

The analysis of the recognition rate of different emotions reveals some additional insights into possible reasons for the approaches' performance: As displayed in Table 2, several emotion expressions achieved only low recognition rates and thus might have impeded the recognition of emotions the video-tapes sentences. As shown in Table 1, the machine learning annotation assigns the emotion "anticipation" very frequently, which seems to be hard to recognize from Nao's posture, while especially LIWC assigns the easier "neutral" emotion very frequently. This explains, at least partially, the bad performance of the machine learning annotations in this study. Reasons for the partially low recognition rates might be the subtle usage of the LEDs. Sometimes two intensity levels of an emotion only differed in NAO's eye color, for instance acceptance and trust (see Appendix A). While this small manipulation might have been recognized when looking at the pictures in the expression selection process, it is likely to be overlooked when watching the short videos. Similarly, Häring et al. [43] reported eye color as unreliable for conveying emotions. Interestingly, the manual annotations performed well despite assigning emotions that are rather hard to recognize from Nao, suggesting that the manual annotations match very well with the sentences they are associated with.

However, it is conspicuous that the initial computed congruence on the sub-emotion level is rather low in all conditions. This could be due to the subject of emotion annotation, especially for human annotators. Francisco et al. [33] and Kolog et al. [51] both report rather low inter-annotator agreement for labeling emotions to texts, too. Using flexible tokenization and a smaller set of 15 emotion labels, [95] achieved higher agreement of $\kappa = 0.34$ but suggested this value being surprisingly high for this difficult task. Not only is emotion expression through text rather complex, leading to even short sentences conveying multiple emotions [75], also the identification of emotions is highly subjective [33]. Thus, low congruence in emotion annotation is quite likely [33, 63]. Regarding our human annotators, we expect at least the student annotators having basic knowledge about emotions and emotional theories due to their field of studies. Although even untrained annotators given considerable freedom can achieve high agreement rates [95], future studies might benefit from trained annotators achieving higher interrater agreements. Further, Kolog et al. [51] found correlations between intra-annotator agreements and the annotators' emotion, indicating that their emotions influence the labels used when annotating emotions in texts. Since the story used in the study dealt with the death of the protagonist's sister, individual participants could have been emotionally aroused by the story content while others were not. This could have led to different emotion perceptions.

Also, the amount of labels could have been a problem. In this study, all of Plutchik's subemotions were used [68] resulting in 24 emotion labels plus the "neutral" label. Volkova et al. [95] state that already 15 emotion labels exceed the typically used number of categories in emotion annotation. Generalizing the labels to the level of basic emotions improved the congruence between initial and re-annotated labels. Providing more labels obviously leads to higher variety which in turn can allow for more detailed representation of emotion. Although, the additional labels which were displayed with only marginal different postures may have impeded emotion recognition in our survey. Thus, the study should be replicated using a smaller subset of emotions.

Furthermore, it is noteworthy that the robot Nao utilized in the experiment only uses bodily expressions and eye LED colors to convey emotions. This could have led to weak emotion identification, since body language alone is not sufficient to recognize emotions and should be accompanied by facial expressions to enhance emotion perception [20, 103]. Our results show that – although only Nao's expression were manipulated – in the machine learning condition, in which the expression for anticipation was repeated 17 times, participants relied on the story's content more often compared to the manual annotation and LIWC conditions. Alike, in the LIWC condition with 17 neutral labels participants referred to the robot's voice more often compared to the other conditions. These results are in line with the finding that, especially when bodily and vocal expression do not match, the semantic modality is predominantly processed [98]. Thus the robot's voice and the thereby told story content may have been used by the participants to substitute misleading expression applied by the robot. In contrast, this result could also be due to a halo effect, the transfer of one known characteristic to other attributes of a person or entity [32]. Halo effects for social robots were already indicated in previous studies. For example, Yamashita et al. [100] found a relationship between robots' touch sensation and their personality impressions. Regarding storytelling robots, Appel et al. [7] found that voice was perceived more congruent to the story when a robotic storyteller used story-congruent facial expressions. This could explain the participants' recourse to the modality of voice even if it was not manipulated.

# 6 Evaluation

A laboratory user study was conducted to evaluate the three different annotation approaches with regard to the quality of the storytelling experience. As within Study I, we hypothesized that the way of annotating emotions leading to different emotion expressions displayed by the robot Nao [78] while telling the same story may influence the storytelling experience regarding transportation and cognitive absorption. Again, since related works highlight advantages of all three approaches, we postulate undirected hypotheses:

**H2a:** Transportation into the story differs between the three different annotation approaches.

**H2b:** Cognitive absorption by the story differs between the three different annotation approaches.

## 6.1 Method

A multivariate single-factor (human annotators vs. LIWC vs. machine learning) between groups design was applied to compare the three annotation approaches concerning their influence on the test persons' transportation and cognitive absorption.

Arriving at the laboratory, participants were first asked to disinfect their hands and have a seat at the sanitized table with the robot Nao [78], a monitor, mouse and keyboard on it. The monitor was turned off before the experiment to avoid distraction. Before taking part in the experiment, participants were both orally and written introduced into the study and gave written and informed consent. Afterwards, they were asked to attentively listen to the robot telling a story. Being randomly assigned to one of the three conditions, the robot told the story using the expressions indicated in the prestudy described in Sect. 4.2 matching the emotions labeled by the respective annotation approach. When the storytelling was finished, participants were asked to turn on the monitor and fill in the questionnaire. After completing the survey, the test persons were verbally informed about the study's research aim and left the laboratory after being thanked for their participation.

### 6.1.1 Measures

The *Transportation Scale Short Form* (TS-SF) [6] was applied to measure the recipients' transportation into the story. It includes six items, e.g. "I could picture myself in the scene of the events described in the narrative", that are anchored by a seven-point Likert-scale from 1 - "not at all" to 7 - "very much". While Appel et al. [6] reported Cronbach's Alpha of.80 up to.87, reliability calculated for the current sample was.90.

To measure the participants' state of involvement into the storytelling experience, the *Cognitive Absorption* (CA) questionnaire, originally developed by Agarwal et al. [1] to investigate software and web usage, was adapted to the robotic storytelling scenario. It comprises five scales, the (1) *Temporal Dissociation* scale which originally includes five items but was cut down to only the three items that refer to the participants' current state, e.g., "Time flied while the robot told the story.", the (2) *Focused Immersion* scale including five items, e.g., "While listening to the robot, I got distracted by other attentions very easily.", the (3) *Heightened Enjoyment* scale comprising four items, e.g., "I enjoyed using the robot.", the (4) *Control* scale including three items, e.g., "I feel that I have no control while listening to the robot.", and the (5) *Curiosity* scale which comprises three items, e.g., "Listening to the robot made me curious.". A seven-point Likert-scale anchored by 1 - "Strongly disagree", 4 - "Neutral", and 7 - "Strongly agree" was applied. Agarwal et al. [1] reported reliability values of.93 for the *Temporal Dissociation* as well as the *Heightened Enjoyment* and *Curiosity* scale,.88 for the *Focused Immersion* scale, and.83 for the *Control* scale. Cronbach's Alpha in the current sample was.85 for the *Temporal Dissociation* scale,.89 for the *Focused Immersion* scale,.93 for the *Heightened Enjoyment* scale,.70 for the *Control* scale, and.88 for the *Curiosity* scale.

Last, participants provided some demographic data, e.g., age and gender, and were able to give feedback in a comment box.

**Table 5** Participants' demographic data per group from Study II - Evaluation

| | ♀ | ♂ | Age | | *n* |
|---|---|---|---|---|---|
| | | | *M* | *SD* | |
| MA | 22 | 9 | 28.65 | 12.90 | 31 |
| LIWC | 22 | 9 | 23.61 | 3.59 | 31 |
| ML | 19 | 13 | 26.00 | 9.40 | 32 |

*MA* manual annotation, *ML* machine learning

### 6.1.2 Participants

Overall, 94 persons with an age ranging from 18 to 67 years (*M* = 26.09, *SD* = 9.40) took part in the study. Thirty-one of them reported themselves as male (age: *M* = 26.48, *SD* = 4.84), and 63 as females (age: *M* = 25.89, *SD* = 11.00). None of the test persons self-reported as diverse gender. Most participants were native speakers (*n* = 91), only three persons reported speaking German for more than three years. Being randomly assigned to one of the three conditions, both the human annotators' version and the LIWC version of the story were received by 31 participants each, while 32 persons saw the robot telling the story based on the machine learned annotations. Demographic data of the three groups are displayed in Table 5.

## 6.2 Results

All analyses were calculated using SPSS 26 and an alpha of.05. Descriptive values, anchored by 1 and 7, are displayed in Table 6. In general, the calculated means exceed the average of the subscales, except for *Control* from the *Cognitive Absorption* questionnaire which was evaluated less positively by the participants. The qualitative analysis of the comments was done using *MAXQDA2018* [93].

**Table 6** Descriptive data from Study II - Evaluation

| | MA | | LIWC | | ML | |
|---|---|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| TS | 4.47 | 2.36 | 3.76 | 1.45 | 3.96 | 1.35 |
| CA: TD | 4.30 | 1.57 | 3.39 | 1.18 | 4.44 | 1.45 |
| CA: FI | 4.37 | 1.38 | 3.89 | 1.24 | 4.07 | 1.37 |
| CA: HE | 4.90 | 1.53 | 4.15 | 1.13 | 4.88 | 1.33 |
| CA: CO | 2.89 | 1.46 | 2.51 | 0.98 | 2.51 | 1.22 |
| CA: CU | 5.11 | 1.38 | 4.41 | 1.40 | 4.57 | 1.49 |

*MA* manual annotation, *ML* machine learning, *TS* transportation, *CA* cognitive absorption, *TD* temporal dissociation, *FI* focused immersion, *HE* heightened enjoyment, *CO* control, *CU* curiosity. Calculated values range from 1 to 7

**Table 7** Verification of assumptions for analyses of Study II–Evaluation

| | *p* from Shapiro Wilk's test | | | *p* from |
|---|---|---|---|---|
| | MA | LIWC | ML | Levene's test |
| TS | .019* | .222 | .155 | .602 |
| CA: TD | .152 | .106 | .128 | .237 |
| CA: FI | .012* | .871 | .292 | .510 |
| CA: HE | .025* | .286 | .218 | .592 |
| CA: CO | .066 | .049* | .019* | .101 |
| CA: CU | .057 | .002** | .320 | .809 |

*MA* manual annotation, *ML* machine learning, *TS* transportation, *CA* cognitive absorption, *TD* temporal dissociation, *FI* focused immersion, *HE* heightened enjoyment, *CO* control, *CU* curiosity
* *p* <.05, ** *p* <.01

First, to compare the recipients' transportation between the three conditions (**H2a**), the respective values from the TS-SF were analyzed. As calculated using a Shapiro-Wilk test, the values were not normally distributed in the manual annotation's group ($p$ =.019), whereas Levene's test reported homogeneity of variances for the transportation values (see Table 7). Accordingly, Welch's ANOVA was calculated due to its robustness against violations of these assumptions indicating no significant effect of annotation approach on the users' transportation, Welch-Test $F(2, 60.51) = 2.15, p = .126$. $\omega^2$ = 0.02.

Second, to determine whether the approach of annotation influences the cognitive absorption (**H2b**), the CA scales were analyzed. Indicated by the Shapiro Wilk test's results displayed in Table 7, values were not normally distributed on the *Focused Immersion*, *Heightened Enjoyment*, *Control* and *Curiosity* scales. Box's test revealed homogeneity of covariance matrices ($p$ =.113) and Levene's test reported homogeneity of variances (see Table 7). Because it is robust against violation of normal distribution, especially when group sizes exceed 30, a one-way MANOVA was calculated. Results indicate no statistically significant difference between the annotations approaches on the combined CA scales, $F(10, 174) = 1.82$, $p$ =.060, $\eta^2$ =.10, Wilk's lambda =.82.. Post-hoc univariate ANOVAs showed significant group differences for the *Temporal Dissociation* scale ($F(2, 91)$ = 5.12, $p$ =.008, $\eta^2$ =.90) as well as the *Heightened Enjoyment* scale, $F(2, 91)$ = 3.23, $p$ =.044, $\eta^2$ =.07. Tukey HSD post-hoc analyses were calculated for both scales revealing significant differences between the manual annotation and the LIWC condition ($p$ =.033, $M_{Diff}$ = 0.914, 95%-CI[−1.77, −0.06 ]) as well as between the machine learning and LIWC condition ($p$ =.011, $M_{Diff}$ = 1.050, 95%-CI[−1.90, −0.20]) but not between the manual annotation and machine learning condition ($p$ =.136, $M_{Diff}$ = 0.020, 95%-CI[−0.71, 0.98]) on the *Temporal Dissociation* scale. Contradictory, pairwise tests did not show significant differences on the *Heightened Enjoyment* scale (*ps* >.05).

Due to technical errors, the robot's movements had to be adjusted during the experiment. The emotional postures were not changed, only the movement of its arms into selected positions was slowed down. To test whether this change in movement had an effect on the storytelling experience, a RM-MANOVA on the TS-SF values and AC scales comparing before and after was calculated. No significant difference was found, $F(6, 83) = 0.92$, $p$ =.484, Wilk's Λ =.94.

Approximately half of the participants ($n = 44$) gave comments after completing the experiment using the comment box at the end of the questionnaire. Most comments were provided by persons assigned to the machine learning condition ($n = 18$), while 14 participants who saw the manual annotated version of the story and twelve persons who received the LIWC version left comments. Only 9 (20.45%) of the comments included positive feedback, the remaining comments were negative. Most of the positive feedback was given for the LIWC version of the story ($n = 5$), whereas the manual annotated as well as machine learned story versions were only positively commented by two persons each. Positive feedback in all groups was related to the likability of the robot Nao ($n = 4$) and the gestures it performed, $n = 3$. Participants in the LIWC condition also liked the story told by the robot, $n = 2$.

The negative feedback was subdivided into four categories: feedback referring to the robot's movement or voice, to the story or other. Most of the negative feedback referred to the robot's movements in the machine learning condition. Most of these comments complained about the robot's motors' sounds being too loud ($n = 6$), also positions were repeated too often ($n = 4$) and the robot's movements were distracting, $n = 4$. Further, participants in this condition feared self-collisions of the robot ($n = 3$) and wished for less ($n = 1$) or other ($n = 1$) movements. In contrast, negative comments on the robot's movements on the manual annotated version only referred to the loud motor sounds ($n = 6$), being distracted by the robot's movements ($n = 3$) and the fear of self-collisions of the robot's body-parts, $n = 1$. Last, participants in the LIWC condition negatively commented on the robot's movements being too loud ($n = 4$), distracting ($n = 4$) and not matching the story told, $n = 4$.

Again, most of the negative comments concerning the robot's voice were provided in the machine learning condition. Fewest complains about the robot's voice were made in the LIWC version. The comments referred to the artificial sound of the text-to-speech voice that is not coming from the robot's mouth, missing pauses, intonation, and fluency of speech, the robot speaking too fast or skipping single words and thus being hard to understand. Frequencies per group are displayed in Table 8. Concerning the story told by the robot, two participants in the machine learning condition mentioned the story to be sad and one person in each of the three conditions stated to be confused by the story. Regarding further negative comments, participants in the machine learning condition complained about "beep"-sound the robot made during the story ($n = 1$), missing eye contact ($n = 1$) and being confused by the changing LED colors in the robot's eyes ($n = 2$). Last, one person who received the manually annotated version of the story described the robot being creepy.

**Table 8** Comments on the robot's voice per group from Study II–Evaluation

| Comment | MA | LIWC | ML |
|---|---|---|---|
| | $n$ | $n$ | $n$ |
| Hard to understand | 4 | 4 | 7 |
| Artificial voice | 1 | 1 | 2 |
| Speaking too fast | 2 | 1 | 3 |
| Missing pauses | 1 | 0 | 2 |
| Missing intonation | 3 | 1 | 3 |
| Skipping words | 1 | 1 | 0 |
| Non-fluent speech | 1 | 1 | 1 |
| Voice not deriving from robot's mouth | 1 | 0 | 0 |

*MA* manual annotation, *ML* machine learning

## 6.3 Discussion

The three annotation approaches of manual labeling by human annotators, annotation via LIWC and machine learned annotations were compared regarding their influence on the experience of a robotic storytelling scenario.

Results revealed no significant differences concerning the recipients' transportation into the story when applying the respective emotional postures. Thus, **Hypothesis 2a** was rejected. By trend the descriptive values for the manual annotation condition were higher compared to the LIWC and machine learning condition. Regarding the cognitive absorption, significant group differences were only found for the *Temporal Dissociation* scale, not for the whole cognitive absorption construct. Although the participants' tendency to loose their sense of time was higher in the manually annotated and the machine learned story version compared to the LIWC version, **Hypothesis 2b** had to be rejected, too.

Nevertheless, the descriptive values of the cognitive absorption dimensions as well as for transportation all show the same pattern. On each scale, the highest mean value was achieved in the manual annotation condition, whereas the LIWC condition always showed the lowest mean value. These results resemble the significant finding on the *Temporal Dissociation* scale, indicating a trend that manual annotations by human annotators might be the most favorable approach to prepare stories' texts for a robotic storyteller.

The low values provided for the *Control* subscale are noteworthy but expectable since participants had no option to influence the robot's storytelling or interact with it. In future iterations, the *Automated Robotic Storyteller* thus should be able to react to the listeners' reactions, such as perceived attention or social eye gaze. Also, it would be interesting to evaluate the three approaches in terms of social engagement with the robotic storyteller instead of only focusing on narrative engagement in the task in form of transportation.

The feedback provided was mostly negative. Remarkable is that participants in the machine learning condition noticed the robot's repeating postures, whereas persons in the LIWC condition commented on the expressions not matching the story. It is possible, that several expressions were applied with a wrong timing due to the implementation on the tokens' start. Thus the gesture-speech co-alignment could be misplaced at some points in the LIWC condition. However, none of these comments were given on the manually annotated version even though the expressions were applied at the beginning of a respective token in each condition. This finding emphasizes the suitability of manual annotations for robotic storytelling compared to the other annotation approaches included in the study. However, it conflicts with the lack of significant differences in transportation. As stated above, the descriptive trends follow the pattern revealed by participants' comments. Perhaps a larger sample size would uncover these differences.

Further, negative feedback was provided on the robot's voice. It is a well-known problem that basic text-to-speech systems only process single sentences without taking the context into account. This results in monotonous and thus tedious speech [24]. Especially in the storytelling use case, prosodic cues such as pitch, intensity, and tempo are important [37, 71] as these parameters are used to convey emotions [94]. Alike, pauses and their absence influence the storytelling experience. They are used for distinction between sections and sentences [94] as well as for backchanneling [90]. The negative perception of Nao's text-to-speech generated voice could impede both transportation and cognitive absorption. Hence the synthetic voice should be addressed in future studies. However, transportation was above average in our study.

## 7 General discussion

The goal of the studies presented within this paper was to determine which annotation approach should be used to develop an *Automated Robotic Storyteller*. The approaches of (1) manual annotation by human annotators, (2) semi-automated word-sensitive annotation using the text analysis program LIWC2015 [66], and (3) fully automated annotation via machine learning were tested using a scripted robotic storyteller. Therefore, emotional non-verbal expressions via body language matching the 24 subemotions of Plutchik's *Wheel of Emotions* [68] were determined in a preliminary survey, first. These expressions were utilized to manually implement three robotic storytelling scenarios for the robot Nao [78] using the same story but different annotation approaches. To identify the most suitable and thus favorable annotation approach two studies were carried out. For validation of the resulted annotations the robotic storytelling was re-annotated in an online study. Furthermore, the resulting storytelling scenarios were evaluated in a laboratory user study.

Taking the results from the validation into account, the machine learning based annotation approach must be disqualified. Participants were badly able to recognize the emotion depicted by the robot in this condition. This might indicate a mismatch between the emotions recipients expected based on the story and emotions labeled to the story in the machine learning approach. Our setting provides some specific challenges for the machine learning approach: unlike the other two approaches, it relies entirely on annotated training data, of which we only have a very limited amount. This is further exacerbated by the rather high number of emotions annotated in our data: some emotions appear less than ten times in the training data, making it very hard for the model to pick up the relevant signals. Although we have already pre-trained the model for the related task of binary polarity classification and additional pre-training on larger related datasets could be helpful, collecting more annotated data and potentially reducing the number of emotions is the most promising approach to improve the performance of the machine learning model. Moreover, several emotion expressions only achieved low recognition rates in the pre-study, which might have influenced the recognition of emotions in the validation study.

Concerning the evaluation, results revealed no significant differences between the three annotation approaches aside from participants more strongly loosing their sense of time when receiving the manually or machine learning annotated story version compared to the LIWC version. However, descriptive values follow this finding, indicating that annotations via LIWC might not be as suitable as the other approaches. Combining the findings from both studies, manually annotations of human annotators seem to be the most favorable way of annotation for robotic storytelling. In line, stories for robotic storytellers are annotated manually by human annotators in most studies, e.g. [7, 35, 84, 86, 87], currently needed to be scripted each time anew. However, our findings are not general: We only tested one specific machine learning approach, which may not be optimal for our setting. While we train the model to predict the most frequently annotated emotion in the human annotations, our goal is not to match the human annotations as closely as possible, but rather to provide labels that are optimal for the robotic storyteller. Therefore, an approach based on prompting a large language model (e.g., ChatGPT) may be more suitable than our current supervised approach.

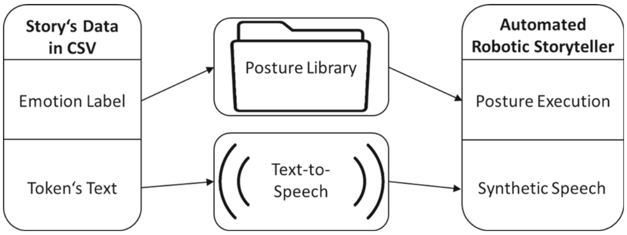

**Fig. 3** Overview of the implementation of the automated robotic storyteller

## 8 Implementation of the automated robotic storyteller framework

On the empirical basis described above the *Automated Robotic Storyteller* framework was developed. Since manual labeling of human annotators seems to be the most promising annotation approach, automatically preparing stories for the *Automated Robotic Storyteller* has to be replaced by generating a large database of annotated texts. Therefore, stories and their emotion labels are stored as CSV-files and can automatically be executed by the robot applying the respective emotional non-verbal expressions via body language. This allows for quick editing of the stories and easy expanding of the database.

The framework was implemented using Python 2.7[4] and the NAOqi Python SDK [77]. First, the emotional postures determined in Sect. 4.2 were stored based on the robot's joints' settings and numerical values. This way, they can be retrieved anytime and thus be used in our framework. Second, stories and their labels are read from the CSV-files. Doing this, the robot Nao converts the sentence automatically into synthetic speech via text-to-speech and plays the respective expression assigned to the emotion label read from the CSV-file, as shown in Fig. 3. This results in an automated robotic storytelling scenario.

The workflow for users of the *Automated Robotic Storyteller* framework can be divided into two parts. First, because manually labeling stories with emotions seems to be the most promising approach, the story which shall be presented by the robotic storyteller, has to be annotated by humans using the subemotions from Plutchik's *Wheel of Emotions* [68]. The tokenization is not restricted to full sentences but can be adjusted to the user's personal needs. Second, the annotated text is saved token-wise in a CSV-file. Line by line, a token's number, text, and the annotated emotion are entered. Last, the user specifies the file's path in the *Automated Robotic Storyteller* system and when executing the program, the story is automatically told by the robot accompanied by emotional body language.

In doing so, the *Automated Robotic Storyteller* framework facilitates the implementation of robotic storytelling. Every story can be played by the system without an individual implementation. This makes robotic storytelling also accessible for non-experts who are not familiar with (visual) coding. Furthermore, due to its basic implementation, the framework can be extended for further research, e.g., integrating voice modulation or sound effects.

## 9 General recommendations

Based in the studies' results, general recommendations on robotic storytelling can be identified. First, the cost of inappropriate body language seems to be higher compared to the cost of omitting emotion expressions. Thus, a confidence parameter for emotion expression should be taken into account acting as a threshold for executing the respective body language. If the threshold is not achieved, the context could be taken into account and emotional body language could be replaced by context-oriented behavior. Contextual movements could be iconic, metaphoric or deictic gestures [29], e.g., looking or pointing in a certain direction matching the text [87]. Alike, the timing of the emotion expression is important. Emotional behavior should be tightly connected to the respective story content to avoid inappropriate body language caused by temporal latency.

Further, a lot of pitfalls can be derived from the test persons' comments. Participants complained about the robot's eye contact and color. Since gazing is an important factor for the robot's anthropomorphism, likeability [48], and persuasion [42], robotic storytellers should be able to keep eye contact. Also, gazing could be used to draw the recipients' attention. Our participants comments indicate that eye color is not a reliable factor to recognize emotions, which is in line with findings by Häring et al. [43]. The authors instead suggest to focus on bodily expressions and sound. In doing so, the robot's synthetic voice produced via text-to-speech should be taken into account. Parameters such as tempo and intensity could be adjusted to the story content [71], e.g., slowing down and decreasing intensity at sad moments and increasing both speed and intensity when fear or anger are conveyed [94]. This could be fostered by applying facial expressions additionally to the body language used [53, 103], facilitating the emotion recognition.

## 10 Future work

The above described limitations and recommendations indicate further possible extensions and revisions of the *Automated Robotic Storyteller* framework. The fluctuating interrater agreement should be included into the execution of

[4] http://python.org/download/

bodily expressions in order to prevent from applying inappropriate expressions. In future iterations, the annotators' agreement will be used as a confidence parameter. Emotion labels with values above a certain threshold can be used for non-verbal expression execution, whereas labels with values under this threshold are ignored and the respective expressions are replaced by contextual movements of the robot. Further, eye color will not be part of the expressions. Regarding the emotion labels, also a smaller subset of emotions may improve the recognition of recipients by decreasing the variety of nearly similar expressions for subemotions from the same family, while at the same time making the training task for the machine learning approach much easier. When replicating the experiments with a smaller subset of emotions, also the tokenization will be changed. As [95] suggest, humans prefer short tokens for annotations of about four to seven words. In addition, by shortening the tokens, the emotions would be more precisely matched to the story's text, e.g. commas could be additionally used to divide tokens instead of using full sentences. Considering the studies' design, the approaches will further be compared to a baseline, in which, for instance, no emotion expressions or random expressions are shown. In addition, a manipulation check verifying the robot as an *emotional* storyteller will be included.

Since relying on the fully automated machine learning approach is a desirable for the future, enabling us to use a much wider range of stories without the need to collect manual annotations, we will also work towards improving the performance of this approach. Apart from reducing the number of emotion classes, mentioned above, it is also promising to explore additional pre-training datasets. Recent research [16] suggests that we may be able to use datasets with different emotion categories that the ones used in our work by mapping them into a shared embedding space, making it much easier to find suitable datasets for additional pre-training. Since out ML-based approach tends to produce more non-neutral labels than the other modalities, we may also consider a two-stage classification, where we first detect whether there is any emotion represented in the text and then classify the specific emotion. This would give us an easier way to control the number of non-neutral emotion labels. In addition, our setting seems particularly suitable for prompting-based methods, which have become popular in NLP [15].

Additionally, future versions of the *Automated Robotic Storyteller* should be able to keep eye contact and shift gazes based on a story's context. Also, voice modulation based on emotion labels could be integrated. In contrast, the robot's changing eye color should be removed from the framework when non-verbal expressions are revised. Last, the robot Nao, which was utilized in this paper, is not capable of showing facial expressions, thus the experiments should be replicated using further robots which allow for mimic art.

## 11 Conclusion

Since conveying emotions in robotic storytelling scenarios is crucial for both story comprehension and storytelling experience, robots should be capable of emotional expressiveness, e.g. using body language. If this emotion expression can be generated automatically when given a story's text, it is open for a wide application context. To develop the *Automated Robotic Storyteller* based on empirical methods, we compared three approaches of annotating emotions to a given text, i.e. manual annotation by human annotators, word-sensitive annotation via LIWC, and a machine learning based approach. The annotations were based on the *Wheel of Emotions* [68] and tied to emotional non-verbal expressions via body language indicated in a preliminary study. Results from the validation show that emotions derived by the machine learning approach are worst recognized, whereas recognition of the emotions labeled by LIWC [66] closely followed by the human annotators achieved the best outcomes. Contradictory, evaluation results from the user evaluation indicated a trend towards manually annotated and machine learning approach, however, no significant differences were found regarding storytelling experience operationalized with transportation and cognitive absorption. Based on these findings, we chose the approach of manual annotations by human annotators for our *Automated Robotic Storyteller* framework because it seems to be most suitable for labeling emotions to stories. We implemented the system using the *Nao* [78] robot. In future investigations, the framework will be expanded using contextual movements of the robots, including confidence of the emotion labels into the behavior decision process, and the framework shall be tested using further robots offering additional modalities, e.g., facial expressions.

**Funding** Open Access funding enabled and organized by Projekt DEAL.

## Appendix A Pose library

See Figs. 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27.

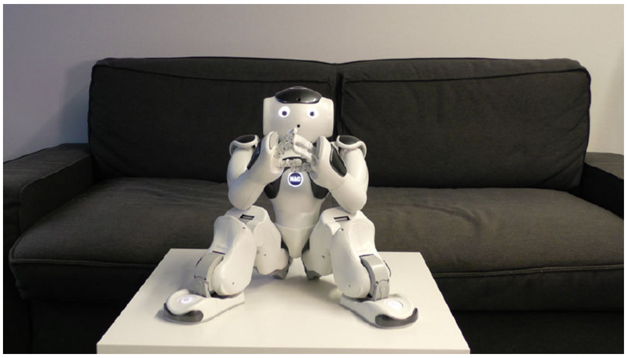

**Fig. 4** Pose for interest

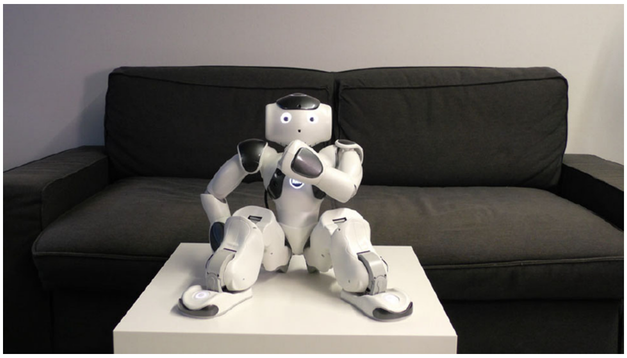

**Fig. 5** Pose for anticipation

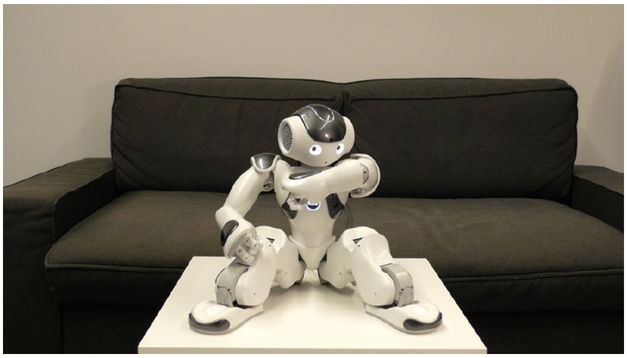

**Fig. 6** Pose for vigilance

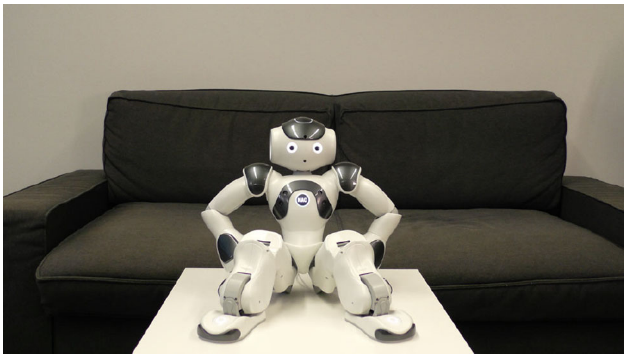

**Fig. 7** Pose for serenity

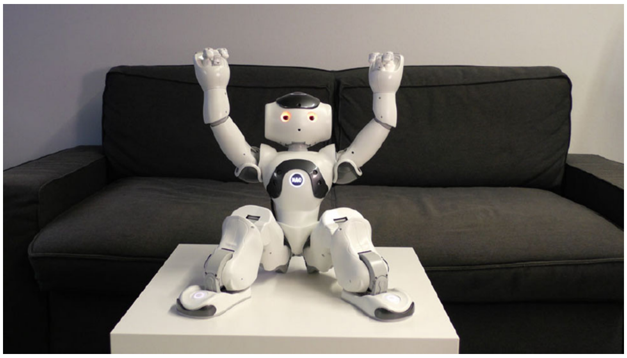

**Fig. 8** Pose for joy

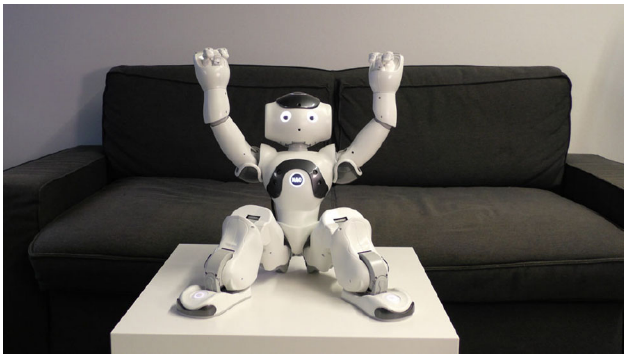

**Fig. 9** Pose for ecstasy

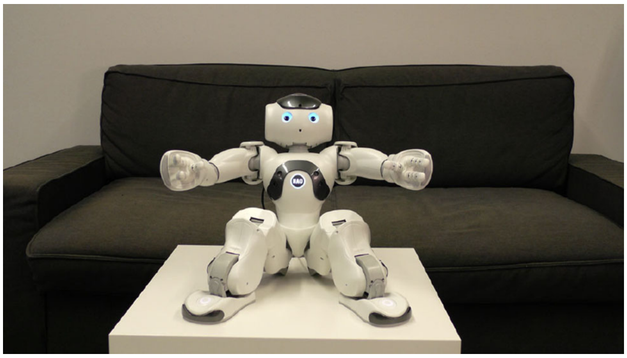

**Fig. 10** Pose for acceptance

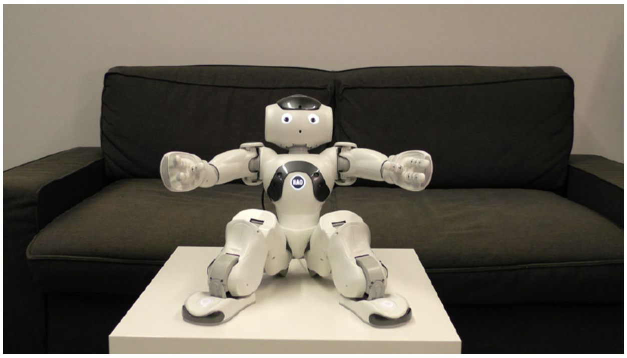

**Fig. 11** Pose for trust

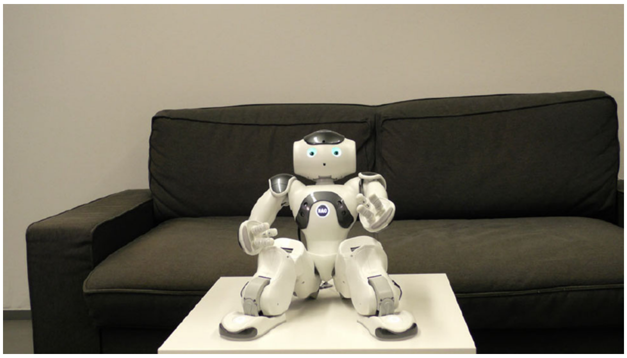

**Fig. 12** Pose for admiration

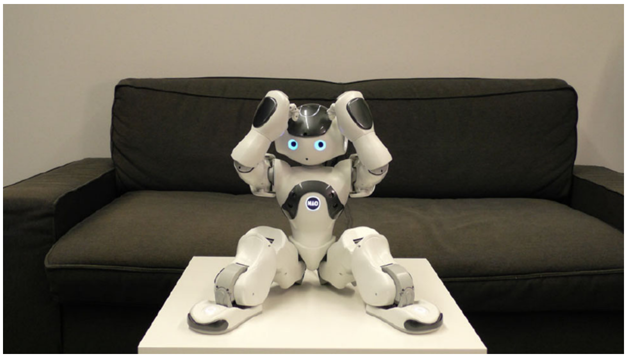

**Fig. 13** Pose for apprehension

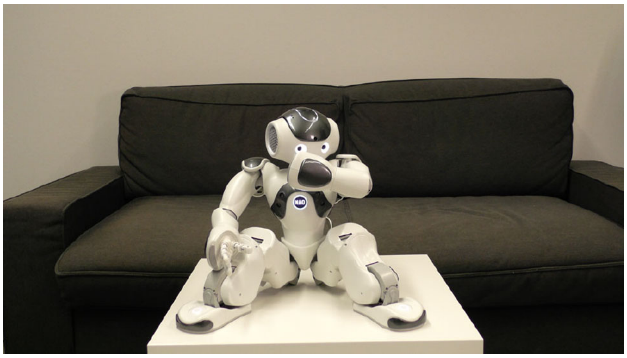

**Fig. 14** Pose for fear

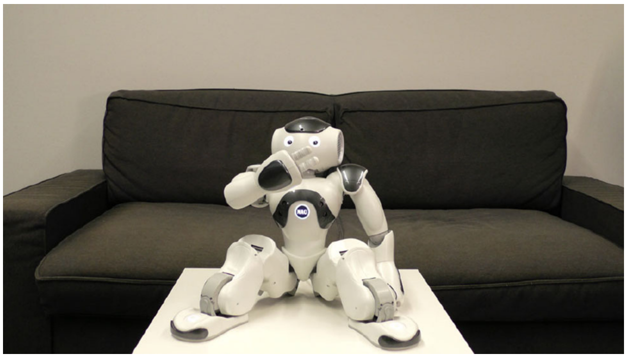

**Fig. 15** Pose for terror

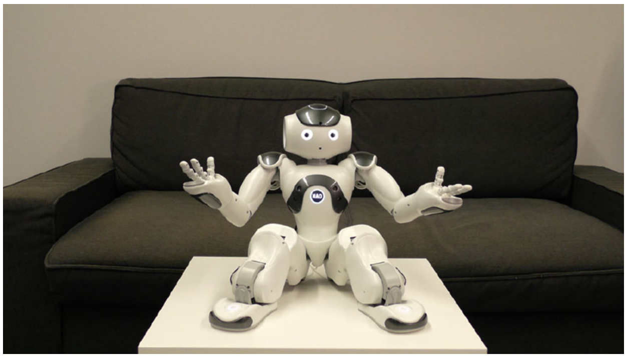

**Fig. 16** Pose for distraction

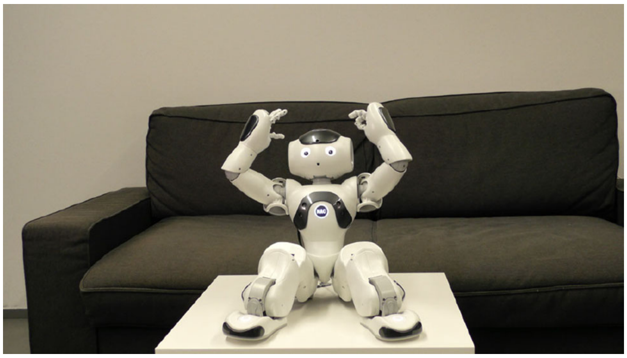

**Fig. 17** Pose for surprise

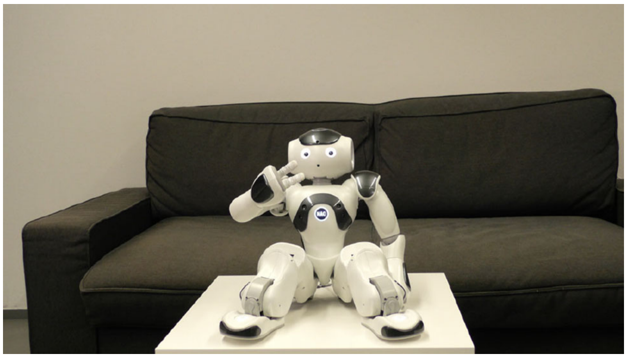

**Fig. 18** Pose for amazement

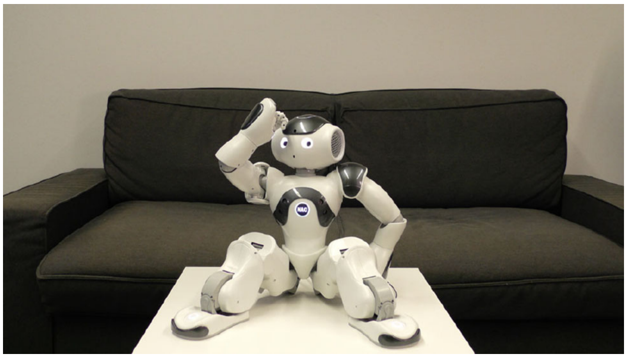

**Fig. 19** Pose for pensiveness

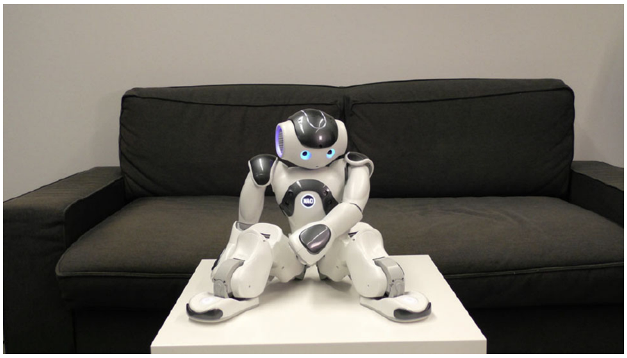

**Fig. 20** Pose for sadness

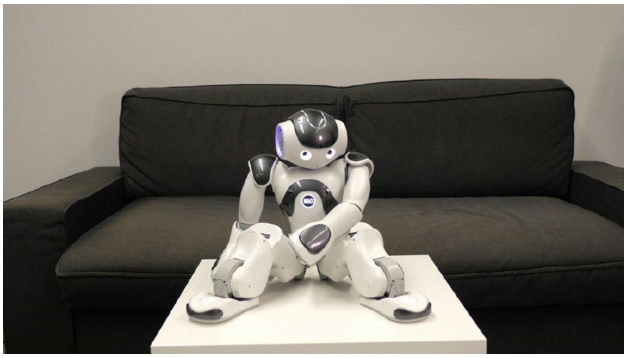

**Fig. 21** Pose for grief

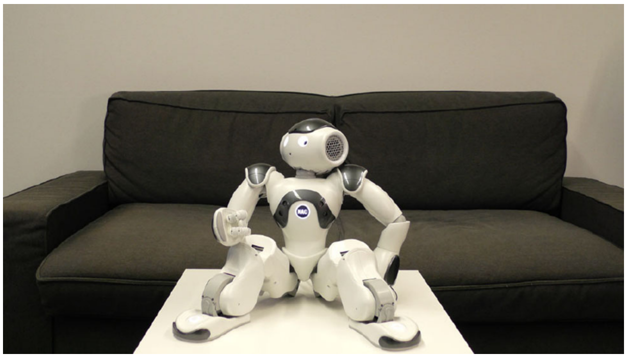

**Fig. 22** Pose for boredom

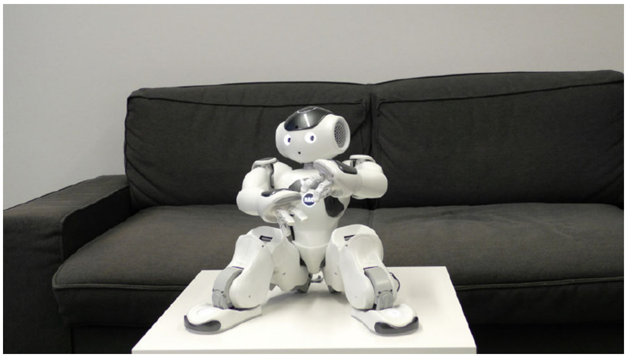

**Fig. 23** Pose for disgust

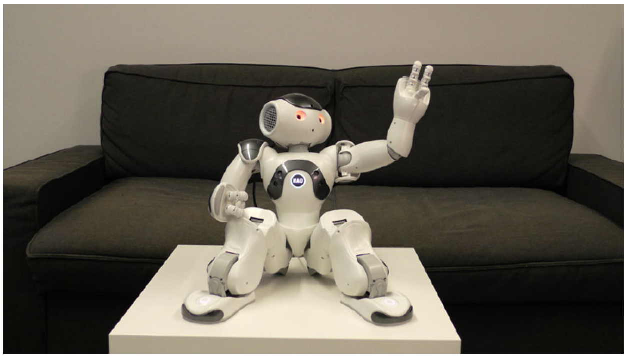

**Fig. 24** Pose for loathing

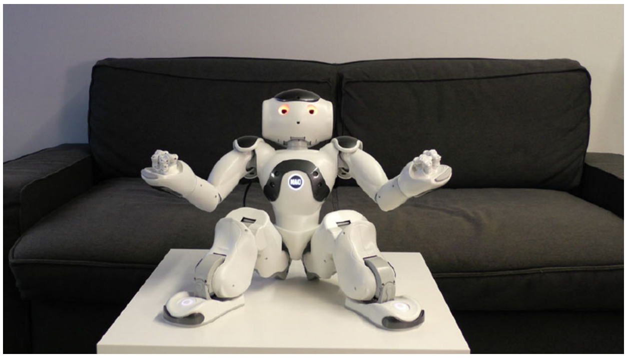

**Fig. 25** Pose for annoyance

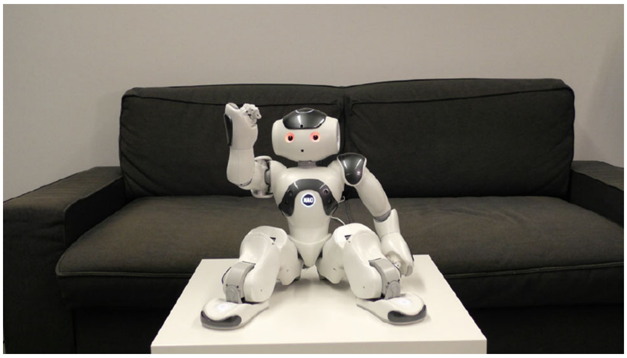

**Fig. 26** Pose for anger

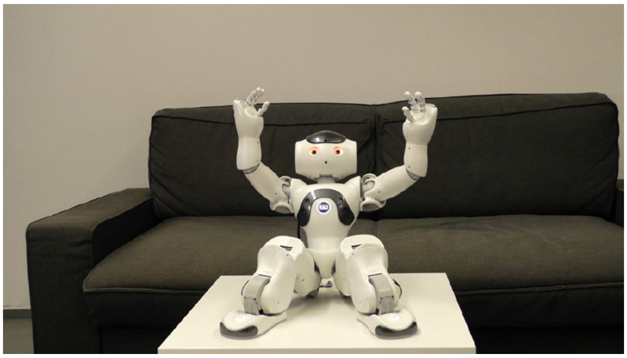

**Fig. 27** Pose for rage

## References

1. Agarwal R, Karahanna E (2000) Time flies when you're having fun: cognitive absorption and beliefs about information technology usage. MIS Q 24(4):665. https://doi.org/10.2307/3250951
2. Ahn Le Q, d'Alessandro C, Deroo O, et al (2010) Towards a storytelling humanoid robot. In: Association for the advancement of artificial (ed) 2010 AAAI Fall Symposium Series
3. Aldebaran Robotics (2016) Choregraphe [Software] https://www.ald.softbankrobotics.com/en
4. Alexandrova IV, Volkova EP, Kloos U, et al (2010) Short paper: virtual storyteller in immersive virtual environments using fairy tales annotated for emotion states. Citeseer 65–68
5. Alm CO, Roth D, Sproat R (2005) Emotions from text: machine learning for text-based emotion prediction. In: Mooney RJ (ed) Proceedings of the conference on human language technology and empirical methods in natural language processing-HLT '05. Association for Computational Linguistics, Morristown, NJ, USA, pp 579–586. https://doi.org/10.3115/1220575.1220648
6. Appel M, Gnambs T, Richter T et al (2015) The transportation scale-short form (TS-SF). Media Psychol 18(2):243–266. https://doi.org/10.1080/15213269.2014.987400
7. Appel M, Lugrin B, Kühle M et al (2021) The emotional robotic storyteller: on the influence of affect congruency on narrative transportation, robot perception, and persuasion. Comput Hum Behav 120(106):749. https://doi.org/10.1016/j.chb.2021.106749
8. Augello A, Pilato G (2019) An annotated corpus of stories and gestures for a robotic storyteller. In: 2019 Third IEEE international conference on robotic computing (IRC). IEEE, pp 630–635. https://doi.org/10.1109/IRC.2019.00127
9. Augello A, Infantino I, Maniscalco U, et al (2019) Narrob: a humanoid social storyteller with emotional expression capabilities. In: Samsonovich AV (ed) Biologically inspired cognitive architectures 2018, Advances in Intelligent Systems and Computing, vol 848. Springer International Publishing, Cham, pp 9–15. https://doi.org/10.1007/978-3-319-99316-4_2
10. Aylett R (2022) Interactive narrative and story-telling. In: Lugrin B, Pelachaud C, Traum D (eds) The handbook on socially interactive agents-volume 2: interactivity, platforms, application. 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics, Association for Computing Machinery, pp 463–491
11. Beck A, Stevens B, Bard KA (2009) Comparing perception of affective body movements displayed by actors and animated characters. In: Proceedings of the symposium on mental states, emotions, and their embodiment, pp 169–178
12. Beck A, Canamero L, Bard KA (2010) Towards an affect space for robots to display emotional body language. In: 19th international symposium in robot and human interactive communication. IEEE, pp 464–469. https://doi.org/10.1109/ROMAN.2010.5598649
13. Beck A, Cañamero L, Damiano L, et al (2011) Children interpretation of emotional body language displayed by a robot. In: Social robotics, lecture notes in computer science, vol 7072. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 62–70. https://doi.org/10.1007/978-3-642-25504-5_7
14. Bono A, Augello A, Pilato G et al (2020) An act-r based humanoid social robot to manage storytelling activities. Robotics 9(2):25. https://doi.org/10.3390/robotics9020025
15. Brown TB, Mann B, Ryder N, et al (2020) Language models are few-shot learners. arXiv preprint arXiv:2005.14165
16. Buechel S, Modersohn L, Hahn U (2021) Towards label-agnostic emotion embeddings. In: Proceedings of the 2021 conference on empirical methods in natural language processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp 9231–9249. https://doi.org/10.18653/v1/2021.emnlp-main.728, https://aclanthology.org/2021.emnlp-main.728
17. Cassell J, McNeill D (1991) Gestures and the poetics of prose. Poetics Today 12(3):375–404
18. Cassell J, Vilhjálmsson HH, Bickmore T (2001) Beat: the behavior expression animation toolkit. In: Pocock L (ed) Proceedings of the 28th annual conference on Computer graphics and interactive techniques. ACM, New York, NY, pp 477–486
19. Cho K, van Merriënboer B, Gulcehre C, et al (2014) Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). Association for Computational Linguistics, Doha, Qatar, pp 1724–1734. https://doi.org/10.3115/v1/D14-1179, https://www.aclweb.org/anthology/D14-1179
20. Clavel C, Plessier J, Martin JC, et al (2009) Combining facial and postural expressions of emotions in a virtual character. In: Intelligent virtual agents, lecture notes in computer science, vol 5773. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 287–300. https://doi.org/10.1007/978-3-642-04380-2_31
21. Costa S, Brunete A, Bae BC et al (2018) Emotional storytelling using virtual and robotic agents. Int J Human Robot 15(03):1850006. https://doi.org/10.1142/S0219843618500068
22. Declerck T, Scheidel A, Lendvai P (2011) Proppian content descriptors in an integrated annotation schema for fairy tales. In: Language technology for cultural heritage. Springer, pp 155–170, Berlin
23. Devlin J, Chang MW, Lee K, et al (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
24. Doukhan D, Rosset S, Rilliard A, et al (2012) Text and speech corpora for text-to-speech synthesis of tales. In: Proceedings of the 8th international conference on language resources and evaluation, pp 1003–1010
25. Duffy BR, Rooney C, O'Hare GMP, et al (1999) What is a social robot? In: 10th Irish conference on artificial intelligence & cognitive science. http://hdl.handle.net/10197/4412
26. Ekman P (1984) Expression and the nature of emotion. Approaches to emotion 3(19):344
27. Ekman P (ed) (1997) What the face reveals: basic and applied studies of spontaneous expression using the facial action coding system (FACS). Series in affective science, Oxford Univ. Press, New York, NY. http://www.loc.gov/catdir/enhancements/fy0605/96036655-d.html
28. Ekman P (1999) Basic emotions. In: Dalgleish T, Power MJ (eds) Handbook of cognition and emotion. Wiley, Chichester, pp 45–60. https://doi.org/10.1002/0470013494.ch3
29. Ekman P (2004) Emotional and conversational nonverbal signals. In: Language, knowledge, and representation. Academic Publishers, pp 39–50, Cambridge
30. El Maarouf I, Villaneau J (eds) (2012) A French Fairy Tale Corpus syntactically and semantically annotated
31. Erden MS (2013) Emotional postures for the humanoid-robot nao. Int J Soc Robot 5(4):441–456
32. Forgas JP, Laham SM (2016) Halo effects. In: Pohl R (ed) Cognitive illusions. Ebrary online, Routledge, Abingdon, Oxon, pp 276–290
33. Francisco V, Hervás R, Peinado F et al (2012) Emotales: creating a corpus of folk tales with emotional annotations. Lang Resour Eval 46(3):341–381
34. Frijda NH (2001) The emotions. In: Studies in emotion and social interaction, Cambridge University Press, Cambridge
35. Gelin R, d'Alessandro C, Anh Le Q, et al (2010) Towards a storytelling humanoid robot. In: Dialog with robots. Association for the Advancement of Artificial Intelligence, pp 137–138

36. Giordano R (2018) Wordlist Maker-list unique words, count total words. https://design215.com/toolbox/wordlist.php
37. Goossens N, Aarts R, Vogt P (2019) Storytelling with a social robot. Robots for Learning R4L
38. Green MC, Brock TC (2000) The role of transportation in the persuasiveness of public narratives. J Pers Soc Psychol 79(5):701–721. https://doi.org/10.1037//0022-3514.79.5.701
39. Grzyb B, Vigliocco G (2020) Beyond robotic speech: mutual benefits to cognitive psychology and artificial intelligence from the study of multimodal communication. https://doi.org/10.31234/osf.io/h5dxy
40. Haas M (2014) Weakly supervised learning for compositional sentiment recognition. PhD thesis, Heidelberg University
41. Habermas T (2011) Moralische Emotionen: Ärger in Alltagserzählungen. Jenseits des Individuums-Emotion und Organisation, Vandenhoeck Ruprecht, Göttingen 329:1–350
42. Ham J, Bokhorst R, Cuijpers R, et al (2011) Making robots persuasive: the influence of combining persuasive strategies (gazing and gestures) by a storytelling robot on its persuasive power. In: Social robotics, lecture notes in computer science, vol 7072. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 71–83. https://doi.org/10.1007/978-3-642-25504-5_8
43. Häring M, Bee N, André E (2011) Creation and evaluation of emotion expression with body movement, sound and eye color for humanoid robots. IEEE, pp 204–209
44. Hashimoto T, Hitramatsu S, Tsuji T, et al (2006) Development of the face robot saya for rich facial expressions. In: 2006 SICE-ICASE international joint conference. IEEE, pp 5423–5428. https://doi.org/10.1109/SICE.2006.315537
45. Hegel F, Muhl C, Wrede B, et al (2009) Understanding social robots. In: 2009 second international conferences on advances in computer-human interactions. IEEE, pp 169–174. https://doi.org/10.1109/ACHI.2009.51
46. Iovino M, Scukins E, Styrud J et al (2022) A survey of behavior trees in robotics and AI. Robot Auton Syst 154:104096. https://doi.org/10.1016/j.robot.2022.104096
47. Izui T, Milleville I, Sakka S, et al (2015) Expressing emotions using gait of humanoid robot. IEEE, pp 241–245
48. Karreman D, Sepulveda Bradford G, van Dijk B, et al (2013) What happens when a robot favors someone? How a tour guide robot uses gaze behavior to address multiple persons while storytelling about art. In: 2013 8th ACM/IEEE international conference on human-robot interaction (HRI). IEEE, pp 157–158. https://doi.org/10.1109/HRI.2013.6483549
49. Kim E, Klinger R (2018) A survey on sentiment and emotion analysis for computational literary studies. arXiv preprint arXiv:1808.03137
50. Kim E, Klinger R (2019) Frowning Frodo, wincing Leia, and a seriously great friendship: learning to classify emotional relationships of fictional characters. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, pp 647–653. https://doi.org/10.18653/v1/N19-1067, https://www.aclweb.org/anthology/N19-1067
51. Kolog EA, Montero CS, Sutinen E (2016) Annotation agreement of emotions in text: the influence of counsellors' emotional state on their emotion perception. In: 2016 IEEE 16th international conference on advanced learning technologies (ICALT). IEEE, pp 357–359. https://doi.org/10.1109/ICALT.2016.21
52. Krcadinac U, Pasquier P, Jovanovic J et al (2013) Synesketch: an open source library for sentence-based emotion recognition. IEEE Trans Affect Comput 4(3):312–325. https://doi.org/10.1109/T-AFFC.2013.18
53. Kret ME, Stekelenburg JJ, Roelofs K et al (2013) Perception of face and body expressions using electromyography, pupillometry and gaze measures. Front Psychol 4:28. https://doi.org/10.3389/fpsyg.2013.00028
54. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33(1):159. https://doi.org/10.2307/2529310
55. LimeSurvey GmbH (2021) LimeSurvey https://www.limesurvey.org/de/
56. Liu B (2020) Sentiment analysis: mining opinions, sentiments, and emotions. Cambridge University Press, Cambridge
57. Lombardo V, Damiano R (2012) Semantic annotation of narrative media objects. Multimed Tools Appl 59(2):407–439
58. Lovecraft HP (1959) The secret cave or John Lee's adventure. Arkham House, Sauk City
59. Lugrin B, Pelachaud C, Traum D (2021) The handbook on socially interactive agents. ACM, New York. https://doi.org/10.1145/3477322
60. Mehrabian A (2017) Nonverbal communication. Routledge, Taylor and Francis Group, Abingdon, Oxon and New York
61. Munezero M, Montero CS, Mozgovoy M, et al (2013) Exploiting sentiment analysis to track emotions in students' learning diaries. In: Laakso MJ, Simon (eds) Proceedings of the 13th Koli calling international conference on computing education research-Koli Calling '13. ACM Press, New York, New York, USA, pp 145–152. https://doi.org/10.1145/2526968.2526984
62. Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends® Inf Retr 2(1–2):1–135. https://doi.org/10.1561/1500000011
63. Park SB, Yoo E, Kim H et al (2011) Automatic emotion annotation of movie dialogue using WordNet. Springer, Berlin, pp 130–139
64. Pelachaud C, Gelin R, Martin JC, et al (2010) Expressive gestures displayed by a humanoid robot during a storytelling application. In: AISB'2010 symposium new frontiers in human-robot interaction. Leicester
65. Pelachaud C, Busso C, Heylen D (2021) Multimodal behavior modeling for socially interactive agents. In: Lugrin B, Pelachaud C, Traum D (eds) The handbook on socially interactive agents. ACM, New York, pp 259–310. https://doi.org/10.1145/3477322.3477331
66. Pennebaker JW (2015) LIWC2015 [Software]. https://liwc.wpengine.com/
67. Pennebaker JW, Boyd RL, Jordan K, et al (2015) The development and psychometric properties of LIWC2015
68. Plutchik R (1982) A psychoevolutionary theory of emotions. Soc Sci Inf 21(4–5):529–553. https://doi.org/10.1177/053901882021004003
69. Plutchik R (2001) The nature of emotions: human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. Am Sci 89(4):344–350
70. Qureshi AH, Nakamura Y, Yoshikawa Y et al (2018) Intrinsically motivated reinforcement learning for human-robot interaction in the real-world. Neural Netw Off J Int Neural Netw Soc 107:23–33. https://doi.org/10.1016/j.neunet.2018.03.014
71. Ramli I, Jamil N, Seman N et al (2018) The first Malay language storytelling text-to-speech (TTS) corpus for humanoid robot storytellers. J Fundam Appl Sci 9(4S):340. https://doi.org/10.4314/jfas.v9i4s.20
72. robopec (2021) Reeti: an expressive and communicating robot [Hardware]. https://www.robopec.com/en/constructions/others/reeti-robopec/
73. Russell JA (1980) A circumplex model of affect. J Pers Soc Psychol 39(6):1161–1178. https://doi.org/10.1037/h0077714
74. Salem M, Eyssel F, Rohlfing K, et al (2011) Effects of gesture on the perception of psychological anthropomorphism: a case study with a humanoid robot. In: Social robotics, lecture

notes in computer science, vol 7072. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 31–41. https://doi.org/10.1007/978-3-642-25504-5_4
75. Seyeditabari A, Tabari N, Zadrozny W (????) Emotion detection in text: a review. https://doi.org/10.48550/arXiv.1806.00674
76. Simon-Kucher & Partners (2020) Welche genres lesen sie unabhängig vom format? https://de.statista.com/statistik/daten/studie/1189038/umfrage/gelesene-genres-von-buechern/
77. SoftBank Robotics (2017) NAOqi Python SDK [Software]
78. SoftBank Robotics (2018) NAO: V6 [Hardware] https://www.softbankrobotics.com/emea/en/nao
79. SoftBank Robotics (2021) Pepper [Hardware]. https://www.softbankrobotics.com/emea/en/pepper
80. Song S, Yamada S (2017) Expressing emotions through color, sound, and vibration with an appearance-constrained social robot. In: Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction-HRI '17. ACM Press, New York, New York, USA, pp 2–11. https://doi.org/10.1145/2909824.3020239
81. Steinhaeusser SC, Lugrin B (2022) Effects of colored leds in robotic storytelling on storytelling experience and robot perception. In: 2022 17th ACM/IEEE international conference on human-robot interaction (HRI), IEEE, pp 1053–1058
82. Steinhaeusser SC, Lugrin B (in press) Eeffects of number of voices and voice type on storytelling experience and robot perception. In: Savery R (ed) Sound and robotics. CRC Press, pp 9–32, Boca Raton
83. Steinhaeusser SC, Gabel JJ, Lugrin B (2021a) Your new friend nao vs. robot no. 783-effects of personal or impersonal framing in a robotic storytelling use case. In: Companion of the 2021 ACM/IEEE international conference on human-robot interaction. ACM, New York, NY, USA, pp 334–338. https://doi.org/10.1145/3434074.3447187
84. Steinhaeusser SC, Schaper P, Bediako Akuffo O, et al (2021b) Anthropomorphize me! Effects of robot gender on listeners' perception of the social robot NAO in a storytelling use case. In: Companion of the 2021 ACM/IEEE international conference on human-robot interaction. ACM, New York, NY, USA, pp p 529–534. https://doi.org/10.1145/3434074.3447228
85. Steinhaeusser SC, Schaper P, Lugrin B (2021c) Comparing a robotic storyteller versus audio book with integration of sound effects and background music. In: Companion of the 2021 ACM/IEEE international conference on human-robot interaction. ACM, New York, NY, USA, pp 328–333. https://doi.org/10.1145/3434074.3447186
86. Striepe H, Lugrin B (2017) There once was a robot storyteller: measuring the effects of emotion and non-verbal behaviour. In: Social robotics, lecture notes in computer science, vol 10652. Springer International Publishing, Cham, pp 126–136. https://doi.org/10.1007/978-3-319-70022-9_13
87. Striepe H, Donnermann M, Lein M, et al (2019) Modeling and evaluating emotion, contextual head movement and voices for a social robot storyteller. Int J Soc Robot 1–17. https://doi.org/10.1007/s12369-019-00570-7
88. Tsiourti C, Weiss A, Wac K et al (2019) Multimodal integration of emotional signals from voice, body, and context: effects of (in)congruence on emotion recognition and attitudes towards robots. Int J Soc Robot 11(4):555–573. https://doi.org/10.1007/s12369-019-00524-z
89. Valdez P, Mehrabian A (1994) Effects of color on emotions. J Exp Psychol Gen 123(4):394
90. van Laer T, de Ruyter K, Visconti LM et al (2014) The extended transportation-imagery model: a meta-analysis of the antecedents and consequences of consumers' narrative transportation. J Consum Res 40(5):797–817. https://doi.org/10.1086/673383
91. Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. In: Advances in neural information processing systems, vol 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
92. Vaughn LA, Hesse SJ, Petkova Z et al (2009) this story is right on: the impact of regulatory fit on narrative engagement and persuasion. Eur J Soc Psychol 39(3):447–456. https://doi.org/10.1002/ejsp.570
93. VERBI GmbH (2018) MAXQDA2018 [Software]
94. Verma R, Sarkar P, Rao KS (2015) Conversion of neutral speech to storytelling style speech. In: 2015 eighth international conference on advances in pattern recognition (ICAPR). IEEE, pp 1–6. https://doi.org/10.1109/ICAPR.2015.7050705
95. Volkova EP, Mohler BJ, Meurers D, et al (2010) Emotional perception of fairy tales: achieving agreement in emotion annotation of text. In: Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text. Association for Computational Linguistics, USA, CAAGET '10, pp 98–106
96. Weisbuch M, Ambady N, Clarke AL et al (2010) On being consistent: the role of verbal-nonverbal consistency in first impressions. Basic Appl Soc Psychol 32(3):261–268. https://doi.org/10.1080/01973533.2010.495659
97. Wolf M, Horn AB, Mehl MR et al (2008) Computergestützte quantitative Textanalyse: Äquivalenz und Robustheit der deutschen Version des Linguistic Inquiry and Word Count. Diagnostica 54(2):85–98
98. Xu J, Broekens J, Hindriks K, et al (2015) Effects of a robotic storyteller's moody gestures on storytelling perception. IEEE, pp 449–455
99. Xu J, Broekens J, Hindriks K, et al (92014) Effects of bodily mood expression of a robotic teacher on students. In: 2014 IEEE/RSJ international conference on intelligent robots and systems. IEEE, pp 2614–2620. https://doi.org/10.1109/IROS.2014.6942919
100. Yamashita Y, Ishihara H, Ikeda T, et al (2016) Path analysis for the halo effect of touch sensations of robots on their personality impressions. In: Social robotics, lecture notes in computer science, vol 9979. Springer International Publishing, Cham, pp 502–512. https://doi.org/10.1007/978-3-319-47437-3_49
101. Yin D, Meng T, Chang KW (2020) SentiBERT: a transferable transformer-based architecture for compositional sentiment semantics. In: Proceedings of the 58th annual meeting of the association for computational linguistics. Association for Computational Linguistics, Online, pp 3695–3706. https://doi.org/10.18653/v1/2020.acl-main.341, https://www.aclweb.org/anthology/2020.acl-main.341
102. Zabala U, Rodriguez I, Lazkano E (2022) Towards an automatic generation of natural gestures for a storyteller robot. In: 2022 31st ieee international conference on robot and human interactive communication (RO-MAN). IEEE, pp 1209–1215. https://doi.org/10.1109/RO-MAN53752.2022.9900532
103. Zecca M, Mizoguchi Y, Endo K, et al (2009) Whole body emotion expressions for kobian humanoid robot—preliminary experiments with different emotional patterns. In: RO-MAN 2009-the 18th IEEE international symposium on robot and human interactive communication. IEEE, pp 381–386. https://doi.org/10.1109/ROMAN.2009.5326184
104. Zehe A, Becker M, Jannidis F, et al (2017) Towards sentiment analysis on german literature. In: Joint German/Austrian conference on artificial intelligence (Künstliche Intelligenz), Springer, pp 387–394
105. Zehe A, Arns J, Hettinger L, et al (2020) Harrymotions-classifying relationships in harry potter based on emotion analysis. In: 5th SwissText & 16th KONVENS joint conference
106. Zhang Z, Niu Y, Wu S, et al (2018) Analysis of influencing factors on humanoid robots' emotion expressions by body language. In: Huang T, Lv J, Sun C, et al (eds) Advances in neural networks–

ISNN 2018, Lecture Notes in Computer Science, vol 10878. Springer International Publishing, Cham, pp 775–785. https://doi.org/10.1007/978-3-319-92537-0_88

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.