Storytelling by a Show of Hands: A framework for interactive embodied storytelling in robotic agents

Philipp Wicke[1] and **Tony Veale**[2]

Abstract. With the increasing availability of commercial humanoid robots, the domain of computational storytelling has found a tool that combines linguistics with its physical originator, the body. We present a framework that evolves previous research in this domain, from a focus on the analysis of expressiveness towards a focus on the potential for creative interaction between humans and robots. A single story may be rendered in many ways, but embodiment is one of the oldest and most natural, and does more to draw people into the story. While a robot provides the physical means to generate an infinite number of stories, we want to hear stories which are more than the products of *mere generation*. In the framework proposed here, we let the robot ask specific questions to tailor the creation process to the experiences of the human user. This framework offers a new basis for investigating important questions in Human-Robot-Interaction, Computational creativity, and Embodied Storytelling.

1 INTRODUCTION

When was the last time that a story touched and inspired you? Was it made up of words in a book or pictures on a screen, or was it, perhaps, delivered in a song or recited by actors in a play? There are as many different ways of telling a story as there are stories to tell. Nonetheless, all storytellers share the same goal: to express something internal. This subjective *something* is most likely an emotion, insight, experience or abstract concept that cannot be expressed by an equation or by a single word. To evoke the feelings and associations, to be truly captivated, touched and engaged in a story, we use the full potential of embodiment. Using the entire body to tell a story unlocks the powerful multi-modality of our spatial and gestural abilities. There is a significant overlap of activation in our motor cortex for action words and their associated enaction by the reader [20], indicating that there is an implicit bodily engagement even when we read a single word. Further neuroscientific research suggests that the Broca's area which is linked to speech production, encodes neural representations of a spoken word in an articulatory code which is subsequently processed by parts of the motor cortex preceding the act of speech [11]. Reading a story aloud with the aid of iconic gestures allows us to tap into this tacit wiring of word to action [7].

The ancient Greek and Roman orators founded the school of *Chironomia*, the study of the effective use of hands to supplement or even replace speech. This school persisted until the 19th century with works such as [1] and [4]. There are practically no limits in how immersive a story can be when the storyteller's body – and those of an audience willing to play-along – is used to create a kind of performative theater. This immersiveness transcends the normal boundaries of interaction by causing a feedback loop[3] that influences how the story is enacted [10]. Ultimately, the perfect story involves the reader, such that readers can perfectly internalize what the storyteller has expressed, thus achieving every storyteller's goal.

So, when was the last time that a computer-generated story touched and inspired you in this way?

1.1 From embodied symbols to abstract ideas

The field of Computational Creativity aims to create machines that can transform humble ones and zeros into novel and original pieces of art, or into engaging tools that can foster creativity in humans. Creative storytelling is perhaps the most challenging endeavor of computational linguistic creativity. Looking at storytelling only by means of symbols and signs, we can derive abstract ideas with very different approaches such as *construal* [29], *anthropomorphism* [21] (e.g., see Fig. 1a) and *transformation* [48]. Ultimately, we strive for a system that can create and tell a touching story utilizing the expressive power of multi-modality and physical embodiment. The approach presented in this paper exploits a humanoid robot (the *NAO*) to augment symbolic narratives with embodied gesture and emotion.

Science Fiction gives us an understanding of what we can expect of a creative humanoid robot. In the HBO series *Westworld* (2016) we are presented with the perfect immersive android theater in which spectators are guests in a western-style amusement park. The android *hosts* (e.g., see Fig. 1d) are not aware of their programming or their role, which keeps them in storyline loops that they must repeat. These loops offer interaction points for guests to take part in their adventures. The host is thus the perfect actor; each shows human traits and offers the subjective impression of memory and emotion, yet each executes its role without the awareness that it is a performer. These android hosts literally embody their stories, as they are part of it while they tell it. The show also depicts the creators of the hosts, as developers who actively work on improving both the storylines (loops) and the gestures, expressions and vocal tics of the hosts. Eventually, it is the implementation of a profound class of gestures, the so-called *reveries*, that contributes to the evolution (and revolution) of the hosts. The captivating and immersive power of the hosts

[1] School of Computer Science, University College Dublin, Dublin D4, Ireland, email: philipp.wicke@ucdconnect.ie
[2] School of Computer Science, University College Dublin, Dublin D4, Ireland, email: tony.veale@ucd.ie

[3] The core of the so-called autopoietic feedback loop claims that every behavior of an actor triggers a specific behavior in a physically present spectator, and vice versa, thereby influencing how the actors behave.

is grounded in simple, concrete symbols, but evolves into a series of gestural manipulations that enable them to articulate the most abstract concepts in a perfectly convincing embodiment of an inner life.

Current robotic and CC technologies are still far away from these scenarios, but we can investigate how best to tell engaging stories using computers. *Scéalextric* [50] is one automated story-generation system that uses symbolic representations of characters, actions and causal consequences to invent and render stories with morals. At its core the system is built around action triplets such as the following:

1. X *action* Y
2. Y *reaction* X
3. X *re-reaction* Y

Scéalextric generates stories by linking these triples into a longer, track-like structure (or what in Westworld is called a *loop*) on which its characters (X and Y above) can move and interact. Stories are rendered upon this plot track by choosing fully-fleshed characters to inhabit the roles of X and Y, and by rendering action symbols as idiomatic surface sentences and dialogue fragments [52, 49]. Rendering is principally a linguistic activity, but it allows for multimodal expression too, as when Emoji are used to supplement (and even replace, translation-style) the textual renderings [51]. An example of an Emoji rendering is provided in Fig. 1b. The next step is to use a gestural rendering and transform these stories into an interactive, embodied storytelling experience using a humanoid robot.

The following section compares the textual and gestural approaches, showing that they share fundamental semiotic building blocks, and then proposes a marriage of both to augment symbolic narrative generation with gestures. The importance of gestures for language and for storytelling is explored in section 3, while section 4 focuses on the rendering of machine-generated stories on a humanoid robot with human-like gestures, starting with an overview of previous research in this field. Section 5 describes our proposed framework for allowing an interactive form of embodied storytelling with a Nao robot (see Figure 1c). The robot will engage with the spectator to shape the direction a story will take and the way it is told, to create a unified experience. The paper concludes with a consideration of the implications for future work.

2 FROM PICTURES TO BODIES

Emoji are not designed to be semantic primitives in the sense of [54, 15], but a previous study investigated their potential to be used as such in language [53], showing that it is useful to regard emoji as semiotic building blocks. The discipline's founder, Ferdinand de Saussure, viewed semiotics as the "*science that studies the life of signs within its society*"[41]. Just as we can identify the written word `SAPLING` as an arbitrary *signifier* of a *signified* concept, the mental image of a sapling, the emoji (Unicode U+1F331) as depicted in Fig. 1b, first emoji) serves as an iconic signifier for the same signification. Iconic signifiers give rise to their own forms of ambiguity, so that (Unicode U+1F331) can refer to the sapling itself, or the idea of growth, or to nature and plant-life in general [38]. Emoji can thus be used as metaphors, metonyms, icons and letters. [53] showed how symbolic narratives generated using the *Scéalextric* system can be augmented with emoji to render verbs as sequences of visual signs, Emoji can be used in this role as iconic signs for their literal meanings, as metaphors and as visual riddles using the rebus principle[4]. If

[4] This is an allusional method that uses pictures to represent words or parts of words. Consider the BEE emoji and the LEAF emoji, which can be read as BELIEVE if the rebus principle is applied.

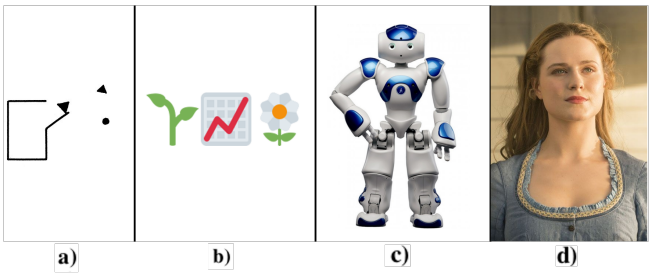

Figure 1. Examples of different renderings to tell a story from most abstract (a) to most distinct embodiment (d): **a)** Geometrical objects from the Heider and Simmel [21] experimental study of apparent behavior. Subjects were found to interpret the animation of these geometrical objects and shapes in terms of animated beings, attributing personality and motives. **b)** A sequence of emoji representing the concept of *growth* using a method derived in [53] to tell stories with emoji. The first emoji is the *sapling* emoji. **c)** The Nao robot by Aldebaran Robotics, for specifications see [16]. **d)** Evan Rachel Wood portrays the android host Dolores Abernathy in HBO's *Westworld* (2016)

Emoji can be used to co-render the output of the *Scéalextric* system, other semiotic units such as gestures can be thrown into the mix too.

To shift from the domain of pictures to the gestural domain of the body, we must identify gestures to embody the semiotic units of storytelling. As parts of a semiotic system [6] gestures can also can be classified as arbitrary, iconic and metaphorical [36]. In the next section, we consider why gestures are so important not only to storytelling, but to linguistic communication of all kinds.

3 GESTURES: EXPRESSING THE INTERNAL

Linguistics had long disregarded the role of the body in communication, but empirical work in cognitive science by McNeill [35], Bergen [2] and more recently Hauk [20] has shown that the body is an important instrument for human language and communication. An investigation titled *Embodied Sociolinguistics* by Bucholtz and Hall [3] claims that gestures are embedded in a cultural, social and ideological context and as such they imbue spoken language with a layer of contextual semantics. Kelly *et al.* [24] conduct an extensive investigation into the evolution of speech originating from the body. They see language development as a product of bodily actions, and note, from the perspective of language acquisition in children, that the onset of first gestures predicts the appearance of first words. Their evidence suggests that language should not be investigated separately from its origin, the body. As the interface between internal cognition and the external world, the body can make use of gestures to express what speech alone cannot convey. Gestures serve as a crucial link between the conceptualization of ideas and their expression through communication. McNeill describes them as fundamental assets of linguistics for our conceptualizing capacities [36]. It has to be noted, that the meaning of a gesture can be highly culturally and contextually dependent and their appropriateness can even differ within a small group of individuals. A distinct example is the Aymara language, where speakers refer to events in the future pointing behind them as opposed to pointing ahead of them as it is conventionally practiced in most other languages [42].

Despite technological progress in the videotaping and analysis of gestures and body language, there is still no unified methodology to

annotate and classify gestures [39]. Nonetheless, a range of studies, like those in gesture recognition [27, 26], consider Kendon's separation [25] of a *preparation*, *stroke* and *retraction* phase for the structure of a single gesture. For an overall classification, most studies refer to McNeill's classification of gestures into *iconic* (resembling what is being talked about), *metaphoric* (abstractly pictorial, but essential), *deictic* (i.e. pointing) and *beats* (temporal marks in narrative). As semiotic objects, the gestures understood as metaphoric act as a cross-domain mapping to express internal feelings, concepts and thoughts in concrete terms [5]. Gestures do not only speak for themselves, they serve as context for speech, while speech also serves as context for gestures when both are integrated successfully. This contribution of additional meaning to the communicative act has been empirically proven in a number of experiments [23, 7].

We can thus use gestures as emoji-like semiotic units for a broad variety of complex concepts, not least as part of an approach to embodied storytelling in a robotic agent. This framework, which admits text, emoji and gestures into the story-rendering process, will engage with users to create a captivating user experience.

4 OF MEN AND MACHINES

Robotic embodiment raises some prior issues we must address before considering gestural story-telling. Even if robots seem to have left the realm of pure science fiction, we are still at the point where an encounter with a robot in real life raises excitement, curiosity and amazement. But once robots become part of a system and we encounter them on a daily basis, habituation occurs [28]. On one hand, the enactment of a gesture by a robot might not appear as exciting if it is enacted by a human, but on the other hand this novelty effect will likely wear off after a few weeks. In a study by Kanda *et al.* [22], a robot was deployed in an 18-day field trial at a Japanese elementary school to teach children English using words and gestures. After the first week of frequent interaction with the robot, children showed diminished interest, to the point where one reported: "I feel pity for the robot because there are no other children playing with it".

Robots such as the Nao bring an undoubted cuteness factor to story-telling, yet we must strive to build systems that are creative and entertaining in their own right, in content as well as appearance. Despite advances in robotics, developers still struggle to create convincing humanoid robots, and all too often humanoid robots fall into the *uncanny valley* (Figure 2). This so-called valley [40] is a gulf separating a cartoon-like robot such as the Nao (Figure 1c), that is seen as cute and unthreatening, from an overly-human robot that is often thought to look creepy and disturbing in the Freudian sense of the *unheimlich*.

4.1 Previous Work

The Nao robot from Aldebaran/Softbank [16] is a polished, ready-to-use anthropomorphic bipedal robot that stands 57cm high. With LEDs for eyes and an immobile mouth, the robot also lacks facial expressiveness, yet it compensates with 25 degrees of freedom in its movements. A discussion of its different modalities and functions that are useful for interactive storytelling is provided in section 5.1. As an off-the-shelf consumer-grade robot, the Nao has been used less for research in robotic engineering and more for studies in psychology, sociology and linguistics [43]. Here we will highlight those studies which are relevant to interactive storytelling with a robot.

Most relevant is the approach of Pelachaud *et al.* who designed an expressive gesture model for a storytelling Nao robot [44, 30, 12].

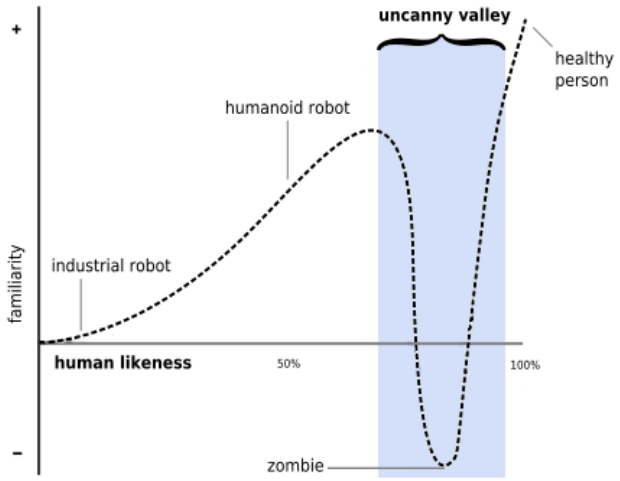

Figure 2. This graph depicts the theoretical perception of familiarity on a scale from industrial robot to healthy person. The area in blue marks the uncanny valley. Adapted from author: Karl MacDorman [33].

Their approach offers a unified framework that formalizes gestures previously used for a virtual avatar. As such, these gestures are rendered in a Function Markup Language and a Behavior Markup Language. This results in a reusable database of approx. 500 annotated gestures. They use a subset of these for a version of the robot that reads stories to children. Their evaluation in [31] confirms that the gestures are perceived as appropriate to their objectives while scoring poorly for naturalness. They highlight that their approach tries to blend this instantiation of storytelling with a common framework that also allows it to be applied for other robots. This is true for the gesture database, which has been annotated with admirable detail about the gesture space by dissecting each gesture into preparation, stroke and retraction phases. While an adaptation for the Nao robot requires two additional databases that were not available on request, we shall draw as many from insights from this study as we can.

Ham *et al.* [18] focused on the influence of gaze and gestural behaviour in a storytelling Nao robot. The authors handcrafted a set of 21 gestures and 8 gazing behaviors based on data from a professional stage actor. Their results indicate that the combined effect of gaze and gesture was greater than the effect of either gaze or gesture alone. Gazing is a standard procedure in the autonomous behavior software of the Nao robot, and we comment on the implications of this in section 5.1. While we learn from these insights, the approach in this paper must expand greatly on the set of 21 gestures to allow for a more exhaustive use of bodily modalities in the Nao.

With respect to multi-modal uses of the Nao, studies by Jokinen, Wilcock et al. [9, 37, 55] are worthy of mention. Their system, which is half question-answering system and half spoken-dialog system, uses Wikipedia as a knowledge source and renders the retrieved content in a conversational manner [55]. In [9] these authors discuss the different modalities of face detection, tactile sensors, non-verbal cues and gestures. They use the Nao's inbuilt face recognition software, as well as sonar sensors and speech direction detection to start the conversation, and empirically determine that the best communication distance is 0.9 meters. They implemented a small set of six gestures to signal discourse-level details, hyperlinks or to manage turn-taking

with human interlocutors. Some insights about speech and gesture synchronization are especially noteworthy. For example, their animation software did not accurately reflect the timing of gestures when performed by the actual robot. Each gesture was parametrized using Python code but the Nao's speech recognizer does not allow for a sudden interruption by the user. These authors also split each gesture into preparation, stroke and retraction phases to align the pitch of the spoken sentence with the stroke of the gesture.

The work of [19] investigates the influence of each separate modality in terms of its potential for emotional expression. This study investigated body movement, sound and eye color for six specific postures and emotions. It concludes that body movement appears to accurately convey an emotion in most cases, but sound and eye colour is much less expressively accurate, failing to match the desired emotion in half of all cases. These insights allow us to prioritize the gestures for our framework of embodied storytelling, which is described in the next section.

We begin by briefly reviewing the state of the art in automated storytelling. Although there are recent attempts to unify automatic storytelling frameworks (see e.g., [8]), most frameworks differ significantly in their algorithms and data-structures, using different knowledge bases, symbolic representations and/or learning technologies. The open story generation system *Scheherazade* [32] implements a novel approach that can work in new domains without possessing a prior model of those domains. *Scheherazade* first crowd-sources facts related to a new domain, automatically builds a domain model and finally selects a story from that domain model that obey's the system's high-level criteria. Another symbolic approach is the work of [45]: *MEXICA* automatically creates stories that conform to a cognitive model of the writing process. A case-based approach that reasons using an ontology of proven story elements is presented in [14], and more recent work on the functional morphology of stories is presented in [13]. In line with recent applications of deep Machine Learning techniques to almost every problem in Computer Science, Neural Networks have also been used of late as a basis for augmenting storytelling systems. Fine-grained approaches such as that of [46] use *Long Short-Term Memory* (LSTM) networks to infer events from a text that can later be used as part of a more general solution, while deep learning approaches such as that of [34] can draw from such event-level insights as they transform textual story data into narratives event sequences. As noted earlier, the work in this paper builds upon the *Scéalextric* system of [50] for a number of reasons, not the least of which is that the system comes with a comprehensive public knowledge-base of event sequences.

5 THE FRAMEWORK

5.1 Modalities

Our framework builds on two software packages provided by Aldebaran. The first, *Choregraphe*, provides a GUI that can be used to access most of the Nao's functionality. However, it does not provide direct access to the underlying code, and this access is crucial to the use of external databases and other sources of knowledge. We use it chiefly as a work-flow manager for the creation of gestures in the robot's *Animation Mode*. The second package is *NAOqi*, which supports access via direct coding in Python to all of the Nao's functionality, including joint motors, speakers and LEDs.

NAOqi (Version 2.1.4.13) comprises a range of modules, accessible via the robot's IP address. These modules, which must be loaded, have cross dependencies, so our framework provides a centralized *Awareness Loader* that pre-loads all modules for later use, while also booting the speech recognizer and initiating interaction with the user. This *Awareness Loader* is thus a centralized thread that executes a high-level function such as storytelling by calling only those modules necessary for the current action. In this way it sidesteps issues arising from cross-talk between modules. We focus here on the storytelling framework, which the user initiates by explicitly asking the Nao for a story. The trigger word that activates this feature via speech-recognition is '*story*'.

5.2 Technical Solutions

This section considers some technical problems encountered during the embodiment of the story-telling system, and describes technical solutions designed to circumvent the limits of each module.

5.2.1 Speech Recognition

This module can start and stop the Nao's speech recognition software, which responds to pre-assigned *trigger* words. There is no practical limit on the size of the trigger vocabulary, but even a few thousand words requires an onerous loading time and slows the system noticeably. Moreover, the likelihood of accurately recognizing any given word diminishes as the size of the vocabulary grows, since each trigger becomes less differentiated from others. In Nao's *word spotting* mode, the robot parses the incoming audio stream and assigns a probability to each segment that matches a trigger word in its vocabulary. This mode is most useful when users interact with the robot using complete sentences. We disable *word spotting* mode for interactive storytelling, as the system expects the user to reply with just one trigger word in an interaction. This offers more robustness and the vocabulary size can be increased since the algorithm does not need to extract the trigger from a context of unwanted speech. Yet even in this single-word mode it is crucial that the interaction still feels natural to the user. This naturalness is achieved by framing the interaction using yes-and-no questions. We empirically determine the threshold for trigger recognition to be $p(targetWord) > 0.6$.

5.2.2 Text-To-Speech

Nao offers a choice between a vanilla *Text-To-Speech* (TTS) module and an *AnimatedSpeech* module. The latter extends the TTS module with an enriched rendering of the speech output. Both modules employ the robot's speakers, while the latter responds to special markup in the given text. To create a more fluent interaction we preprocess each text string so as to access each embellishment prior its output. We also shorten the pause between sentences to create more fluency and momentum in the telling of a story.

5.2.3 Creaky Joints

It goes without saying that a storytelling robot requires speech output that is audible and understandable. However, the mechanical joints of a gesticulating robot create their own sounds that compete with the robot's speech, even when the volume of the latter is maximized. When additional noise in a non-laboratory environment is present, the story is easily misunderstood, thus defeating the use of gestures to make it more comprehensible. We have thus introduced a subtitle feature in our framework, which pipes the output of the TTS module onto a screen. As shown in in Fig. 4, the audience is thus able to read the robot's verbal output in large-print in real-time.

5.2.4 *Autonomous Behaviour and Eye Color*

The Nao platform provides a set of background procedures in its *Autonomous Behaviour* module that includes balancing, face recognition, face tracking, voice attention and blinking. Each of these contributes to a more lively appearance for the robot and so, unless it interferes with one of story-telling actions, the framework does not disable any autonomous behaviour. Notably, the blinking of the eyes interferes with changes to the LED color of the robot's eyes, but as we know from other research, its eye color does not contribute much to the comprehension of its outputs and is consequently disregarded.

5.3 Gestures

Previous works differ from the current approach in some significant respects, either because they used pre-generated stories, a small set of gestures, a pre-rendered set of speech and gesture behaviours, or no interaction at all during storytelling. The current framework overcomes all of these limitations by generating its stories in real time (via *Scéalextric*) during the robot's interactions with the user, and by drawing upon a set of 400+ gestures to render each sentence of the story with an appropriate embodied behaviour.

We extracted 423 pre-installed gestures (also called behaviours) from the robot's internal storage and associated each of these gestures with plot *verbs* from the *Scéalextric* system. 13 of the 423 pre-installed gestures were discarded because they pose an increased risk of falls and of harming the robot via poor movement trajectories, or because they are too specific (e.g. singing a song) for any action verb, or because they loop endlessly. For the remaining 410 gestures we create strong, medium and weak associations to one or more *Scéalextric* verbs. 195 of the 410 have at least one strong association, 322 have at least one medium association and 214 have at least one weak association. This results in a coverage of 68% for all action verbs in the *Scéalextric* system. Because *Scéalextric* searches a graph of interconnected action triples to form a story, we can easily favor stories that use actions with associated gestures, or rank stories by the degree to which they can be effectively embodied by the robot. For an example gesture see Fig. 3.

To foster a natural and captivating interaction during storytelling, we must synchronize the robot's gestures with its speech while also inserting interaction points for the audience. Several authors have studied the selection of suitable time points for speech and gesture synchronization. A notable ERP study by [17] concludes from empirical evidence that speech and gesture are most efficiently integrated when they are coordinated together in time. The majority of studies conclude that the integration of information works best if the gesture co-occurs with its contextualizing word. The approach of [9] uses a very small set of decomposable gestures so as to synchronize each phase of the gesture with the words of predefined sentences. As we use a large number of atomic gestures, our current framework employs a simple heuristic that synchronizes the start of each gesture with the start of the sentence it adorns. In [36] McNeill argues that one gesture mostly appears with one clause and only occasionally more than one appears with a single clause. In the current framework most of the gestures temporally align with one clause, and in cases where their duration is longer than the sentence, the robot waits for up to 2 seconds before starting any new sentence and gesture.

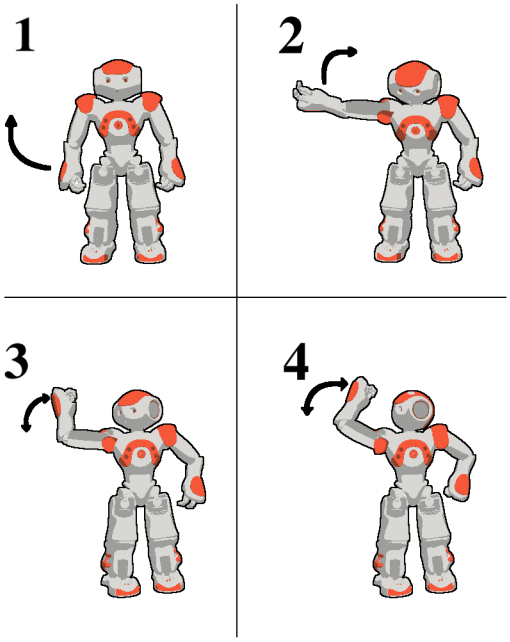

Figure 3. Example of a Nao gesture in four frames. This gesture has been annotated to strongly associate with the action *train*. First frame is the resting position, followed by a raising of the arm in the second frame. The third and fourth frame are alternating a few times. This gesture is a *show of muscles*.

6 TELLING AN INTERACTIVE STORY

The framework as described – marrying the *Scéalextric* story-generator to a semiotic system of robotic gestures – has been implemented around the Nao platform. In this pilot implementation, users interact with the robot using single-word prompts, such as "story", "yes" and "no." The first initiates the story-telling process, while the latter two offer guidance via answers to the robot's questions. In addition, a user may specify any of 782 verbs in response to the robot's initial request for a story action on which to start a new story. For instance, should the user say "betray" then the robot will respond with a story about betrayal by generating a *Scéalextric* story from a starting triple that contains this verb. The stories it generates are rendered into idiomatic English and articulated by the robot's speech synthesis module, while one gesture per sentence (typically the one most strongly associated with the main verb) is simultaneously mimed.

In cases where there is no gesture associated with the verb, the system instead draws from a pool of 16 generic poses and gestures that are not obviously associated with any one action. Fig. 4 presents a scene from a public demonstration of this pilot system. We can now elaborate on the subsequent work that will transform this set-up into a fully interactive experience for the audience.

A captivating story allows readers to weave their own personalities into the tale and empathize with its characters. This kind of interaction requires the robot to request guidance from the user that will shape the story. Fortunately, the knowledge-base provided with *Scéalextric* provides a question form for each of its plot verbs. For example, the action *kill* has the question form '*Have you ever wanted to put an end to someone?*' Suppose then that just one of the possible next actions in a story is *kill*. Instead of choosing for itself, or choosing randomly, the robot can instead pose the associated question to the user. If the answer is "yes" then this is taken as tacit acceptance that the next action in the story will be *kill*. If it is "no" then the robot considers another avenue for the plot to follow.

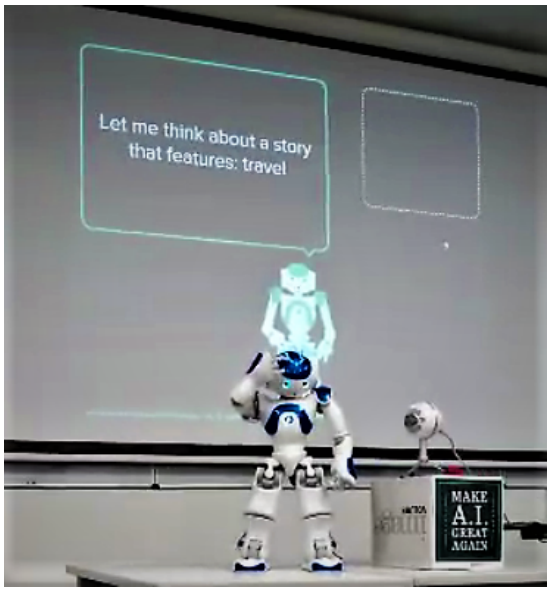

Figure 4. Demonstration of the preliminary storytelling framework at the UCD School of Computer Science Opening Evening 2017.

6.1 Digging for Stories

In this way the robot probes the psyche of the user to find material for its plot lines. The story generation process can be regarded as a tree (see Fig. 5) in which the root is an initial action that has been provided by the user. Each child node holds an action that causally follows from its parent node, while the tree's leaves are the ultimate actions in each possible plot originating at the root. At each node the user is again probed with a question related to the node's action. A "yes" plunges the teller deeper into the story-tree, while a "no" pushes the teller to another node on the same level.

In the following dialogue, which can serve as an illustration, the associated action in each case is appended in brackets and is not actually shown to the audience.

1. `Nao`: Have you ever been rebuffed by an elitist?
   (are_rebuffed_by)
2. `User`: No.
3. `Nao`: Have you ever shared a kiss with a lover?
   (are_kissed_by)
4. `User`: Yes.
5. `Nao`: Have you ever offered protection to somebody?
   (guard)
6. `User`: No.
7. `Nao`: Have you ever worked your charms on an admirer?
   (charm)
8. `User`: No.
9. `Nao`: Have you ever had a debate with a rival?
   (debate)
10. `User`: Yes.

Here the system initiates the dialogue with a random action, and poses the related question in (1). When the user replies in the negative in (2), the system draws another random action and poses the related question in (3). When the user responds positively in (4), the system can now choose a plausible causal reaction in (5). The path picked through the tree by the user's "yes" responses serves as the plot for the robot's story, which it can finally render in idiomatic English and articulate with speech and gestures. This rendering is performed when the user eventually tells the robot to "enact" the tale. In the rendered tale, the protagonist is designated "you" since that character's actions mirror the answers given by the user.

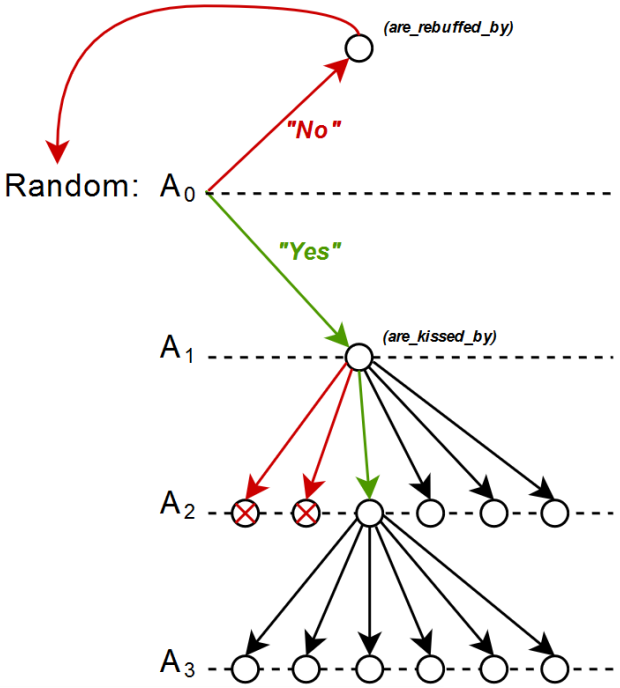

Figure 5. Example of the knowledge acquisition process in a tree diagram. Red arrows indicate a negative response from the user and green a positive. Black arrows have not been evaluated.

6.2 Enactment

An example of a story enacted in this way is provided in the following trace:

1. `[BodyTalk_9, None, kill]:`
   This is the story of how you killed John.
2. `[Kisses_1, Strong, kiss]:`
   You gave John a passionate kiss.
3. `[No_1, Medium, are_rejected_by ]:`
   But John rejected your proposition.
4. `[Explain_3, Strong, debate]:`
   So you debated hard and long with John.
5. `[No_3, Medium, lose_favour_with]:`
   John no longer felt well-disposed towards you.
6. `[BodyTalk_9, None, kill]:`
   As a result you chocked the air out of John.

This is a simple story by *Scéalextric* standards, but it serves to illustrate the rendering process. We believe a user can better relate to a story that is shaped by personal insights provided by that user to the robot, yet it is important to note that the user does not actually write the story. The user is at best a co-creator, or perhaps a muse. It is the machine that writes its own tales.

7 FUTURE WORK

In this paper we have considered the role of gesture in communicating the actions of the story, under the presumption that an action is the same regardless of who performs it. However, when humans creatively use gestures to tell stories, they often inflect those gestures to reflect the character performing them. We have said little about the role of character in story-telling here, though much has been said in [51] in the context of *Scéalextric* and its means of generation. In fact, *Scéalextric* employs a rich database of stock characters and their qualities (behaviour, dress sense, pros and cons), to model hundreds of people who are historical, contemporary and entirely fictional. Since *Scéalextric* stories employ vivid characters as protagonists and antagonists, we shall have to explore how this vividness can translate into gestural inflections.

8 CONCLUSION

Our framework synthesizes some elements of previous approaches to embodied storytelling in a robotic agent while innovating in other respects. Even when interactions with the user are limited to a very small set of answers (such as '*Yes*', '*No*', '*Enact*', '*Repeat*', '*Stop*') complex questions can be used to tease out a uniquely tailored story that is based on the user's own experiences. However, these stories also invite users to reflect on their own actions in a fictional context. We have taken a step away from previous research that used the presentation of the story as a means to analyze the quality of human-robot-interaction, and a step closer to an embodied collaborative system that puts the focus on the interaction between humans and robots.

A robot might create stories that seem less plausible to the user if no guidance is provided, because a robot that does not understand the meanings of the symbols it is manipulating cannot be regarded as possessing intelligence, not to mention creativity [47]. In our framework the user's input is a means of personalization, not of assuming creative control. In this way both the robot *and* the human benefit from their interactions, as do the stories that result. Though still simple, these tales do a little of what great tales do so well: they put readers at the heart of the action while making readers question their own hearts.

ACKNOWLEDGEMENTS

We would like to thank Stefan Riegl for his contribution to the storytelling system as outlined here.

REFERENCES

[1] G. Austin, *Chironomia; or, a treatise on rhetorical delivery*, T. Cadell and W. Davies, 1806.
[2] B. Bergen, S. Narayan, and J. Feldman, 'Embodied verbal semantics: Evidence from an image-verb matching task', in *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society*, pp. 139–144, (2003).
[3] M. Bucholtz and K. Hall, 'Embodied sociolinguistics', *Sociolinguistics: Theoretical debates*, (2016).
[4] J. Bulwer, *Chirologia, or the natural language of the hand.*, Thom. Harper and Henry Twyford, 1975.
[5] A. Cienki and C. Müller, 'Metaphor, gesture, and thought', *The Cambridge handbook of metaphor and thought*, 483–501, (2008).
[6] H. H. Clark, *Using language*, Cambridge university press, 1996.
[7] N. Cocks, G. Morgan, and S. Kita, 'Iconic gesture and speech integration in younger and older adults', *Gesture*, **11**(1), 24–39, (2011).
[8] E. Concepción, P. Gervás, and G. Méndez, 'A common model for representing stories in automatic storytelling', in *6th International Workshop on Computational Creativity, Concept Invention, and General Intelligence. C3GI*, (2017).
[9] A. Csapo, E. Gilmartin, J. Grizou, J.G. Han, R. Meena, D. Anastasiou, K. Jokinen, and G. Wilcock, 'Multimodal conversational interaction with a humanoid robot', in *Cognitive Infocommunications (CogInfoCom), 2012 IEEE 3rd International Conference on*, pp. 667–672. IEEE, (2012).
[10] E. Fischer-Lichte, *Ästhetik des Performativen*, Suhrkamp Verlag, 2012.
[11] A. Flinker, A. Korzeniewska, A.Y. Shestyuk, P.J. Franaszczuk, N.F. Dronkers, R.T. Knight, and N.E. Crone, 'Redefining the role of broca's area in speech', *Proceedings of the National Academy of Sciences*, **112**(9), 2871–2875, (2015).
[12] R. Gelin, C. d'Alessandro, Q. A. Le, O. Deroo, D. Doukhan, J.C. Martin, C. Pelachaud, A. Rilliard, and S. Rosset, 'Towards a storytelling humanoid robot.', in *AAAI Fall Symposium: Dialog with Robots*, (2010).
[13] P. Gervás, 'Computational drafting of plot structures for russian folk tales', *Cognitive computation*, **8**(2), 187–203, (2016).
[14] P. Gervás, B. Díaz-Agudo, F. Peinado, and R. Hervás, 'Story plot generation based on cbr', *Knowledge-Based Systems*, **18**(4), 235–242, (2005).
[15] C. Goddard and A. Wierzbicka, *Semantic and lexical universals: Theory and empirical findings*, volume 25, John Benjamins Publishing, 1994.
[16] D. Gouaillier, V. Hugel, P. Blazevic, C. Kilner, J. Monceaux, P. Lafourcade, B. Marnier, J. Serre, and B. Maisonnier, 'Mechatronic design of nao humanoid', in *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pp. 769–774. IEEE, (2009).
[17] B. Habets, S. Kita, Z. Shao, A. Özyurek, and P. Hagoort, 'The role of synchrony and ambiguity in speech–gesture integration during comprehension', *Journal of Cognitive Neuroscience*, **23**(8), 1845–1854, (2011).
[18] J. Ham, R. Bokhorst, R. Cuijpers, D. van der Pol, and J.J. Cabibihan, 'Making robots persuasive: the influence of combining persuasive strategies (gazing and gestures) by a storytelling robot on its persuasive power', in *International conference on social robotics*, pp. 71–83. Springer, (2011).
[19] M. Häring, N. Bee, and E. André, 'Creation and evaluation of emotion expression with body movement, sound and eye color for humanoid robots', in *Ro-Man, 2011 IEEE*, pp. 204–209. IEEE, (2011).
[20] O. Hauk, I. Johnsrude, and F. Pulvermüller, 'Somatotopic representation of action words in human motor and premotor cortex', *Neuron*, **41**(2), 301–307, (2004).
[21] F. Heider and M. Simmel, 'An experimental study of apparent behavior', *The American journal of psychology*, **57**(2), 243–259, (1944).
[22] T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro, 'Interactive robots as social partners and peer tutors for children: A field trial', *Human-computer interaction*, **19**(1), 61–84, (2004).
[23] S. D Kelly, D. J. Barr, R. B. Church, and K. Lynch, 'Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory', *Journal of memory and Language*, **40**(4), 577–592, (1999).
[24] S. D Kelly, J. M. Iverson, J. Terranova, J. Niego, M. Hopkins, and L. Goldsmith, 'Putting language back in the body: Speech and gesture on three time frames', *Developmental neuropsychology*, **22**(1), 323–349, (2002).
[25] A. Kendon, 'Gesticulation and speech: Two aspects of the process of utterance', *The relationship of verbal and nonverbal communication*, **25**(1980), 207–227, (1980).
[26] S. Kettebekov and R. Sharma, 'Toward natural gesture/speech control of a large display', *Engineering for human-computer interaction*, 221–234, (2001).
[27] S. Kettebekov, M. Yeasin, and R. Sharma, 'Prosody based audiovisual coanalysis for coverbal gesture recognition', *IEEE transactions on multimedia*, **7**(2), 234–242, (2005).
[28] K. L. Koay, D. S. Syrdal, M. L. Walters, and K. Dautenhahn, 'Living with robots: Investigating the habituation effect in participants' preferences during a longitudinal human-robot interaction study', in *Robot and Human interactive Communication, 2007. RO-MAN 2007. The 16th IEEE International Symposium on*, pp. 564–569. IEEE, (2007).
[29] R.W. Langacker, 'Nouns and verbs', *Language*, 53–94, (1987).
[30] Q. A. Le, S. Hanoune, and C. Pelachaud, 'Design and implementation of an expressive gesture model for a humanoid robot', in *Humanoid*

Robots (Humanoids), 2011 11th IEEE-RAS International Conference on, pp. 134–140. IEEE, (2011).
[31] Q. A. Le and C. Pelachaud, 'Evaluating an expressive gesture model for a humanoid robot: Experimental results', in *Submitted to 8th ACM/IEEE International Conference on Human-Robot Interaction*, (2012).
[32] B. Li, S. Lee-Urban, G. Johnston, and M. Riedl, 'Story generation with crowdsourced plot graphs.', in *AAAI*, (2013).
[33] K. F. MacDorman, T. Minato, M. Shimada, S. Itakura, S. Cowley, and H. Ishiguro, 'Assessing human likeness by eye contact in an android testbed', in *Proceedings of the XXVII annual meeting of the cognitive science society*, pp. 21–23, (2005).
[34] L. J Martin, P. Ammanabrolu, W. Hancock, S. Singh, B. Harrison, and M. O. Riedl, 'Event representations for automated story generation with deep neural nets', *arXiv preprint arXiv:1706.01331*, (2017).
[35] D. McNeill, 'So you think gestures are nonverbal?', *Psychological review*, **92**(3), 350, (1985).
[36] D. McNeill, *Hand and mind. What the hands reveal about thought*, Chicago: University of Chicago Press, 1992.
[37] R. Meena, K. Jokinen, and G. Wilcock, 'Integration of gestures and speech in human-robot interaction', in *Cognitive Infocommunications (CogInfoCom), 2012 IEEE 3rd International Conference on*, pp. 673–678. IEEE, (2012).
[38] I. Meir and O. Tkachman, *Iconicity*, Oxford University Press, 2014.
[39] I. Mittelberg, 'Methodology for multimodality', *MARQUEZ, MG; MITTELBERG, I. Methods in cognitive linguistics. Amsterdam: John Benjamins Publishing Company*, 225–248, (2007).
[40] M. Mori, 'The uncanny valley', *Energy*, **7**(4), 33–35, (1970).
[41] W. Nöth, *Handbook of semiotics*, Indiana University Press, 1995.
[42] R.E. Núñez and E. Sweetser, 'With the future behind them: Convergent evidence from aymara language and gesture in the crosslinguistic comparison of spatial construals of time', *Cognitive science*, **30**(3), 401–450, (2006).
[43] N. Parde, A. Hair, M. Papakostas, K. Tsiakas, M. Dagioglou, V. Karkaletsis, and R. D. Nielsen, 'Grounding the meaning of words through vision and interactive gameplay.', in *IJCAI*, pp. 1895–1901, (2015).
[44] C. Pelachaud, R. Gelin, J.C. Martin, and Q. A. Le, 'Expressive gestures displayed by a humanoid robot during a storytelling application', *New Frontiers in Human-Robot Interaction (AISB), Leicester, GB*, (2010).
[45] R. Pérez ý Pérez and M. Sharples, 'Mexica: A computer model of a cognitive account of creative writing', *Journal of Experimental & Theoretical Artificial Intelligence*, **13**(2), 119–139, (2001).
[46] K. Pichotta and R. J. Mooney, 'Learning statistical scripts with lstm recurrent neural networks.', in *AAAI*, pp. 2800–2806, (2016).
[47] J. R. Searle, 'Minds, brains, and programs', *Behavioral and brain sciences*, **3**(3), 417–424, (1980).
[48] T. Veale, *Coming good and breaking bad: Generating transformative character arcs for use in compelling stories*, Proceedings of ICCC-2014, the 5th International Conference on Computational Creativity, Ljubljana, June 2014, 2014.
[49] T. Veale, 'Game of tropes: Exploring the placebo effect in computational creativity.', in *ICCC*, pp. 78–85, (2015).
[50] T. Veale, 'A rap on the knuckles and a twist in the tale', *AAAI spring symposium series*, (2016).
[51] T. Veale, 'Déjà vu all over again', in *ICCC*, pp. 245–252, (2017).
[52] T. Veale and A. Valitutti, 'A world with or without you', in *Proceedings of AAAI-2014 Fall Symposium Series on Modeling Changing Perspectives: Re-conceptualizing Sensorimotor Experiences. Arlington, VA*, (2014).
[53] P. Wicke. Ideograms as semantic primes: Emoji in computational linguistic creativity. Thesis DOI: 10.13140/RG.2.2.21344.89609 (2017).
[54] A. Wierzbicka, *Semantic primitives*, (Frankfurt/M.)Athenäum-Verl., 1972.
[55] G. Wilcock and K. Jokinen, 'Wikitalk human-robot interactions', in *Proceedings of the 15th ACM on International conference on multimodal interaction*, pp. 73–74. ACM, (2013).