Show, Don't (Just) Tell: Embodiment and Spatial Metaphor in Computational Story-Telling

Philipp Wicke
School of Computer Science
University College Dublin
Dublin, Ireland
philipp.wicke@ucdconnect.ie

Tony Veale
School of Computer Science
University College Dublin
Dublin, Ireland
tony.veale@ucd.ie

Abstract

To a human storyteller, a story is more than a textual artifact. Rather, as stories are both generated *and* generative, each is also a blueprint for performances to come. Tellers must draw on their own bodily affordances – from voice and gesture to movement around a stage – to bring stories to life, much as a conductor and an orchestra must translate a written score into actual music. This paper explores the creative challenge of translating from a story-*text* to a story-*performance*, from words to physical actions and characters to embodied actors. The mapping requires distinct models for gesture, narration, dialogue and stage direction if computer-generated tales are to transcend the limitations of their production process. Using the *Scéalability* framework, we evaluate the interlocking role of spatial metaphor and pantomime in turning a narrative artifact into a coherent performance.

All The World's A Stage

Cinema is a visual medium, so filmmakers understandably live by the old Hollywood maxim, *"Show, Don't Tell."* Why use exposition or dialogue to tell of a dramatic event when you can show it directly on screen? Whether on stage or on screen, actors don't just speak their lines; they live them out, with gesture, posture and meaningful spatial movement. The same can be said of storytellers of any kind. We humans put our backs into telling a story, so that our audiences can experience narrative events as though they were really there. Tellers are also performers, and their embodiment is key to the audience's identification with the characters in a tale.

Mapping from a narrative text – of the kind typically produced by story generation systems – to an embodied performance requires a process not unlike the translation of a musical score into an orchestral performance. The composer's vision for the piece must be respected, but difficult decisions that affect its realization must also be made. How many performers are needed, who will play what, and how must they be arranged, in space or in time? All the tricks of the theatre must be used to get the most out of what is available. When a piece has more roles than the cast has performers, the mapping is not an isomorphic one. As some roles are prioritized (or focalized) over overs, embodied characters must allude to the actions or feelings of those we cannot directly see or hear. Physical actors must show *and* tell, to convey a story world that is much bigger than them and their stage.

Figure 1: Two embodied, robotic agents telling a story. One enacts a marriage proposal by iconically bending its knee.

We focus here on the embodied realization of computer-generated stories with a mix of physical devices, specifically Amazon's *Echo/Alexa* and the *NAO* anthropomorphic robot. Alexa serves as our omniscient narrator, while two NAOs, named *Kim* and *Bap*, speak and act as the story's characters. Unlike past approaches to robotic storytelling, our robots are actors who take on specific roles in the narrative, and so they must act accordingly. Each robot speaks with a gender-appropriate voice for their character, making apt and often iconic gestures as they do, while moving about the stage in ways that metaphorically convey their relationship to other characters. In addition to speaking any dialogue given in the story itself, they must often add their own so that each physical action is paired with a naturalistic speech-act. They use this supplementary dialogue to comment on the tale itself, or to compensate for the physical absence of lesser characters.

In some ways, compensation is key to the whole process. Embodiment compensates for the weaknesses of the underlying story, by drawing our attention to how it is physically enacted rather than how it was automatically created. It may seem that clunky robots can do little to salvage a clunky tale, but each one adds to the innate comedy potential of the other. Consider *Sunspring*, an AI-generated sci-fi script that was filmed as a short movie by Ross Goodwin and Oscar Sharp. On the page, *Sunspring* is by turns absurd and seemingly unfilmable, with stage directions such as "He is standing in the stars and sitting on the floor." Yet the human actors lean into

Proceedings of the 11th International Conference on Computational Creativity (ICCC'20)
ISBN: 978-989-54160-2-8

the script, using their faces, gestures and body-language to make the incongruous seem relatable on a human level. That audiences find the experience affecting and memorable has more to do with how it is embodied than how it was written. Our performance framework, named *Scéalability*, makes use of the *Scéalextric* story-generator (Veale 2017) for its tales, but it can, in principle, use any automated generator that provides access to a tale's surface form and its deep structure. As in the case of *Sunspring*, our goal is to add coherence and entertainment value to computational stories by appealing to how we humans use gesture and space to create meaning.

This goal is unpacked in the following sections. We first explore research in automated story generation, and on how robots and other devices can be used to tell stories. We then present the *Scéalability* framework, with a specific emphasis on its models of dialogue, gesture and spatial metaphor. As our concerns go beyond system-engineering, *Scéalability* is also used to explore some hypotheses regarding the relative merits of iconic/pantomimic gesture and spatial metaphor in an embodied performance. Empirical studies are conducted to determine if audiences really can discern and appreciate the coherent use of space and gesture in the telling of a tale.

Related Work and Ideas

Computational Storytelling

The generation of stories by mechanical means is a practice that predates AI and the advent of the modern computer. In 1928, the Canadian author William Wallace Cook marketed a system named *Plotto, the master book of all plots*, which gave budding authors more than 1600 plot schemas, cross-indexed for easy retrieval and recombination (Cook 1928). Cook's system was just one of several that exploited nothing more sophisticated than the filing cabinet and the card index, placing the emphasis squarely on good data over algorithms. Yet, in retrospect, *Plotto* anticipates the AI approaches that would follow, from case-based-reasoning to story grammars.

Early AI approaches would be just as schematic as *Plotto*, while automating the tasks of planning and schema combination. One of the first, the *Novel Writer* system of (Klein et al. 1973), generated murder-mystery plots, while the more influential *TALE-SPIN* produced tales of woodland creatures by first building a world of goals and related characters for them to explore (Meehan 1977). Genre is also a tacit *meta*-schema in its own right, one that lends the weight of convention to otherwise insubstantial tales. So, just as the *Universe* system cranked out soap opera plots (Lebowitz 1985), the *Minstrel* system navigated a very different genre with just as many conventions: tales of courtly knights (Turner 1993).

The creative practices of human authors offer insights into how machines can write (or aspire to) tales of comparable quality. Dehn's *Author* was the first AI system to explicitly model authorial goals in story creation (Dehn 1981), though more elaborate cognitive models have since been developed. The *Mexica* system of (Pérez and Sharples 2001) posits a two-phase cycle of creation called E-R, for *Engagement-Reflection*, in which the story generator alternates between bouts of incremental story development and subsequent consideration of the new opportunities that these may open up.

Cook's development of *Plotto* in 1928 coincided with the flourishing of academic interest in the structuralist analysis of folk tales and similar cultural artifacts. In his *Morphology of the Folk Tale*, (Propp 1928) identified an inventory of recurring building blocks from which old tales are built, and from which new ones might be composed. His analysis remains relevant today, and forms the schematic basis for such generators as the *PropperWryter* system of (Gervás 2013).

These models are each symbolic in nature, and use logical forms that would pose little difficulty to the users of Cook's *Plotto* book. In contrast, non-symbolic approaches sacrifice this interpretability for trainability, robustness and scale. For instance, the GPT-2 neural language model of (Radford et al. 2019) is trained on 40Gb of web text, and can "hallucinate" coherent continuations to arbitrary story text prompts. For instance, these authors show how continuations preserve the genre and guiding conceit of even speculative prompts, such as one that imagines the discovery of unicorns in the Andes. But GPT-2 and its kind are *text-in, text-out* generators that do not provide access to the plot-structure of the narrative text. While GPT-2 has hidden depths, its outputs are all surface.

Relation to the previous research project *Scéalability* needs access to the surface *and* deep structure of a story, so that it can choose gestures, dialogue and spatial movements to match the narrative intent behind the words. For this reason, we opt instead for the *Scéalextric* generator of (Veale 2017). Inspired by Cook's *Plotto*, *Scéalextric* constructs its narratives from prefabricated segments of plot, which it connects end-to-end or expands top-down using recursive descent. Each plot segment comprises a sequence of transitive story verbs, from an inventory of 800 possibilities including *fall_in_love_with*, *rescue* and *murder*. Each verb has two pariticipants, either the generic A and B or functionally-dependent roles such as A_{spouse} and B_{friend}. At heart, *Scéalextric* is a story-grammar for generating plots that are then rendered into a narrative text using an idiomatic mapping of story verbs to phrasal forms. It is its scale, ease of extensibility and modularity that distinguishes it most from other story generators. For instance, it provides a large database of famous characters for use in its stories, to instantiate the generic roles A and B (and their dependents) and to lend vivid colour (specific locales, weapons, vehicles, clothing, etc.) to the rendered text. Additional modules can also be inserted with relative ease, to attach physical gestures, spoken dialogue, and other stage directions so as to turn a narrative into a performance that shows as well as tells.

Embodied Storytelling

Story-telling with performing robots has been studied from a number of perspectives using robots of different kinds and varying physical affordances. The storytelling capabilities of an expressive robot face, called *Reeti*, was investigated by (Striepe and Lugrin 2017). In their comparative study, test subjects were presented with stories delivered by the *Reeti*, an audio book, and a neutral robot speaker. These authors use the same *AttrakDiff* questionnaire (Hassenzahl, Burmester, and Koller 2003) for their study as we shall employ in our own evaluation to follow. While the *Reeti* lacks a

body, it has an emotive face that is expressive enough to convey even an ironic intent (Ritschel et al. 2019). We shall use *NAO* robots that lack facial expression, with the expectation that their spatial mobility will compensate in other ways.

Storytelling is more than reading a text, as stories are enriched by the most basic visual cues. Early studies by (Heider and Simmel 1944) show that humans readily imbue simple geometric shapes that move about a screen with human-like intentions and mental states. Audiences are just as willing to attribute intention and emotion to anthropomorphic robots that purposely gesticulate and stride about a stage.

Studies show that apt gestures can increase the expressiveness of an embodied storyteller (Csapo et al. 2012). Yet most robotic storytellers draw upon a small set of predefined gestures. For instance, the *WikiTalk* project of (Meena, Jokinen, and Wilcock 2012) relies on just seven gestures, which it uses to indicate discourse structure, as in the *Open-Hand-Palm-Up* gesture to mark the start of a new paragraph. *WikiTalk* uses a NAO robot to present the results of Wikipedia queries with a mix of voice and gesture, and shows how this integrated multimodality supports a natural interaction between human and machine (Jokinen and Wilcock 2014). But there have also been attempts to create custom gestures in order to suit an arbitrary speech context. Recent work by (Rodriguez et al. 2019) uses a Generative Adversarial Network (GAN) to produce apt gestures for a *Pepper* robot. Human gestures do not follow clear universal rules, so the generation of gestures is a non-trivial task. Nonetheless, there is some cognitive evidence for a schematic basis to many human gestures (Cienki 2005; Mittelberg 2018), and the generation of non-verbal behaviors for a virtual avatar based on such schemas is presented in (Ravenet, Clavel, and Pelachaud 2018).

Embodied Performance

Proxemics is the study of space for social interactions and their actors in a mutual environment. Research by Pope *et al.* investigated those interactions by theatre practitioners in physical spaces, using 360-degree filming and virtual reality (VR) (Pope et al. 2017).

Theatrical performances of robots on a stage require technological developments and a robust software framework. A focus on those low-level challenges, which require strong technical coordination has been presented in (Lin et al. 2009). In their study, a twin-wheeled, two-armed robot and a bipedal robot were set in a theatrical performance to show different performative challenges, e.g. story-telling concluding with a kiss between two robots.

A co-creative approach with a human and artificial performers has been investigated as improvisational theatre by (Mathewson and Mirowski 2017). The authors present two versions of AI-based chat-bots. *Pyggy* is an "Artificial Improvisor" using speech synthesis and speech recognition to communicate with an audience. It is capable of an open dialog interaction and *Pyggy* is embodied by a virtual avatar (a face with mouth movements synchronized to the speech). A different version of the robot called *A.L.Ex.* utilized Neural Language Model-based Text Generation to overcome *Pyggy*'s limited set of trained sentences.

In our research, we bring additional meaning from motion between our robotic actors, as well as orientation, pantomime and gesture in a study of space, taking other research, e.g. (Pope et al. 2017) into a computational domain. Our focus is not on these low-level challenges (Lin et al. 2009), but on simpler and more general uses of space and gesture in story-telling with improvisational elements between the robots.

Relation to the previous research project Embodied storytelling within a generative framework has previously been studied in (Wicke and Veale 2018b; 2018a; Veale, Wicke, and Mildner 2019), who combined the *Scéalextric* story-generator with a single anthropomorphic NAO robot. Over 400 predefined gestures of the NAO are mapped onto almost 800 story verbs of the plot-generation system, so that a single robot teller makes an apt movement for every action it narrates. An interactive variant was later presented in (Wicke and Veale 2018a), in which the robot is guided through the story-space by a user's answers to the robot's questions. In a process that might be considered *co-creative*, the robot uses the *Scéalextric* plot graph to ask its questions and then branch according to the answer it receives. Once all questions have been answered, the plot is assembled and the tale is performed. A second device – the smart speaker *Alexa/Echo* – is added to the mix in (Veale, Wicke, and Mildner 2019). This pairing allows for banter between the devices, who now share the responsibility of narrating the tale (Alexa) and responding to it physically (NAO). It allows for a comparative analysis of the two devices, contrasting the disembodied voice of *Alexa* with the embodied antics of *NAO*. Although both contribute to the performance, it is always the embodied robot that adds the strongest humorous dimension. For this reason, *Scéalability* doubles down on its use of a robot by orchestrating the actions of two *NAOs* in showing and telling a tale.

The novel contribution of this research over the closely related research project is the addition of spatial movements by the robotic actors, which is baked into the storytelling system *Scéalextric* and coordinated by the *Scéalability* framework. Those changes and additions will be outlined in the next section. Moreover, an empirical evaluation presents its successful integration.

The *Scéalability* Embodiment Framework

Scéalability is conceived as an approach to storytelling-as-performance that places a definite emphasis on computer-generated narrative. This emphasis has a practical rationale: human-authored stories lack the semantic markup to allow performers to look beyond a surface text to see the plot logic within, while symbolic story generators offer transparent access to any level of the story at which the machine can reason. So, at the core of *Scéalability* sits a transparent generator of just this kind. Specifically, the structured outputs of the *Scéalextric* generator comprise a sequence of multilevel *story beats*, each of which contains the following elements:

1. A single narrative event, framed by a single story verb
2. Generic case roles for this story verb (e.g., A and B_{spouse})
3. The specific characters that fill these case roles (e.g. *Neil Armstrong* and *Princess Leia*, or *C-3PO* and *HAL 9000*)
4. A surface-level textual rendering for this single event, that incorporates character-specific details where possible
5. A logical connective that links this story beat to the next one (e.g., *so, then, but, yet, because, and*)

These strands are easily unpicked by *Scéalability* so as to weave its own performative elements into a narrative. While any system that facilitates this separability of levels can be used as the generative heart of the framework, we evaluate the approach here with *Scéalextric*. Its various symbolic levels allow us to strengthen its existing narrative model while hooking in additional gesture, dialogue and spatial models. Let's now look at each of those additional models in turn.

The Gestural Model

As shown in (Wicke and Veale 2018b; 2018a; Veale, Wicke, and Mildner 2019), storytelling can be gesturally-enhanced by mapping (almost) every story verb onto one or more gestures in the robot's repertoire of physical flourishes. However, in those earlier systems, a single robot was expected to gesticulate for *all* of the characters in a story, with no regard for which character was making which gesture. With two robots, one for each of the two central parts A and B, gestures must be linked to specific case roles in each story verb so that they are performed by the right robot, and at the right time relative to the actions of other performers.

For a performance using n robots, we assume that only n characters will be embodied on stage. Human theatrical performances are more flexible than this, as real actors can play different parts in different scenes (with costume changes, accents and makeup to match). But our rigid *NAO* robots are not so flexible, and we wish to avoid confusing the audience with double-jobbing performers. So, with $n = 2$ robots, one can play role A as the other fills role B, and only the gestures associated with those roles are ever performed. The model must also indicate the order of gestures for a given verb (e.g., should A gesture before B for the story verb *propose_to*?), and whether they should be enacted before or after the narrator (in this case, *Alexa*) vocalizes its narrative text.

The Dialogue Model

Scéalextric does not have its own dialogue model, since its stories are rendered in a neutral third-person voice. While we can expect a robot's gestures to speak for themselves, to an extent, such actions are rarely unambiguous, and it is just more natural for gestures to accompany live speech than omniscient narration. Moreover, spoken dialogue and physical action enrich each other when they are performed together.

The picture is complicated somewhat by the use of characters in supporting roles that are *not* embodied in the show. Roles such as A_{spouse} and B_{lawyer} have no robot presence, no gestures to perform, and no dialogue to utter, yet their presence in the narrative must still be felt by the audience. The model thus encompasses two kinds of dialogue: that which is uttered by embodied actors as they enact an event in which they appear, and that which they say to each other to comment on the unseen actions of other, disembodied roles. The former is *embodied dialogue*, the latter *meta-dialogue*.

Gestures and speech acts are expressions of the same urge to communicate, albeit in different modalities, so embodied dialogue is modeled in much the same way as physical gestures. For every story verb in the generator's inventory, we simply define a set of apt vocalizations for the roles involved. Take, for example, *propose_to*: an actor in the agentive role may say "Will you marry me?" or even "It's time to take our relationship to the next level," while the actor in the patient role may reply "Wow, I don't know what to say." A reply must sound natural while being suitably vague, since the actors don't yet know how the plot will unfold; the proposal may well be rejected in the next beat. In any case, all speech acts must be delivered in an appropriate sequence, and timed to enhance the gestures that are linked to the verb.

Meta-dialogue is a special case that comprises speech-acts that are uttered by actors in the central roles A and B, about characters that cannot be seen. Although the narrator tells us about these characters, the actors cannot show them to us. Suppose the next story beat is A_{Spouse} *cheat_with* B_{Lawyer}. Our robot actors do not portray these characters, and cannot speak or gesticulate for them. Worse still, they have nothing to do or to say when the narrator speaks of these characters. Meta-dialogue allows A and B to talk to each other about A_{Spouse} and B_{Lawyer}, as though they were a Greek chorus. For example, A may say "I could kill that lawyer of yours," to which B might reply "Just wait until you see the bill!" These jokes are baked into the dialogue model, as speech acts associated with the action A_{Spouse} *cheat_with* B_{Lawyer}. Our goal here is not to invent new speech acts, but to give our actors stock dialogue for events the generator can anticipate.

Nonetheless, not every speech act is entirely scripted in advance. We allow our actors to ad-lib, by giving them underspecified dialogue of the form "You are $+quality$" or "You are $-quality$." At the time it is spoken, $+quality$ is replaced with a simile that accentuates *quality*, while $+quality$ is replaced with one that ironically undermines it. For instance, $+welcome$ may be replaced by "as welcome as a cool breeze on a summer's day," while $+welcome$ can make way for "about as welcome as a skunk at a garden party." The system has a large stock of 1000s of similes to draw upon, but can also search on the web for fresh ones.

The Spatial Model

Our robot actors can do more than wave their arms and bend their legs; they can move about the stage in ways that meaningfully reflect their relationship to each other. Space is rich in metaphorical potential, so we speak of close ties and distant acquaintances, of losing touch and of coming together, of keeping our friends close and our enemies closer. These spatial metaphors are rooted in deep-seated image schemas (Johnson 1987) that conceptualize abstractions such as *love* and *hate*, *trust* and *fear* in experiential terms. A basic image-schematic model for reasoning about spatial metaphors was developed in (Veale and Keane 1992). Dubbed *Conceptual Scaffolding*, the model allows the semantics of non-spatial

verbs to be specified using spatial primitives such as *up* and *down*, *connect* and *disconnect*, and *contain* and *release*. By retrofitting this model onto *Scéalextric*'s story verbs, we can enable our robot actors to move about the stage in accordance with the actions they are enacting in the narrative.

We focus here on the *connect* and *disconnect* primitives, which allow us to signal the current state of the narrative via the relative closeness of the robot actors. Certain *disconnect* verbs, such as *compete_against*, cause both to move apart, as each takes a step back, while asymmetrical verbs, such as *resent* and *distrust*, cause just one of the actors to move back. Similarly, some *connect* verbs, such as *live_with*, cause both robots to move closer together, while others, such as *spy_on*, cause just one to take a step closer, and yet others, such as *pursue*, cause one to move closer as the other moves away. In every case, the unit of relative movement is a single step. Each robot begins the performance at a distance of six steps from the other. We hypothesize that audiences will register their spatial dance at a semantic level, for as the plot brings actors closer together and further apart, space will serve as a conceptual scaffolding for the twists and turns of the plot.

Empirical Evaluation

Scéalability is designed to turn storytelling into a show. The narrative text of a story is augmented with spoken dialogue, gestures and stage directions for the actors to perform. Our major concern here is the value that embodied actors add to the telling, and so we focus on the relative contribution of gestures, which are often showy and pantomimic, and spatial movements, which more subtly achieve a cumulative effect. We expect each form of embodiment to be more effective when used coherently – that is to say, in line with the plot – and to add to the audience's appreciation of the story. We also expect their contributions to be additive, so that a performance with both is to be preferred over just one alone.

A Pilot Study

Each of our studies is based on the same *Scéalextric* story. Raters do not view the performances live, but watch video recordings of different settings. In a pilot study that we will only briefly summarize here, we showed our raters a recording of a complete *Scéalextric* story, which takes 3 minutes to view. All ratings are crowd-sourced on the *Amazon Mechanical Turk* (AMT) platform, and all questions are posed in a random order. We have paid 53, 52, 53 and 52 workers in 4 conditions ($Spatial_{Coherent}$, $Spatial_{Incoherent}$, $Pantomimic_{Coherent}$, $Pantomimic_{Incoherent}$). The length of this full story and its recording serves to dilute the effect of key actions and their embodied delivery. Nonetheless, the pilot shows that audiences prefer gestures a little more than spatial movements, and coherent over incoherent uses of embodiment. Statistically significant differences are found for each of these contrasts in user ratings on the *AttrakDiff* scale. Specifically, a post-hoc t-test shows a significant difference between *Spatial Movement* and *Pantomimic Gesture* ($p = 0.002$, Cohen's D$= 0.094$) with means and standard deviations $\mu_{Gesture} = 4.591$, $\sigma_{Gesture} = 1.636$ and $\mu_{Spatial} = 4.430$, $\sigma_{Spatial} = 1.785$, and a significant difference ($p = 0.002$) in favour of coherent action (Cohen's D$= 0.094$). Coherent spatial movement scores significantly better on the *AttrakDiff* scale than incoherent movement (Cohen's D$= 0.272$).

A Refined Study Protocol

Since *Scéalextric* stories often contain many story beats, with various twists and turns, we build on the pilot study to show raters the following story excerpt that highlights just two story beats, which they rate on the same questionnaire:

A=Hillary Clinton; **B**=Donald Trump;
B-*friend*=Melania Trump; **N**=Narrator

(*two robots embody* A *and* B)
N: Say hello to Hillary Clinton.
A:{waving}
N: And let us welcome Donald Trump.
B:{waving}

(*white text on a black screen, also spoken by Alexa*)
N: What if Hillary Clinton fell in love with Donald Trump? Hillary was attracted to Donald because he was rich, wealthy and privileged. In response, Donald flirted outrageously with Hillary. So Hillary went down on bended knee and proposed to Donald.

story beat: < A propose_to B >
A: {gesture: move_closer} "It's time we took our relationship to the next level. Will you marry me?"

(*white text on a black screen, also spoken by Alexa*)
N: But Donald felt a deep love for Melania. So Donald turned a cold eye to Hillary's entreaties. Well, Hillary took Donald hostage ...

story beat: < A release B >
B: {gesture: back-away} "Just let me go."
A: "I will release you."

(*white text on a black screen, also spoken by Alexa*)
N: Thereafter Hillary would say that it was the other way around: that it was Hillary who dumped Donald!

Raters in the following two studies view this excerpt enacted with a mix of narration, dialogue, gesture and/or spatial movement, and complete a questionnaire based on the *AttrakDiff* model of (Hassenzahl, Burmester, and Koller 2003) as previously used by (Striepe and Lugrin 2017). All ratings are crowd-sourced on the *Amazon Mechanical Turk* (AMT) platform, and all questions are posed in a random order. All crowd-sourced evaluations carry the risk that some raters will not fully engage, since they are paid a small sum per task, and provide unvarying or random responses. So, to manage the risk and exclude potential "scammers," we add three gold-standard questions to the questionnaire. Each has an obvious answer for those engaged in the task, such as "How many robots are visible on screen?" (answer: 2). Those who fail the gold-standard questions are excluded

from the study. Furthermore, we only allow raters with a *Master Worker Qualification* on AMT to submit responses. This qualification is granted to those who have provided a large number of valid responses in previous tasks.

The same questionnaire is used in each study, and comprises 2 parts of 7 items each (excluding the 3 gold-standard questions). In the first part, raters answer 7 questions of the form '*The performance of the robots is ...*' by choosing a value from 1 to 7 on these 7 *AttrakDiff* dimensions: **(I)** unpleasant/pleasant; **(II)** ugly/attractive; **(III)** disagreeable/likeable; **(IV)** rejecting/inviting; **(V)** bad/good; **(VI)** repelling/appealing; and **(VII)** discouraging/motivating. For the second part, raters signal their agreement with the following 7 statements with a value from 1 (strongly disagree) to 7 (strongly agree): **(VIII)** The robots appear human-like; **(IX)** The robots show their intentions; **(X)** I could act like one of the robots;**(XI)** The robot mirrored how I would react; **(XII)** I sided with one of the robots; **(XIII)** I am curious as to how the story continues; and **(XIV)** The robots' movements are appropriate to events in the story.

Study I: Coherence of Performative Elements

Physical movements by robot actors can be eye-catching, and reinforce the embodied nature of the performance. But do they also add to the narrative in any semantic fashion? To test whether gestures and spatial movements are understood as meaningful contributions to the tale, we evaluate each under two conditions: the coherent condition, in which gestures or spatial movements are chosen to suit the semantics of each story verb; and the incoherent condition, in which gestures are chosen randomly, and spatial movements are performed contrary to the underlying image schema (so *connect* verbs are treated as *disconnect* verbs, and vice versa).

Methods: We present relevant parts of a sample story with both coherent and incoherent embodiment to human raters in a crowd-sourcing study on AMT. Two conditions (coherent / incoherent) for two embodiment strategies (gesture and spatial movement) necessitates four independent trials. Each rater is shown a 1-minute video that narrates the story with on-screen text and a synthesized voice-over. To focus their attention, the performance of just two story beats is presented on screen. These involve the story verbs *propose_to* and *release*, which are rendered in the four trials as follows:

1. *Coherent Spatial Movement*: Robots move closer together (*propose_to*) and later move further apart (*release*).
2. *Incoherent Spatial Movement*: Robots move further apart (*propose_to*) and later closer together (*release*).
3. *Coherent Pantomimic Gesture*: Robot A bends its knee (*propose_to*). Later robot B opens both its arms (*release)*.
4. *Incoherent Pantomimic Gesture*: Robots A and B perform random pantomimic gestures (for each verb).

40 raters were recruited for each trial ($N = 40*4 = 160$), and each was paid 0.40$ for completing the questionnaire.

Analysis: Human ratings for each trial were acquired over several weeks. Not counting excluded responses, the four trials elicited 29 valid responses for *Coherent Gesture*, 28 for *Incoherent Gesture*, 32 for *Coherent Spatial Movement* and 29 for *Incoherent Spatial Movement* (N = 118). For an overview of the statistical test results, see Table 1. The factor of coherence shows a significant p-value for an ANOVA test (p-value = 0.000061, mean squares = 48.138 and F-values = 16.147). A post-hoc t-test results in significant differences for all coherent and incoherent conditions with Cohen's $D = 0.197$ (small to medium effect size) in favor of the coherent conditions. This is reflected in the average rating for all coherent conditions $\mu_{Coherent} = 3.820$ and average rating of all incoherent conditions $\mu_{Incoherent} = 3.480$. The coherent condition for both *Pantomimic Gesture* and *Spatial Movement* performs significantly better (Bonferroni-corrected $p - value = 0.001$ and $p - value = 0.047$, respectively) than the incoherent equivalent.

Type	A	B	A: mean/std	B: mean/std	p-value	Cohen D
	Coh.	Incoh.	3.820/1.749	3.480/1.704	0.000066	0.197
Space	Coh.	Incoh.	3.728/1.794	3.458/1.675	0.047*	0.155
Gesture	Coh.	Incoh.	3.921/1.693	3.503/1.734	0.001*	0.244

Table 1: Post-hoc test for comparison of coherent and incoherent modes of spatial movement and pantomimic gestures. *Bonferroni corrected p-value.

Results: Our hypothesis is thus supported, since we have shown that coherent uses of our two embodied narration strategies outperform the incoherent uses.

Study II: Relative Value of Embodied Strategies

In this experimental study, we focus on the value that coherent gestures and spatial movements add to a performance, whether individually (just one or the other) or both together.

Methods: This study evaluates three performance modes:

1. *Pantomimic Gesture*: the tale is performed with narration, dialogue and gesture, but no schematic spatial moves.
2. *Spatial Movement*: the tale is performed with narration, dialogue and schematic spatial moves, but no gestures.
3. *Combined Action*: the tale is performed with narration, dialogue, gesture and schematic spatial movements.

As before, in the *Spatial Movement* condition the robots face each other and move closer or further away as the plot progresses. The relative position of the robots at any time offers a spatial summary of their relationship status. For the *Pantomimic Gesture* condition, the robots do not alter their position in space, but do use iconic and showy gestures to communicate each story verb. For the *Combined Movement* condition, the robots apply both strategies together, i.e. they move to and fro as the plot demands, and they also make pantomimic gestures for each story verb in the plot.

The three one-minute videos[1] are presented to 120 raters (or 40 for each) on the AMT crowd-sourcing platform. Each rater is shown just one of the three performances and then asked to evaluate it using our $14 + 3$ item questionnaire. In return, each AMT rater is paid 0.40$ per questionnaire.

[1] See all of the recordings here: https://tinyurl.com/wpes3jl

Condition L	Condition R	L: Mean/Std	R: Mean/Std	p-value	Cohen D
Space	Gesture	3.728/1.792	3.921/1.691	0.316*	-0.111
Space	Combined	3.728/1.792	4.131/1.762	0.002*	-0.227
Gesture	Combined	3.921/1.691	4.131/1.762	0.206*	-0.121

Table 2: Post-hoc test for all three conditions. L and R denote conditions named in first (L) and second (R) column. *Bonferroni corrected p-value.

Analysis: All ratings were acquired over several weeks. Not counting excluded responses from those who failed the gold-standard questions, there are 32 valid responses for the *Spatial Movement* condition, 29 for the *Pantomimic Gesture* condition and 33 for the *Combined Movement* condition, for a total of $N = 94$ valid responses. An ANOVA reveals significant differences between the conditions, with $p = 0.0019$ (Sum of squares = 38.686, F-values = 6.292). A post-hoc t-test results in a significant difference between the *Spatial Movement* and *Combined Movement* conditions ($p = 0.002$ Bonferroni corrected). With a mean value $\mu_{Spatial} = 3.728$ and standard deviation $\sigma_{Spatial} = 1.792$ for *Spatial Movement* and a mean value $\mu_{Combined} = 4.131$ and standard deviation $\sigma_{Combined} = 1.762$ for *Combined Movement*, the effect favours the latter (Cohen's D = 0.227). Pairwise comparisons of *Spatial Movement/Pantomimic Gesture* and *Pantomimic Gesture/Combined Movement* do not reveal any significant results. An overview of our analysis can be found in Table 2. Statistical tests have been conducted on the accumulated test construct (of all 14 items) and the results are visualized in Fig. 2. The whiskers indicate the standard error of the mean ($\frac{\sigma}{\sqrt{N}}$).

Results: Our findings suggest that a mix of embodiment strategies – what we have called the *Combined Movement* condition – is more appealing to viewers than *Spatial Movement* alone. However, there is no significant difference between the latter and *Pantomimic Gesture*. It seems that the subtlety of the actors' image-schematic use of space is just as effective as their more showy pantomime actions, whether that is going down on one knee to propose, or making a servile bow to a dominant character. This result is important for more practical reasons too. Pantomime is achieved by a careful mapping of each story verb to one or more motor scripts in the robot's repertoire. The results are eye-catching but often ad hoc, and depend more on cultural associations than semantics. In contrast, the robots' spatial movements are governed by verb semantics, and follow generically from those semantics without the need for ad-hoc mappings.

But it must also be noted that *Combined Movement* does not significantly outperform pantomime on its own. Figure 2 shows that the margin of standard error around the mean for *Pantomimic Gesture* overlaps with that of the other two conditions. Even though the mean values $\mu_{Spatial} = 3.728$, $\mu_{Pantomime} = 3.921$ and $\mu_{Combined} = 4.131$ obey an ascending order, a significant statistical difference can only be found for the first and last of these. While a sample size of $N = 94$ is enough to show some effect, a future study on a larger scale should be more convincing on this front.

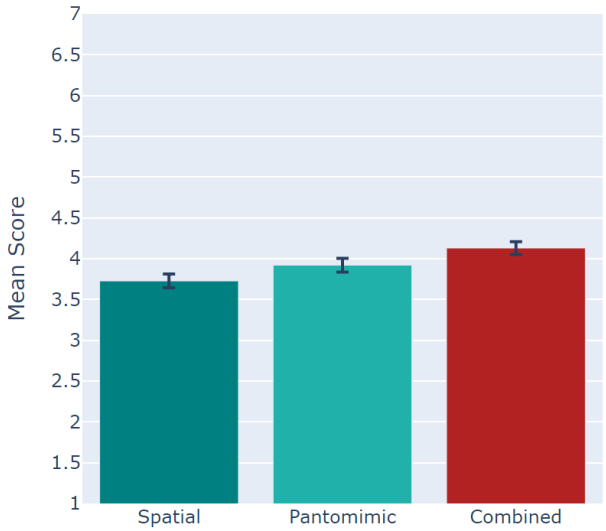

Figure 2: Mean ratings for the *Spatial Movement, Pantomimic Gesture* and *Combined Movement* conditions. The y-axis: mean ratings on the scale 1 to 7. The x-axis: three conditions. Whiskers show the standard error of the mean.

Conclusions and Future Work

The core insight of the presented research shows how spatial movement can be used to improve computational, embodied storytelling. So far, related works use a minimal set of scripted gestures tied to specific actions, which mostly show combinatorial novelty or mere generation. Scaling these storytelling performances to include new stories, new actors and more actions is problematic for scripted movements. Here, our approach of primitive motions shows its strength by allowing scaling with multiple actors, new stories and actions that only need to be associated on one dimension (*connect / disconnect*). We have provided empirical evidence that such approach is comparable with purely pantomimic performances (Study II).

With regards to interactive performance, we also see a role for gestures and spatial movements by the human audience. Robot performances are noisy affairs, and spoken dialogue must be timed so as to not overlap with the din of gears and servos in motion. So, what better way for the audience to convey their inputs to the story than by their own use of gesture and spatial metaphor? We are currently experimenting with visual analysis of the audience, and using emotion detection and pose estimation to recognize non-verbal inputs in the form of facial expressions (of surprise, anger, joy and sadness) and hand-gestures (thumbs up and down, rude finger gestures, aggressive fist motions, etc.). These won't just allow audience members to naturally influence the story line. They will make the audience performers in their own right.

References

Cienki, A. 2005. Image schemas and gesture. *From perception to meaning: Image schemas in cognitive linguistics*

29:421–442.

Cook, W. W. 1928. *Plotto: The Master Book of All Plots*. Tin House Books reprints 1981, 1981 edition.

Csapo, A.; Gilmartin, E.; Grizou, J.; Han, J.; Meena, R.; Anastasiou, D.; Jokinen, K.; and Wilcock, G. 2012. Multimodal conversational interaction with a humanoid robot. In *2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)*, 667–672. IEEE.

Dehn, N. 1981. Story generation after tale-spin. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'81, 16–18. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Gervás, P. 2013. Propp's morphology of the folk tale as a grammar for generation. In *Proc. of the 2013 workshop on Computational Models of Narrative*. Dagstuhl, Germany.

Hassenzahl, M.; Burmester, M.; and Koller, F. 2003. Attrakdiff: Ein fragebogen zur messung wahrgenommener hedonischer und pragmatischer qualität. In *Mensch & computer 2003*. Springer. 187–196.

Heider, F., and Simmel, M. 1944. An experimental study of apparent behavior. *The American journal of psychology* 57(2):243–259.

Johnson, M. 1987. *The Body In The Mind: The Bodily Basis Of Meaning, Imagination, And Reason*. University of Chicago Press.

Jokinen, K., and Wilcock, G. 2014. Multimodal open-domain conversations with the nao robot. In *Natural Interaction with Robots, Knowbots and Smartphones*. Springer. 213–224.

Klein, S.; Aeschlimann, J.; Balsiger, D. F.; Converse, S.; Foster, M.; Lao, R.; Oakley, J.; and Smith, J. 1973. Automatic novel writing: A status report. Technical report, University of Wisconsin-Madison Dept. of Computer Science.

Lebowitz, M. 1985. Story-telling as planning and learning. *Poetics* 14(6):483–502.

Lin, C.-Y.; Tseng, C.-K.; Teng, W.-C.; Lee, W.-C.; Kuo, C.-H.; Gu, H.-Y.; Chung, K.-L.; and Fahn, C.-S. 2009. The realization of robot theater: Humanoid robots and theatric performance. In *2009 International Conference on Advanced Robotics*, 1–6. IEEE.

Mathewson, K. W., and Mirowski, P. 2017. Improvised theatre alongside artificial intelligences. In *Thirteenth Artificial Intelligence and Interactive Digital Entertainment Conference*.

Meehan, J. R. 1977. Tale-spin, an interactive program that writes stories. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence vol 1*, IJCAI'77, 91–98. San Francisco, CA: Morgan Kaufmann.

Meena, R.; Jokinen, K.; and Wilcock, G. 2012. Integration of gestures and speech in human-robot interaction. In *2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)*, 673–678. IEEE.

Mittelberg, I. 2018. Gestures as image schemas and force gestalts: A dynamic systems approach augmented with motion-capture data analyses. *Cognitive Semiotics* 11(1).

Pérez, R. P. y., and Sharples, M. 2001. Mexica: A computer model of a cognitive account of creative writing. *Journal of Experimental & Theoretical Artificial Intelligence* 13(2):119–139.

Pope, V. C.; Dawes, R.; Schweiger, F.; and Sheikh, A. 2017. The geometry of storytelling: theatrical use of space for 360-degree videos and virtual reality. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 4468–4478.

Propp, V. 1928. *Morphology of the Folktale (translated by Laurence Scott)*. University of Texas Press, 1968 edition.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners.

Ravenet, B.; Clavel, C.; and Pelachaud, C. 2018. Automatic nonverbal behavior generation from image schemas. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 1667–1674. International Foundation for Autonomous Agents and Multiagent Systems.

Ritschel, H.; Aslan, I.; Sedlbauer, D.; and André, E. 2019. Irony man: Augmenting a social robot with the ability to use irony in multimodal communication with humans. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 86–94. International Foundation for Autonomous Agents and Multiagent Systems.

Rodriguez, I.; Martínez-Otzeta, J. M.; Irigoien, I.; and Lazkano, E. 2019. Spontaneous talking gestures using generative adversarial networks. *Robotics and Autonomous Systems* 114:57–65.

Striepe, H., and Lugrin, B. 2017. There once was a robot storyteller: measuring the effects of emotion and non-verbal behaviour. In *International Conference on Social Robotics*, 126–136. Springer.

Turner, S. R. 1993. *Minstrel: A Computer Model of Creativity and Storytelling*. Ph.D. Dissertation, Los Angeles, CA, USA. UMI Order no. GAX93-19933.

Veale, T., and Keane, M. T. 1992. Conceptual scaffolding: A spatially founded meaning representation for metaphor comprehension. 8(3):494–519.

Veale, T.; Wicke, P.; and Mildner, T. 2019. Duets ex machina: On the performative aspects of "double acts" in computational creativity. In *Proc. of ICCC'19, the International Conf. on Computational Creativity*.

Veale, T. 2017. Déjà vu all over again: On the creative value of familiar elements in the telling of original tales. In *ICCC*, 245–252.

Wicke, P., and Veale, T. 2018a. Interview with the robot: Question-guided collaboration in a storytelling system. In *Proc. of ICCC'18, the International Conf. on Computational Creativity*, 56–63.

Wicke, P., and Veale, T. 2018b. Storytelling by a show of hands: A framework for interactive embodied storytelling in robotic agents. In *Proc. of AISB'18, the Conf. on Artificial Intelligence and Simulated Behaviour*, 49–56.