<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>Computational Storytelling as an Embodied Robot Performance with Gesture and Spatial Metaphor</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Authors(s)</strong></td>
<td>Wicke, Philipp</td>
</tr>
<tr>
<td><strong>Publication date</strong></td>
<td>2021</td>
</tr>
<tr>
<td><strong>Publisher</strong></td>
<td>University College Dublin. School of Computer Science</td>
</tr>
<tr>
<td><strong>Item record/more information</strong></td>
<td><a href="http://hdl.handle.net/10197/12891">http://hdl.handle.net/10197/12891</a></td>
</tr>
</tbody>
</table>

The UCD community has made this article openly available. Please share how this access benefits you. Your story matters! (@ucd_oa)
# CONTENTS

1 INTRODUCTION ................................................................. 1
   1.1 Computational Story Generation ................................. 2
   1.2 Computational Story Telling .................................. 5
      1.2.1 Signs: Symbols and Icons ................................. 6
      1.2.2 Image Schemas .............................................. 9
      1.2.3 Conceptual Structures ................................... 10
   1.3 Embodied Story Telling ......................................... 12
      1.3.1 Embodied Language ........................................ 12
      1.3.2 Gesture and Sign ........................................... 13
      1.3.3 Embodiment with Nao .................................... 17
      1.3.4 Movement in Space ....................................... 20
      1.3.5 Embodied Creativity ...................................... 22
   1.4 Research Questions .............................................. 23
   1.5 Structure and Contribution .................................. 25

2 BACKGROUND RESEARCH .................................................... 28
   2.1 Storytelling .......................................................... 28
      2.1.1 Computational Story-Generation ......................... 29
      2.1.2 The Scéalextric System .................................. 36
      2.1.3 Embodied, Automated Storytelling ....................... 41
   2.2 Embodiment .......................................................... 43
      2.2.1 Embodied Robotic Performances ......................... 43
      2.2.2 Meaning-Making in Movement ............................ 45
      2.2.3 Multi-Agent Systems .................................... 46
   2.3 Space ................................................................. 47
      2.3.1 Embodied Cognition ....................................... 47
      2.3.2 Proxemics and the Stage ................................ 48
      2.3.3 Image Schematic Uses of Space ......................... 49
   2.4 Gesture ............................................................... 51
      2.4.1 Types of Gestures .......................................... 51
      2.4.2 A Computational Look at Gesture Generation ....... 53
   2.5 Image Schemas ...................................................... 55
      2.5.1 Properties of Image Schemas .............................. 56
      2.5.2 Quantity and Category of Image Schemas ............... 57
      2.5.3 Application of Image Schemas ............................. 59
      2.5.4 Image Schemas in Gestures ............................... 61
      2.5.5 Embodied Storytelling and Image Schemas ............. 63
   2.6 Creativity and Embodiment ................................... 64
      2.6.1 Computational Creativity ................................ 64
      2.6.2 Co-Creativity .............................................. 67
      2.6.3 Embodied Creativity Research ........................... 68
   2.7 Conclusion: Literature Review ................................ 71
3 A CONCEPTUAL FRAMEWORK FOR EMBODIED PERFORMANCE WITH ROBOTIC ACTORS

3.1 Performance Framework for Robotic Actors 74
3.1.1 Taxonomy of Robotic Movement 76
3.1.2 Combinations of Robotic Movement 81
3.1.3 Performance Framework at Work 83

3.2 Interpretation Framework for Storytelling Robots 85
3.2.1 The Representation of Gestures within the Framework 86
3.2.2 Selective Projection for Creative Interpretation 87
3.2.3 Emotional Valence in Story Progression 90
3.2.4 Opportunities for Metaphor and Irony 93

3.3 Integrative Framework for Robotic Performers 95
3.3.1 The Scéalability Architecture 95

4 PRACTICAL IMPLEMENTATION 99

4.1 Idea, Depth and Structure 99

4.2 Scéalability 100
4.2.1 The Gesture Module (GM) 100
4.2.2 The Spatial Module (SM) 103
4.2.3 Evaluative Aspects 106

4.3 Therapist Storyteller 107
4.3.1 Baseline Mode 107
4.3.2 Therapist Mode 108
4.3.3 Evaluative Aspects 111

4.4 Double Act 111
4.4.1 Storytelling with Alexa 112
4.4.2 Adding Nao to the Double Act 113
4.4.3 Domain Model and Meta Model 113
4.4.4 Implementing the Double Act 115
4.4.5 The Double Act in Action 117
4.4.6 Evaluative Aspects 120

4.5 Story Enactment 121
4.5.1 Evaluative Aspects 122

4.6 Interactive Storytelling 122
4.6.1 System Implementation 124
4.6.2 Inserting Interaction Points 124
4.6.3 The Neural Network Models 125
4.6.4 Evaluative Aspects 127

4.7 Nao Robot as Storyteller 128
4.7.1 Technical Solutions 128
4.7.2 Why Choose the Nao? 130

4.8 Demonstration 131

5 EMPIRICAL EVALUATION 133

5.1 Experiments: Embodiment 133
5.1.1 Hypothesis 133
5.1.2 Experimental Setup 134
## CONTENTS

5.1.3 Methodology ........................................ 135
5.1.4 Analysis ........................................... 137
5.1.5 Results ........................................... 138
5.2 Experiments: Space & Gesture ......................... 139
5.2.1 Hypothesis ......................................... 139
5.2.2 Pilot Studies ....................................... 141
5.2.3 Questionnaire Items ................................. 141
5.2.4 Preliminary Findings ............................... 142
5.2.5 Revised Presentation Format ......................... 145
5.2.6 Main Study ......................................... 145
5.2.7 Methods ........................................... 147
5.2.8 Analysis ........................................... 148
5.2.9 Results ........................................... 149
5.2.10 Complementary Study ............................. 151
5.3 Recapitulation ......................................... 154

6 DISCUSSION AND CONCLUSION .......................... 155
6.1 Discussion ........................................... 155
   6.1.1 Connecting Start and End ....................... 155
   6.1.2 A Conceptual Framework ....................... 156
   6.1.3 Implemented Performances ...................... 157
   6.1.4 Addressing the Research Questions ............ 158
6.2 Conclusion ........................................... 165
   6.2.1 Contributions .................................. 165
   6.2.2 Future Work .................................. 168

A ETHICAL APPROVAL / EXEMPTIONS ....................... 170

B QUESTIONNAIRES ..................................... 171
   B.1 Embodiment Experiments Questionnaire ........... 171
   B.2 AttrakDiff - Items ................................. 172
   B.3 Space & Gesture Experiments Questionnaire ...... 173

C EXAMPLE: GENERATED STORY ............................ 174

BIBLIOGRAPHY ........................................... 176
ABSTRACT

A story comes to life when it is turned into a performance. Computational approaches to storytelling have primarily focused on stories as textual artifacts and not as performances. But stories can become much more when they are augmented with actors, dialogue, movements and gestures. Where artificial intelligence research has previously investigated these individual layers, this thesis presents an overarching framework of computational storytelling as an embodied robot performance with a focus on gesture and spatial metaphor.

This work regards storytelling as a performative act, one that combines linguistic (spoken) and physical (embodied) actions to communicate concepts from performer to audience. The performances can feature multiple robotic agents that distribute the different storytelling tasks across themselves. The robots narrate the story, move across the stage, use appropriate gestures, interpret the actions of the story, present dialogue or give the audience an opportunity to interact with verbal or non-verbal cues, while an underlying system provides the story in an act of computational creativity. The performances are used to evaluate the links between concepts, words and embodied actions.

In particular, the robots connect two movement types with the underlying plot: Gestures to enhance theatricality, and spatial movements to mirror character relations in the plot. For both types, we present a comprehensive taxonomy of robotic movement. Moreover, we argue that image schemas play a profound role in the understanding of movement and that, based on this claim, the coherent use of schematic movement is beneficial for our performances and for researchers in the field of robotic performances.

To test these claims, the thesis outlines the Scéalability framework for turning generated stories into performances, which are then evaluated in a series of studies. In particular, we show that audiences are sensitive to the coherent use of space, and appreciate the schematic use of spatial movements as much as gestures.

Keywords: computational creativity, automated storytelling, story-generation, gesture, image schema, human-robot interaction
DECLARATION

I hereby certify that the work contained in this document is my own work which was completed while registered as a candidate for the degree stated on the Title Page.

No portion of the work contained in this document has been submitted in support of an application for a degree or qualification of this or any other university or other institution of learning.

All verbatim extracts have been distinguished by quotation marks, and all sources of information have been specifically acknowledged.

__________________________
Philipp Wicke,
August 11, 2021
COLLABORATIONS

Where the research activity for the thesis was undertaken jointly with others, I list below my collaborators and/or co-authors, including a short description of the nature of the contribution made by each author.

- **Prof. Tony Veale** contributed as a supervisor and to the design and development of various ideas of this thesis. As a co-author of all published research papers, he provided feedback and revisions in both design and implementation of the research. In particular, the implementation of the story-generator *Scéalextric* (Sec. 2.1.2) is his work.

- **Thomas Mildner** contributed as a M.Sc. student to the design and development of the *Double Act* (Sec. 4.4). In particular, Mildner implemented the software that links an *Amazon Alexa* and the storytelling performances as published in [1].

- **Stefan Riegl** contributed as a visiting research student to the design and implementation of the single storyteller conditions (Sec. 4.3). In particular, Riegl has supported the programming required to connect the storytelling system Scéalextric with a search algorithm that supports the retrieval of stories with particular topics and plot configurations.

Particular implementations by my various collaborators have been highlighted at the appropriate sections in this thesis.
Parts of this thesis have been published in the following peer-reviewed papers.


## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>The different layers of an embodied performance. Story and audience can be connected through shared ideas grounded in multiple modalities. The modalities are the layers through which a story can communicate its themes to the audience and through which the audience can perceive the action. On the left, five examples and their use of modalities are depicted: Wicke and Veale [2], Radford et al. [3], Ham et al. [4], Braitenberg [5] and Heider and Simmel [6].</td>
</tr>
<tr>
<td>Figure 2</td>
<td>A sign is the result of the association between a signified (a mental image) and a signifier (the sound or word) [7].</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Storytelling in three different warrior signs. The drawing depicts an iconic representation of a warrior. This drawing of Mabuyu at Ubirr is an Aboriginal rock art painting. The word is a symbolic representation of ‘warrior’ in the Arabic language. There is no resemblance between the word and what it signifies. The items are an indexical representation in the sense that helmet, shield and sword are non-arbitrarily connected to warrior and the concept of warrior must be inferred from them.</td>
</tr>
<tr>
<td>Figure 4</td>
<td>Spatial image schemas as visual representation. Studies in cognitive linguistics argue that the relations UP/DOWN, LEFT/RIGHT, BACK/FRONT, CENTER/PERIPHERY and other spatial schemas are learned through recurring physical experiences [8].</td>
</tr>
<tr>
<td>Figure 5</td>
<td>The spatial operator CONNECT/DISCONNECT applied to metaphoric and literal events. The first two examples show the CONNECT operation in a literal and metaphoric case respectively. The latter two examples provide the same for the DISCONNECT operation. Adapted from Veale and Keane [9].</td>
</tr>
</tbody>
</table>
Figure 6  American Sign Language (ASL) sign that can only be understood by ASL signers. The meaning of the sign cannot be inferred by the movement, because it is an arbitrary sign, a symbol in semiotic terms.  

Figure 7  Sign used by non-signer and signers alike. This movement of the hand to the mouth has iconicity and can generally be well understood, because the movement directly relates to the described act.  

Figure 8  Three social robots: Amazon Alexa is an immobile speaker with high-quality speech recognition and production capabilities. Reeti by Robopec has human-like facial expression capabilities and has been used by Striepe and Lugrin [10] as a robotic storyteller. The Nao robot is a bidpedal humanoid robot by SoftBank, which has been used in many studies as a storyteller [4, 11, 12, 13].  

Figure 9  Nao has 25 degrees of freedom in its movements. Its hands, legs and head can turn and rotate. The screenshots show the motor access through SoftBank’s Choregraphe software.  

Figure 10  The Nao robots Kim and Bap in a storytelling performance. Kim (left) is accusing Bap (right) by pointing at Bap. Bap is pulling its arms in front of the chest and placing its head down low in a defensive manner.  

Figure 11  Interpersonal distances, after Hall [14]. Proxemics, the study of how humans use space, indicates how embodied performances can make appropriate use of spatial zones.  

Figure 12  Each ring is one robot-step wide. The expansiveness of a robot’s gesture is tailored to its ring.  

Figure 13  Example of the ambiguity of gestures. The raised arm above head can vary in its interpretation, when presented with a different accompanying speech act. On the left, the gesture is iconic and the hand signifies the plane flying above the clouds. On the right, the gesture is a metaphoric signal of superiority.  

Figure 14  Depictions of the image schemas CONTAINMENT, IN and OUT adapted from Johnson [8](p.23, p.32 and p.32 respectively). These present only one possible instantiation of each schema.
Figure 15  Related image schemas can be combined and hierarchically arranged. Adapted from Hedblom et al. [15], who apply image schemas for concept invention. The image schemas at the top are the most general, most primitive schemas. The ones at the bottom are the most specific, e.g. CLOSED PATH MOVEMENT with an additional distinguishing point. 60

Figure 16  Motion-Capturing evidence for image schemas in gestural enactment. Identified schemas include PATH, EXTENSION, STRAIGHT, BOUNDEDNESS, CONTAINER, HORIZONTAL, and LEFT-RIGHT. Adapted from Mittelberg [16]. 63

Figure 17  Suggested mechanisms that creative systems can implement to evolve from merely generative to truly creative systems. Some of these steps are discussed in detail by Ventura [17]. Notably, the role of consciousness is not considered by Ventura. 65

Figure 18  The conceptual framework is presented in a division of three parts. Each part provides a different argument: theoretical, practical and empirical evaluations are the outputs of the framework. 75

Figure 19  A taxonomy of robotic movements and their properties. Nodes to the left and right depict broad types of robotic movement and their specific sub-types (rotational, spatial, iconic gestures, narrative beats, etc). Shared properties are indicated with horizontal connections (dashed lines). 77

Figure 20  Combination of spatial and rotational movement in a temporal representation. Spatial movement can co-occur with rotational movement. Different directions can be achieved by combining rotations with positive or negative spatial moves. 82

Figure 21  The result of parallel spatial and rotational movement is a curved walk. Here $t_0$ is the robot’s position prior to movement. $t_1$ is the point when the robot shifts from forward movement to backward movement, whilst completing half of the 90 degree rotation. 82
Figure 22
Diagram of four possible tasks that can be accounted for by the framework. From left to right, each task poses some prototypical requirements for the corresponding software module, which can be addressed by the performance framework properties.

Figure 23
Depending on the narrative context, the action B insult A can cause A to feel "insulted" \((C(x) \rightarrow E(x))\) or to feel "attacked" \((C(x) \rightarrow E(\bar{x}))\). A’s emotional response will then dictate the actor’s physical reaction on stage.

Figure 24
Selective projection in a situation in which “being insulted” is presented as “being attacked”. Situational relevance is determined by the emotional valence of the current action. Figure based on context-dependent blending theory by Brandt and Brandt [18], and adapted from Li et al. [19].

Figure 25
Interpretative decision points in a script with 31 plot actions (all actions listed in Appendix C). Dotted circles mark dramatic shifts of inertial valence, which enable interpretative decision points. The blue line tracks the inertial valence of character A; the orange line tracks that of B.

Figure 26
Overview of the Scéalability framework. A story-generation system creates the stories that will be enacted by agents/devices that cooperate via a blackboard architecture.

Figure 27
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 frames alternate a few times. This gesture is a show of muscles.

Figure 28
Scene from a performance. The left robot (A) proposes to the right robot (B). In the background, an Amazon Alexa device narrates the action.

Figure 29
Schematic depiction of the spatial map that both robots use to signify their character relations. The robots will not cross the "+3" line in order to avoid bumping into one other. Moreover, their gestures become more subtle when the distance between robots is less than 3 units.
Figure 30  Example of a tree traversal for the story generation process. Grey nodes and edges indicate that a path has not been explored by the system. Red nodes and edges denotes those that have been answered with a “No” by the user. Green nodes have been validated by the user, giving the sequence Action 01, Action 02, Action 03 and Action 04 that start the story. An orange node is one that has been rejected by a direct question, but accepted by the user to be a hypothetical possibility. .......................... 109

Figure 31  Double Act of Nao and Amazon Alexa on an Echo device. .......................... 112

Figure 32  Blackboard logic for the system’s meta-models. Idea and conceptualisation of the blackboard logic has been developed jointly by Veale et al. [1]. Graphical rendering by Thomas Mildner, co-author of [1]. .......................... 116

Figure 33  Example of the pose recognition and facial expression recognition software with a cartoon inspired by Gladiator (2000) [20]. In this image the emperor decides the fate of the gladiator Maximus with a verso pollice, a turned thumb. Drawing by Tony Veale, co-author of [21]. .......................... 123

Figure 34  Two Nao robots act out a story using space and gesture. A display with a webcam provides live-feedback to users while tracking their emotions and gestures. An Echo device narrates the tale. .......................... 124

Figure 35  The user provides a thumbs up during storytelling, which alters the course of the story. .......................... 125

Figure 36  Demonstration of the storytelling robot with super-titles on a backing screen to increase comprehensibility. .......................... 130

Figure 37  Differences in likeability ratings for “Story” and “Performance” in the four conditions of the first experiment. Ratings for appreciation of the story are shown in blue, and for the performance in red. .......................... 138
Figure 38  Comparison of ratings for “... greatly contributed to the performance” (upper two) with respect to Alexa (orange) and Nao (blue), and for “... added a humorous dimension to the performance” (lower two) with respect to Alexa (orange) and Nao (blue). Each plot shows mean Likert-Scale ratings for the Double Act condition of Nao and Alexa. .......................................................... 139

Figure 39  Three AttrakDiff[22] constructs have been rated in two pilot studies. The first pilot presented a one-robot performance with spatial movements and gestures. The second pilot presented two robots with spatial movements and gestures. All 21 questionnaire items can be found in the Appendix: HQS & HQI in App. B.2 and ATT in App. B.3. ................................. 144

Figure 40  A storyboard of the key video segments presented to participants. Each black-screen panel provides context for the action to be enacted by the robots in the clip to follow. .......................... 146

Figure 41  Mean ratings for each item in all four test conditions. The items of the questionnaire (App. B.3) are shown on the x-axis, while the mean rating for each is shown on the y-axis. .......................... 148

Figure 42  Accumulated mean ratings for each item in all four test conditions. Note, y-axis displays cropped values of the actual 1-7 rating scale. The whiskers indicate the standard error of the mean. .......................... 149

Figure 43  Average ratings of all questionnaire items (Sec. 5.2.3) for Incoherent (cyan) and Coherent (purple) conditions. ................................. 150

Figure 44  Average ratings of all questionnaire items (Sec. 5.2.3) for Spatial (blue) and Gesture (green) conditions. ................................. 150

Figure 45  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. .......................... 153
## LIST OF TABLES

| Table 1 | A listing of spatial schemas with their robotic realizations, indicating an additional potential for metaphorical meaning. Each row contains a schema that has an inverse motion in another row. | 50 |
| Table 2 | Relevant image schemas from Johnson [8] and adapted from Cienki [23]. This list is not exhaustive. | 57 |
| Table 3 | Different story or plot generators listed with their deep structure, surface form and dialogue capabilities. The deep structure elements can be accessed by the performer with different degrees of freedom and reliability. | 72 |
| Table 4 | Examples of gestures: Iconic, Deictic, Metaphoric, Cohesive and Beats by a robot. Each gesture is ascribed a general property, along with its hardware requirements. | 79 |
| Table 5 | Depiction of the body movements Spatial and Rotational with their corresponding name, depiction with a physical robot, movement vector and properties in the respective columns. | 81 |
| Table 6 | Possible combinations of movement types. comb. (green) are combinable movements, restr. (yellow) are combinations that are only possible to some restricted extent and excl. (red) are movements that are mutually exclusive. | 83 |
| Table 7 | Configurations of Scéalability result in interactive/non-interactive, embodied/non-embodied and single/multiple agent implementations. | 96 |
| Table 8 | A 10-point valence scale is used to rate how an action affects a character emotionally. Apt gestures are associated with points on this scale. | 103 |
| Table 9 | Three speech acts and the corresponding uses of space by the performer. | 104 |
| Table 10 | The 9 chosen image schemas and their corresponding gestures. If a schema requires a direction along an axis, two gestures are provided, otherwise the same gesture is given in both columns. S-P-G is the Source-Path-Goal image schema (see Sec. 2.5). | 105 |
Table 11  Signal encoding examples of hand gestures and facial expressions for decisive questions during storytelling. .......................................................... 127
Table 12  Storytelling modes with their respective properties and links to recordings. The Robot icon stands for a Nao robot as an embodied actor. The Amazon Alexa icon indicates that an Amazon Alexa is part of the performance. If the mode is interactive, the -icon signals that an audience is participating in this mode. .......................................................... 131
Table 13  A total of four studies tests the different storytelling aspects of Scéalability. These studies are concerned with the perception of multi-agent story-telling in different configurations. .... 135
Table 14  Above: four experimental conditions for evaluating the contribution of Scéalability’s SM and GM models to a story performance. Below: two examples, “A likes B” and “A hates B”. .... 147
Table 15  Overview of the t-test statistics for all four conditions. The four factors concern spatial movement (coherent vs. incoherent), gesture (coherent vs. incoherent), embodiment type (space vs. gesture), and coherence (coherent vs. incoherent). Cohen’s D is the effect-size. .... 149
Table 16  Post-hoc test for all three conditions. L and R denote conditions named in first (L) and second (R) column. *Bonferroni corrected p-value. 153
ACRONYMS

AI  Artificial Intelligence
ASL  American Sign Language
AWS  Amazon Web Services
BDI  Believe-Desire-Intention
CC  Computational Creativity
CD  Conceptual Dependency
ERP  Event-related Potential
fMRI  Functional Magnetic Resonance Imaging
GPT  Generative Pre-trained Transformer
HRI  Human Robot Interaction
ICCC  International Conference on Computational Creativity
ISL  Image Schema Logic
LMA  Laban Movement Analysis
MAS  Multi-Agent System
ML  Machine Learning
NLP  Natural Language Processing
NOC  Non-Official Characterization
OWL  Web Ontology Language
RNN  Recurrent Neural Network
TTS  Text-to-Speech
ACKNOWLEDGEMENTS

First and foremost, I have to thank my research supervisor Tony Veale. Without his assistance and dedicated involvement in every step throughout the process, this thesis would have not been accomplished. I would like to thank you very much for your support, encouragement, puns and metaphors over these past four years. I would also like to show gratitude to my Doctoral Study Panel, including Prof. Silvestre and Prof. O Cinnéide. Prof. Silvestre always provided useful feedback and a new perspective.

Over the past four years, many researchers visited our lab and supported me in my work. First, I need to express my gratitude for Stefan Riegl who has been unwavering in their personal and professional support during the first months in Dublin. Without them, those months of rain and pints would not have been the same. Second, I am grateful that Marta Ziosi found her way to our lab. We have shared great months with inspiring discussions and fantastic movies. Third, I thank Thomas Mildner, without him the Double Act (Sec. 4.4) would not have manifested. Finally, I would like to thank Kevin Trebing, who has been an enthusiastic colleague, but more importantly a great friend for many years.

Along this path, I have met some truly inspiring colleagues with whom I have been honored to collaborate with. Marianna Bolognesi did not hesitate to start our collaboration on emojis irrespective of my early career stage. With curiosity and excitement, we went to publish three journal papers together, whilst she would always treat me as an equal contributor. She is and was always available for brainstorming and academic advise. Likewise, I would like to thank João M. Cunha for being a great colleague and an even better friend. He has been the best company at various conferences and always provided me with a helping hand. Lastly, I would like to thank my friend Hannes Wiesner for his ability to teach me statistics. I am, of course, also very grateful to count him as one of my friends. There are many other friends I would like to thank for listening to and supporting me over the past four years. Without my flatmates Kety, Francesca, Nabila and Eloi, my months in Dublin would not have been the same and I am truly grateful to have met these wonderful people. I am also grateful for Till, Katherine and Anna, who helped me getting through some cold, rainy winters. Thanks to Paulina for supporting me down-the-line.

Most importantly, none of this could have happened without my parents and my friend Nadja. Danke, Ute und Norbert - ohne Euch hätte ich es nicht geschafft. I am eternally grateful for their support and love.

This research is supported by a School of Computer Science Ph.D. Scholarship.
INTRODUCTION

They’re not looking for a story that tells them who they are - they already know who they are. They’re here because they want a glimpse of who they could be.

*Anthony Hopkins as Dr. Robert Ford in HBO’s Westworld (2016)*

The story of *Westworld* tells of a futuristic Western-themed amusement park in which androids embody cowboys, sheriffs, bandits and ordinary characters of the American frontier. In this story, human visitors can enter the park and enjoy live-action role-play. Where the android actors, called hosts, are following predetermined story loops which reset after every cycle, the human visitors can interfere with these loops or become part of them. Eventually, each story loop is designed by the makers of the park to lure the visitors into entertaining adventures. Even if the visitor enters the story-line and seemingly impacts the outcome of a story, the ending will always be predetermined by the makers of the park. The opening quote of this chapter cites Anthony Hopkins from HBO’s *Westworld*, in which Hopkins portrays Dr. Robert Ford, the maker of the park. In this quote, Ford expresses why their storytelling is so alluring to the visitors. Stories are being told not only as a mirror of who we are, but as a fantasy of who we want to be. *Westworld* shows that its visitors enjoy becoming heroes and villains in their own stories. The visitors attribute meaning to the signs of respect, reverence and submission that the android actors depict, but they can decide to deny any guilt or responsibility towards the robots. In the show, one of the visitors becomes hopelessly lost between reality and the fictional world of the park, as lines between visitors and hosts start to blur. The visitor gets lost because of the incredible amount of immersion and emulation in which their own life’s story intertwines with the hosts’ loops.

*Westworld* depicts the pinnacle of embodied, robotic performances. The robotic actors bring the ability to perform tirelessly, without conflict of emotions that may arise for human actors and their power to read, analyze and interpret the humans’ emotions and behaviour to change the story dynamically. While modern embodied performances with androids are far away from the quality depicted in *Westworld*, the individual modalities are not. Robotic movements have become very versatile, e.g. *Boston Dynamics’* robots can already perform fast and ro-
bust human-like movements [24]. Immersive storytelling experiences that capture users in days of adventures with artificial agents can already be experienced with the latest triple A video games. There are robotic improvisation/theater performances [25, 26], dancing with robots [27, 28] and many more robotic performances being developed. All of these individual research efforts provide the layers of an ultimate, unified robotic performance framework, which allows us to enact any sort of story. If we place Westworld at the top of embodied, robotic performances, it requires all of those multi-modal layers to synchronize and unify in order to realize a storytelling experience. These layers of performative features are the core of automated storytelling and the focus of this thesis. When we strip all the layers of performative features away, we can reveal the minimal requirements for storytelling and show what each individual layer adds to the performance. Only the interplay of those layers allows the story to convey meaning. If only actors, text and movement are present, without any interplay, the actors are nothing but marionettes executing a script. These considerations are especially salient in embodied performances of artificial agents, but they are much older than modern AI. For instance, the marionette theater of the 19th century moved Heinrich von Kleist to consider the naturalness of puppet storytelling [29]. Kleist believed that a self-aware actor cannot perform with complete naturalness, but a marionette - or, indeed, a robot - lacks the self-consciousness that can impede a graceful and fully immersive performance. Our robots lack human grace when they act, but we wish them to use their body in space in ways that seem natural and intuitive. Hence, the layers of performance that are especially important to this thesis concern gesture and space. At the same time, it is the automated writing process, which provides the script, that precedes the embodied performance and enables co-creative synergy between audience and actors.

1.1 COMPUTATIONAL STORY GENERATION

The story loops in Westworld are written by a designated Narrative Designer in order to bake the desires and demands of the guests into creative adventures. Modern research in computational storytelling tries to model this creative writing process to substitute the Narrative Designer with an automated narrative generator. These efforts have progressed from early attempts at automatic novel writing [30] to recent story generation systems based on large language models such as GPT-3 [3]. Computational storytelling is a subfield of Computational Creativity (CC), which Veale and Cardoso [31] define as the study of computers as autonomous creators and co-creators, which goes above the potential of computers to be mere feature-rich tools. These broader studies of CC encompass any domain with creative
potential: music composition [32, 33, 34], artistic creations [35, 36, 37], mathematical discoveries [38, 39] and poetic writing [40, 41, 42], just to name a few. Among those, the study of narrative generation based on language brings important implications relevant not only for embodied performances, but CC and beyond.

Textual storytelling is an exercise of language. The challenge of this exercise comes from the complexity and boundaries of language - to quote Wittgenstein [43] "The limits of my language mean the limits of my world". This alludes to linguistic relativity, which claims that complexity, boundaries and structure of language affect its speakers’ worldview or reasoning [44]. As such, being creative with language in storytelling taps into cognitive foundations of human thought. As studies from narrative psychology show, artificial agents are more comprehensible if their signified behaviour supports human, narrative interpretation [45, 46]. Turner [47] claims that narrative imagining enables humans to predict, plan and explain, which makes stories an essential tool of thought and reasoning. At the same time, the cognitive grounds of language offer many creative instruments such as metaphor and conceptual blending [48]. The best known example of narratives infused with metaphor and blends are fables, which date back to the ancient Greece. Early storytelling systems such as Tale-Spin [49] created Aesopian fables. Similarly, modern storytelling systems like those by Veale [50] fuse affective metaphors to generate stories with a moral. The generation of stories through computational generators depends a lot on their architects’ definition of a story.

In \textit{Westworld}, stories are narrative loops originating from the backstory of each character. These loops will follow a clear trajectory from a starting point to an end point and then reset to the starting point again, e.g. if a character dies at the end of their story, they start back alive without any knowledge of the loop. Human interaction can cause the story to deviate from its basic trajectory, but eventually the loop will always reach its final destination. This form of predestination can be used to limit the complexity of the systems. Moreover, tales and fables often rely on the fate and destiny of their characters and it is the art of writing which augments their predetermination with unique and profound adventures. The famous James Bond movies always start under the same assumption that Bond will defeat the villain and save the world. Every Bond movie relies on the same recurring themes and even alludes to them, e.g. when Bond orders his Martini shaken, not stirred. The symbols remain the same (the villain, the "Bond girl") but their manifestations change. One would expect that any dangerous situation lacks the thrill of an actual threat to the protagonist if the protagonist is known to survive and meet their predetermined goals. This predetermination can facilitate the generation of coherent stories by the most minimal setup [51]. Notably, prede-
1.1 COMPUTATIONAL STORY GENERATION

termination is obvious on the level of the plot, but not necessarily on
the story-level. The plot is the sequence of events where each causes
the next one, whereas the story or narrative is the telling of the plot
through verbal, visual or written means. Gervás [52] refers to the dis-
tinction between fabula and discourse. The former is the "what is told"
and the latter the "how it is told". Under these considerations, it is
sufficient for a plot generator to produce the causal actions that can
be fleshed out by story writers. One of the earliest approaches to plot
generation goes back to William Wallace Cook’s 1928 book “PLOTTO:
the master book of all plots” [53]. In PLOTTO, Cook provides a large col-
lection of so-called master plots each comprising three parts. These
triplets can be used as plot skeletons for writers like Cook himself,
who needed to write novels in quick succession. This structural ap-
proach to plot generation heavily relies on the creative value of fa-
miliar narrative elements [54]. Moreover, this approach informs the
plot creation of the computational story generation system Scéalextric
[50], which is supporting the embodied performances studied in this
thesis.

Scéalextric\(^1\) is an automated knowledge-based system for generat-
ing plots and rendering them as idiomatic English texts. Analogous
to Plotto, the system uses action-triplets which causally connect the ac-
tions of one character with the reaction of another. The resulting chain
connects actions into a moral narrative, e.g. the heroine redeemed, the
villain defeated. Scéalextric provides a solid plot structure which has both,
a surface linguistic form and a deeper conceptual representa-
tion. Being a symbolic AI system, both of these, the surface form and
deeper structure, are readily accessible and can be altered to study fab-
ula and discourse. For example, on the linguistic surface Veale [55] has
shown that the generated stories are rated higher on dimensions such
as entertainment, laughter and drama when they use characters from
a list of well-known fictional and non-fictional characters. At this sur-
face level, the layers of performance can largely change and augment
the plot in different ways, e.g. embodied actors, visual symbols, move-
ment and more. In contrast, the deeper conceptual representation (e.g.
the heroine’s journey) should remain stable no matter what layers
augment it. Computational storytelling is not only the study of well-
written coherent narratives, but its relevance for narrative studies, cre-
ativity research and cognitive science becomes apparent in its theory
and practice with the deeper conceptual mechanisms of its artifacts
and algorithms. With a focus on embodied performance, this thesis
investigates how a computational storytelling system can make use

\(^1\) The name is derived from the conceptual blend of scéal (the Irish word for story) and
Scalextric (a track based slot car racing system, which allows gamers to connect track
pieces to create racing circuits)
The different layers of an embodied performance. Story and audience can be connected through shared ideas grounded in multiple modalities. The modalities are the layers through which a story can communicate its themes to the audience and through which the audience can perceive the action. On the left, five examples and their use of modalities are depicted: Wicke and Veale [2], Radford et al. [3], Ham et al. [4], Braitenberg [5] and Heider and Simmel [6].

of its deeper representations on the different layers of performance (see Fig. 1), most importantly movement and gesture.

1.2 COMPUTATIONAL STORY TELLING

The first sentence of Ford’s introductory quote states that the audience of the storytelling performance (visitors to the Westworld amusement park) is not looking for a story to tell them who they are, since they already know who they are. Every human observer who participates in an embodied performance will always bring their own story. The story of each human is a narrative spun from their past experiences. Past experiences, the mental model we construct about the world, and our memories, have a profound and direct influence on our perception of the world [56, 57]. This influence, through the narratives that people bring with them, is the most foundational ingredient of storytelling. Humans attribute narrative intent to simple inanimate
objects, even in the absence of an actual plot or story. This effect was first described in the 1944 experimental studies of Heider and Simmel [6], in which observers attribute a narrative to the movement of simple geometric shapes. Just as humans interpret the movement of symbols on a screen as narrative processes, they bring their own experience of the world to the interpretation of an overt narrative. The narrative emerges from the movement of symbols. Hence, the foundation layer of automated storytelling is the symbolic layer. Storytelling using static symbols is as ancient as humanity itself, e.g. cave paintings of the Aboriginal Australians tell the stories of challenges faced by previous generations. Eventually, these symbols evolved into complex, logographic writing systems such as the Egyptian hieroglyphs and Mayan glyphs. Signs and symbols carry metaphoric and schematic meaning, which connect with an observer’s experiences, allowing one to radically reduce complex, abstract concepts into a discrete set of symbols. This form of reduction allows complex narrative structures to be represented using a small number of primitives [58, 59]. Modern versions of automated, visual storytelling with symbols can involve Emoji [60], Mayan-like narrative art pieces [61] or comic strip generation [62]. In particular, Emoji show how modern symbols at our fingertips can carry multi-modal meaning and depict both concrete and abstract concepts [63]. Through these symbols, the ideas of the story and the ideas of the audience are connected, since they rely on the same set of conventionalized meanings.

1.2.1 Signs: Symbols and Icons

At this point, it is important to disambiguate the semiotic terminology further. There are some important distinctions between a sign, a symbol and an icon. Especially the notion of iconicity is important in the context of iconic movements and gestures. One relevant distinction about signs is the one on signifier and signified. For example, the written word HOUSE evokes the mental image of a house. The HOUSE is the signifier and the mental image of the house the signified. This dyadic approach goes back to de Saussure’s 1916 Course in General Linguistics [7] which distinguished the sound-image (signifier) from the concept (signified). The two relate through their conventionalized relationship. The combination of the two constitutes the sign (see Fig. 2). In order to understand the meaning of a sign, its relation to other signs within the system has to be understood. For de Saussure the spoken sign, the sound pattern, was the primary sign system and the linguistic sign a separate secondary sign system. The written letters (secondary signs) signified sounds in the primary sign system. Yet, most semioticians and this work refer to the form of signs as spoken or written [64].
The distinction between signifier and signified allows a more fine-grained understanding of the ways in which meaning can be evoked through signs. In particular, a sign differs in how arbitrary or conventional it is. Instead of a continuous spectrum, Peirce [65] at some point concludes that there must be almost sixty thousand classes of signs on the spectrum from arbitrary to conventional sign. So, Peirce provides a more useful categorization of the sign into icon, symbol and index. Chandler [64] relates this categorization to the Saussurean terms of signifier and signified in the following definitions:

- **Icon/Iconic:** Describes the sign in which the signifier directly resembles the signified. The signified can be understood by the similarity of the signifier (e.g. cartoon, drawing, model, onomatopoeia).

- **Symbol/Symbolic:** Describes the arbitrary sign. In the sense that the signifier does not resemble the signified, but the connection is purely based on convention and has to be learned (e.g. alphabetic letters, words, numbers or road signs).

- **Index/Indexical:** An index sign is not arbitrary, but the signifier is connected in some physical or causal way to the signified. For example, footprints signify a path, a clock can signify time and a pointing (index) finger can signify a direction.

The combination of signifier and signified define the class of sign. A signifier is not a sign, only the combination of a signified allows a signifier to work as part of a sign. Returning to storytelling, a story can be told with an icon, symbol or index. Consider the sign for ‘warrior’ depicted as icon, symbol and index in Figure 3. The icon (Fig. 3 left) is an Aboriginal rock art painting depicting a warrior. The signifier resembles the signified. The icon shows a stylized stick-figure with weapons. The iconicity of this depiction is more apparent in the context of other signs around it, e.g. stick-figures without weapons or drawings of a battle. Yet, this icon can tell the story of a warrior and the drawing shares enough similar properties with ‘warrior’ to be iconic. The symbol (Fig. 3 center) is the Arabic word ‘muharib’ which is a signifier for ‘warrior’. It is a symbol, because the connection between the signifier and signified is arbitrary and has to be learned.
Figure 3: Storytelling in three different warrior signs. The drawing depicts an iconic representation of a warrior. This drawing of Mabuyu at Ubirr is an Aboriginal rock art painting. The word is a symbolic representation of ‘warrior’ in the Arabic language. There is no resemblance between the word and what it signifies. The items are an indexical representation in the sense that helmet, shield and sword are non-arbitrarily connected to warrior and the concept of warrior must be inferred from them.

Even an abandoned shield, sword and helmet (Fig. 3 right) on a battlefield can tell the story of a warrior. Those signifiers are physically and causally connected to the warrior, hence they support an indexical sign of a warrior.

Signs play a fundamental role in storytelling, because they express ideas in either arbitrary or direct ways to the audience. Iconicity allows the storyteller to directly express meaning, which follows the performative credo “Show, don’t tell”. Storytelling with symbols allows an easy identification of signifier and signified, as exemplified by Fig. 3. This identification becomes more difficult when the story is augmented with different modalities. The various ways in which performances can augment a story are visualized in Fig. 1, along with a key to how different performances draw from the various layers. Notably, in Heider and Simmel’s experiments the story emerges through the movement of the shapes. This movement allows for anthropomorphism and augments the story with an additional performative layer. Introducing movement brings a new set of implications for storytelling. At the same time, movement can be iconic or indexical, relational or arbitrary as much as signs are.

If movement is introduced in a performance, it necessitates the concept of space. In this space, the movement of objects, signs or agents is inherently relational, i.e. the movement of one entity changes its proximity to all other entities or observers. Here again, the movement of the entity can be schematic in nature, and tap into complex meaning systems. Those movement schemas are not just bound to symbols in the two dimensional plane. Movement can also be layered with agents instead of symbols. A prominent example, that takes Heider and Simmel’s two dimensional idea to the three dimensional plane,
comes from Valentino Braitenberg’s vehicles [5]. A Braitenberg vehicle is an autonomous agent that can move freely in its environment, steering its path based on sensory information. In its most simplistic version a Braitenberg vehicle’s motion is directly controlled by some sensors, for example light sensors and heat sensors. Yet, the resulting behaviour might appear intentional, complex or even intelligent. Each of these kinds of performances, Heider and Simmel’s moving shapes and Braitenberg’s moving agents, evoke a narrative for the observer through the schematics of the movement. These schematics mirror patterns of understanding and reasoning in the observers and can be described as image schemas.

1.2.2 Image Schemas

The deeper representations of a plot should become apparent irrespective of the type of performance. Hence, it should not matter whether the story is being performed by embodied or non-embodied actors, whether there are speech acts or movement, the representations should be communicated regardless. The language, movement or visuals should therefore rely on conceptual patterns which are stable across the modality. It has been argued that conceptual patterns are built from bodily interactions and linguistic experiences [66]. Johnson first described those imagistic patterns as Image Schemas [8] and they have since been described in more detail by Lakoff [67] and became a field of study in Cognitive Linguistics.

For example, all human beings are subjected to Earth’s gravity, therefore there is an inherent, experienced sensation of what is UP and what is DOWN. Other bodily interactions necessitate an understanding of what is NEAR and FAR or FRONT and BACK. By living and moving in the same physical space, humans share some common concepts of their bodily experience. These spatial schematics are argued to be fundamental semantic building blocks of human reasoning and concept formation [67]. The metaphoric power of those spatial schematics within the layers of performance is a key topic of investigation in this work. Image schemas are powerful abstractions that capture the spatial/temporal logic of an action, as shown in the following instantiations of the outward motion:

John went out of the room.
Pump out the air.
Let out your anger.
Pick out the best theory.
Drown out the music.
(Selected examples from Johnson [8] p. 32 - Fig. 4)

Although there is no fixed set of image schemas, there is a set of reliably recurring schemas that have been studied in cognitive linguistics:
CONTAINER, OBJECT, PATH, PART-WHOLE and UP-DOWN [8, 67]. The ideas, categories and use cases of image schemas will be discussed in detail in Section 2.5. To narrow down the various different image schemas provided by theory and empirical research, this thesis considers a set of schemas defined by practical considerations. Hurtienne and Blessing [68] group the following image schemas that relate to space: UP-DOWN, LEFT-RIGHT, NEAR-FAR, FRONT-BACK, CENTER-PERIPHERY, CONTACT, PATH, SCALE. Those schemas are the basis for the empirical work and investigation of this thesis, and are depicted in Figure 4.

1.2.3 Conceptual Structures

Returning to the case of Heider, Simmel and Braitenberg, the observer infers a story, or rather attributes character and intentional-ity from their own perspective. Automated storytelling works from the other end, communicating an idea through a story to the audience. The conceptual primitives provided by a layer of symbols and movements can thus be utilized by an automated storytelling system to formulate a story expressing its target ideas and themes. But not only movement and symbols draw from these schemas. Words, gestures, speech, in fact all of the layers, have a schematic underpinning that can be revealed and used in embodied storytelling. When words and speech convey meaning, their conceptual structures are often rooted in metaphor [48]. The conceptual organization of our most imaginary and abstract thoughts can be investigated through spatial schemas [9]. What is depicted as a simple visual representation can reach the imagination of complex narratives and abstract concepts. The deeper conceptual representations of a story can be visualized with image schemas. A good example of this is the Conceptual Scaffolding proposed by Veale and Keane [9]. In their theory, the authors argue that spatially-founded representations can be exploited for metaphor comprehension. Through this, the scaffolding relies on a small set of core metaphors to generate visual representations of complex situations. The basic example that Veale and Keane [9] provide
is of the schemas CONNECT/DISCONNECT and shows literal and metaphorical uses of spatial collocation (see Figure 5).

The schemas described in Fig. 4 can also be used to describe events as spatial operations in literal and metaphorical ways. In other words, the schemas can be used to make the implicit or internal representations of a situation explicit. For storytelling, this means that on a plot level (a sequence of events) these schemas are apparent and should be brought forth in implicit ways on any layer of the performance.

For textual storytelling, Veale and Keane [9] provide further examples of how narratives can be represented using spatial operators. For example, the story trope “From Rags to Riches” can be represented visually as “Rags ← ◦ → Jerry ← ◦ → Riches” (see Fig. 5). In this case there is a metaphorical use of the CONNECT/DISCONNECT spatial schemas for Jerry representing his transition from rags (a metonym for poverty) to riches. Not only do image schemas provide scaffolding for narrative processes, they can also provide an algebraic basis for modeling complex processes and situations [69]. It allows us to use simple schemas as primitive building blocks of larger, more complex structures [70]. Likewise, there are examples of agent-based performances that rely on schematic movements. For example, Singh et al. [71] present a co-creative agent that interacts with users by classifying and responding to schematic movements in a 2D virtual environment. But if layers of symbols, words and speech are assumed to share the same conceptual schemas, what about more bodily layers of the performance?
1.3 Embodied Story Telling

1.3.1 Embodied Language

The layers of performance (Fig. 1) include embodiment and, in the most distant layer, gesture. Embodiment is deeply connected with language and the way we think [72]. The philosophical stance of Embodied Cognition roots our thinking and language to the body. The term Embodiment is a broad one that can be understood in colloquial terms with very few implications, but it can also be understood in philosophical terms with some profound implications for cognition and reasoning. In a general sense, embodiment describes a relational dependency which is enabled through a bodily manifestation. To state that embodiment requires a body is shifting the problem to the definition of a body. In colloquial terms, it suffices to leave embodiment to an open-definition of a body. In philosophical terms, embodiment was most prominently introduced by Varela et al. [73]. The authors differentiate the term embodiment as follows:

"First that cognition depends upon the kinds of experience that come from having a body with various sensorimotor capacities, and second, that these individual sensorimotor capacities are themselves embedded in a more encompassing biological, psychological and cultural context. Varela et al. [73] p. 172-173"

This definition of embodiment suggests that cognitive, conceptual representations are shaped and formed by the environment that the body interacts with. Moreover, Varela et al. [73] speak of Enaction in the sense that an agent’s perception is creatively adjusting its actions to match the needs of its situation [74]. The agent is creating their subjective experience through their own actions and interactions with the environment. Consequently, the subjective experience is limited by the affordances of the agent’s body. At the same time, the subjective experience forms memories which in return predict the processing of the perception. This again alludes to the same bottom-up and top-down processes that play a role in anthropomorphizing moving symbols. The notion of embodiment relevant in this work is focused on the connection between the linguistics of storytelling and the body in movement through locomotion and gesture.

The link between language and embodiment is supported by some compelling neuroscientific evidence. Rüschemeyer et al. [75] show greater activity in the sensorimotor and somatosensory cortex when participants processed basic motor verbs (to grasp) versus processing of basic abstract verbs (to think). Similarly, Pulvermüller and Fadiga [76] lay out the cortical anatomy underlying language processing. In this process the human motor system extracts phonological information relevant for speech perception. More precisely, Hauk et al. [77]
provide evidence from brain imaging studies which show the connection between language and the body more distinctly. In their study, the authors present words to be passively read by participants in an fMRI scanner. They show that words referring to a specific part of the body (e.g. pick, kick or lick) elicit a higher activity level in areas of the motor cortex. Moreover, these areas overlapped or were adjacent to those parts of the motor cortex which show heightened activation when the respective movement is performed. This means that reading the word to pick activates very similar brain regions as moving fingers. Because of this underlying connection movements and words have a shared conceptual basis. To look at embodiment in computational storytelling is to observe the body and language in a coupled play. This work considers the storytelling performance as a Language Game [78]. As such, storytelling becomes a simple example of language use, one in which language and the embodied performance are interwoven. The thesis will show how this sense of the term embodiment contributes to a storytelling performance.

This connection between the body and language can be observed by the appearance of the same image schematic notions. These linguistic image schemas can be observed in embodiment when looking at gesture. Often, gestures are only considered an epiphenomenon of speech, a mere by-product of articulation. Being the outer-most layer in most performances, the seemingly ancillary implementation of gestures often plays a subsidiary role. Yet, where the augmentation of the story seems to be most abstract, the schematic foundations and its ideas can still be identified. Gestures are interwoven with their cultural and ideological contexts and therefore they augment language with contextual semantics [79]. Language and thought are tightly connected with neural correlates of bodily motion [77]. In fact, there is evidence that image schemas are some of the underlying patterns of gestures themselves [16, 80].

1.3.2 Gesture and Sign

When we look at gesture and space, the literature considers both with respect to their effects in human-human interaction or human-robot interaction (HRI) [11, 13, 81, 82]. Often, they are added to enhance the performative value or communicative intent, and many gestures are merely ornamental. But the body in motion, specifically when gesturing, can be linked more deeply to the underlying meaning of a story. The deeper conceptual structures can be preserved on the level of embodiment. Narration can have expressive needs that are met by the affordances of the body. As such, gestures have the ability to be iconic, e.g. holding both hands ready, next to the hips to express one’s readiness for a cowboy quick draw duel. A gesture like that exemplifies the complex nature of these movements, which is often overlooked when
gestures are simply auxiliary. An actor might hold both hands ready, next to their hips in order to metaphorically depict readiness for a challenge. Alternately, the actor could literally hold both hands ready, next to their hips in order to quick draw a weapon. The metaphorical case transforms the visual movement into an imaginary threat, whereas the literal case is visual in movement and threat. The context which ultimately determines the meaning of the movement cannot be provided by gestures alone. Gestures can be constrained and informed by the words they co-occur with. These limits can also be overcome by other affordances, e.g. the actor could wear a cowboy hat and holster a gun. Alternatively, gestures can be deictic, i.e. pointing gestures.

Now, the context is inherently referential. Arguably, if an actor points with a real pistol or a finger pretending to be a pistol, it does indeed make a difference to the meaning, yet the referent is still the same.

These examples show the complex nature of gestures as semiotic signs, as embodied signs. Importantly, the notion of sign in the context of hand movement usually refers to Sign language and therefore some further distinctions need to be made. According to the Merriam Webster Dictionary Sign language is defined as

"any of various formal languages employing a system of hand gestures and their placement relative to the upper body, facial expressions, body postures, and finger spelling especially for communication by and with deaf people"[83]

The polysemy of the terms sign and gesture must be addressed in this context and the difference between telling a story using Sign language or telling a story using gestures must be clarified. In each of these situations, a storyteller may use their hands, arms and entire upper body to convey meaning. While some of the gestures in Sign language can be meaningful to signers and non-signers, most non-signers cannot guess the meaning of arbitrary signs (in American Sign Language (ASL) [84]). Conversely, an audience is unlikely to guess the meaning of naturalistic gestures without some accompanying speech if the movement is not iconic or otherwise mimetic. Moreover, a set of gestures is not, in itself, a language. A language requires morphology and a grammar. Sign language exhibits all of the fundamental properties of languages [85] while gestures alone do not. Sign language uses gestures as part of a systematic whole that involves grammar and morphology. Telling a story in Sign language allows the teller to convey the exact elements of a story and simultaneously bring a form of bodily enactment due to the iconic nature of certain signs. Telling a story using only gestures and enactment requires a high degree of iconicity and physical mimesis, but will not fully convey the elements of a story with the same linguistic precision.
The signs of a Sign language are often *symbolic* in the Saussurean sense, i.e. the movement or hand shape is an arbitrary sign where the signifier does not resemble the signified. This is true for most, but not all signs of Sign languages. Some signs are iconic and resemble the signified, much like the onomatopoeia of certain spoken words, and others employ obvious bodily metaphors. For example, ASL signs that relate to emotions such as *happy*, *angry* and *feel* have their signed space on the chest, whereas more cognitive signs, such as *know*, *think* and *understand* have their signed space around the head [86]. So, iconicity is clearly present in Sign language [87], but most Sign language signs are symbolic. The signers’ symbols are meaningful units (morphemes), which are themselves combined from the Sign language equivalent of “phonemes”. Phonemes of sign language are combinations of features such as hand-shape, orientation, location or movement.

In contrast, the cowboy example shows that gestures are often not symbolic but iconic or indexical. **Iconic gestures**, as opposed to the signs of Sign language, do not need to be learned in order to be understood, because the gesture can be understood based on the similarity of the signifier, to that which it signifies. The indexical **deictic gesture** signifies a direction and is therefore not arbitrary, i.e. the pointing also connects in some physical way to the signified. Apart from iconic and deictic gestures, gesture research often identifies the additional classes *metaphoric*, *cohesives* and *beats* as defined by McNeill [88]. A **metaphoric gesture** is a figurative gesture that should not be taken literally, yet it communicates a truth about the situation, as for example showing a trajectory with the hand when talking about a trip. A **cohesive gesture** binds two temporally distant but related parts of a narrative. For example, a storyteller might make the same hand movement whenever the same character appears. **Beats** mark a narrative time. For example, a rhythmic arm movement can indicate time passing. The classification of gestures is often based on the tripartite concept of Peirce’s semiosis (i.e. the production or formation of meaning in a process that involves signs). In this semiosis, the relation between sign form and the signified object is related through similarity, contiguity and conventionality [89].

In the context of storytelling, verbal and signed storytelling can both use symbols (words and signs respectively) to tell the story. While the latter features more iconic signs than the former, gestures are present in each kind of communication. Moreover, some gestures are strictly dependent on the linguistic expression (word or sign) they co-occur with [90]. The distinction between arbitrary movement and iconic movement is of special importance, when investigating the meaning making process through body movement in storytelling.
Figure 6: American Sign Language (ASL) sign that can only be understood by ASL signers. The meaning of the sign cannot be inferred by the movement, because it is an arbitrary sign, a symbol in semiotic terms.

Figure 7: Sign used by non-signer and signers alike. This movement of the hand to the mouth has iconicity and can generally be well understood, because the movement directly relates to the described act.

performances. This distinction is especially apparent between Sign language and gestures as exemplified in Figures 6 and 7. Figure 6 depicts how STORY can be signed in ASL. For this sign, the signer closes both of their hands at their fingertips in front of the lower chest. The hands then circle outwards, opening slightly before circling back in front of the chest to repeat this movement twice. The signifier (movement of hands) does not resemble the signified (STORY). The meaning of the moving hands cannot be inferred by non-ASL signers, because the sign has an arbitrary connection to STORY. It should be noted that the same sign in Fig. 6 can also signify STORYTELLING, since STORY and STORYTELLING employ the same base sign. The difference is conveyed via non manual markers. The sign for STORY is signed more quickly, while STORYTELLING is signed more slowly and expansively.

In contrast, Figure 7 shows an iconic gesture, which also happens to be an iconic sign in ASL. Here, the hand is closed at its fingertips and raised to the mouth in repeated, short movements. This gesture can be understood by most as EAT or EATING without knowledge about ASL. This gesture is iconic, because the movement shares sufficient similarity with the physical act of eating.
At the same time, there are many seemingly arbitrary gestures that co-occur with speech acts, which highlights another key difference between signing and gesturing. The arbitrary signs in Sign language denote meaningful units that must be learned in order to be understood. In unsigned conversation, gestures may seem arbitrary while actually arising from deeper conceptual thinking \[16, 91, 92, 93\]. These conceptual structures, as presented in Section 1.2.3, become apparent in gestures through image schemas. Here, two studies are of particular relevance. First, Cienki \[80\] provides empirical evidence from video-recorded conversations in which participants show a high agreement in the use of image schemas to categorize gestures. Cienki argues that gestures provide additional schematic information "in terms of which to interpret the entity, action, or relation being talked about" \[80\] (p. 438). Second, Mittelberg \[94\] presents evidence from motion-capture data that illustrates how image schemas "operate in - and also guide the interpretation of - gestural (inter-)action" \[94\].

When using manual signs to tell a story, movements often become symbolic in the sense that the embodied sign must be learned in order to be understood (e.g. as in ASL). Signed and Spoken languages make use of gestures that suggest meanings in ways that are both iconic and arbitrary, though even the latter can build on image schematic conceptual structures. This thesis does not concern itself with the translation of text-to-sign/sign-to-text \[95\] or text-to-speech, but seeks to fill a gap in the literature on robotic performances, by exploring embodiment that can be pantomimic (that is, using iconic gestures) or based on spatial schematics. Studies in robotic performances have previously regarded gestures as auxiliary additions. But these studies focused on pantomimic or deictic gestures and did not concern themselves with the implicit, deeper meaning making process that can emerge from less iconic, schematic gestures. Hence, this thesis conducts empirical studies to contrast the use of both, to show how their combination can be of benefit in embodied storytelling performances.

1.3.3 Embodiment with Nao

With the focus of embodiment on the connection between the linguistics of storytelling and the body in movement (Section 1.3.1), an appropriate robot must be chosen to support this focus. The optimal robot for our purposes must be sufficiently embodied to enable body movement, gesture and speech. 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 evokes excitement, curiosity and amazement. But once robots become part of a system and we encounter them on a daily basis, habituation naturally results \[96\]. The
enactment of a gesture by a robot might not appear as exciting if it is enacted by a human, but this novelty will likely wear off after a few weeks. In a study by Kanda et al. [97], 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” [97]. Robots often bring a cuteness factor to a performance, 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. This so-called valley [98] is a gulf separating a cartoon-like robot such as the Nao (Figure 8), 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.

We need our robot to be sufficiently autonomous, such that the storytelling procedure is executed without continuous human interaction. Yet we also want the robot to communicate with, interact with, and learn information from people. Breazeal [99] refers to this class of robots as social robots. Many different social robots are available to choose from and they vary in the properties they bring to a performance, e.g. mobility, facial expression, degrees of freedom, height and more.
With recent trends in smart homes and smart speakers, social robots such as Amazon Alexa or Google Nest have become widely available (for a sketch of Alexa, see Figure 8). Their high-quality speech production and speech recognition provides them with two very practical storyteller properties. With no limbs, a can-shaped appearance and only an LED to signal discourse markers, these robots cannot be considered embodied in our terms (Sec. 1.3.1). Instead, we consider them a kind of non-embodied storyteller (Sec. 4.4).

A robot with more obvious embodiment was chosen as a storyteller by Striepe and Lugrin [10]. They chose the Reeti robot, a 40cm small, static robot with human-like facial expression capabilities (for a sketch of Reeti, see Figure 8). Its cartoon-like appearance avoids the uncanny valley. The robot has no limbs attached to its white, gender neutral body. Although, the authors provide various studies showing the effectiveness of Reeti in a storytelling context [10, 100], this robot is not feasible for our research, because it cannot move in space or use gestures.

With two legs, two arms and 25 degrees of freedom in its movements, the Nao robot from Aldebaran/SoftBank [101] is a polished, ready-to-use anthropomorphic social robot that stands about 58cm high. With LEDs for eyes and an immobile mouth, the robot lacks facial expressiveness, but it compensates for this with robust, bipedal walking capabilities. Nao can move its arms, legs and head in various directions. The hands have three fingers that can only open or close. The torso cannot rotate, hence a rotation can only be achieved through a circular walk. All programmable and movable joints are depicted in Figure 9.

Its movement capabilities are sufficiently advanced that Nao is regularly used at the Robot Soccer World Cup (RoboCup²). Nao’s speech recognition and production capabilities, its humanoid form and overall friendly appearance has convinced previous researchers to use Nao as a storyteller [4, 11, 12, 13]. Moreover, the Nao has been used in a variety of other research studies (for an overview of research on Nao as a socially assistive robot, see [102]), which makes our Nao-based results regarding movement and gestures useful and reproducible for

---

2 https://www.robocup.org/
There are certain limitation to Nao as a storyteller, which we discuss in Sec. 4.7. Nonetheless, we chose the Nao because of its gestural capabilities, which are the perfect fit for our practical implementation and experiments. These studies focus on the different movement types in embodied performances with robots, primarily in storytelling. Figure 10 shows the two Nao robots used in our experiments performing different gestures. The robot on the left is using one arm for a pointing gesture, whilst opening up the hand and dropping the other arm. The context of the story makes the opening of the hand meaningful, whereas the pointing refers to the other robot regardless of its narrative purpose. The robot on the right is pulling both arms in front of its body with its head lowered to its chest. The underlying conceptual schemas at work are suggestive of a defensive stance. All movement trajectories are highlighted using bold arrows to indicate the schematics at play.

1.3.4 Movement in Space

Embodied performances can not only alter the meaning of gestures, they can also use a specific stage to alter the perception of space. For example, acting on a stage and acting in front of a camera deliver a completely different viewpoint to the audience. Even a stage can provide different modes of observation, e.g. a theater in the round or a traverse can create different viewpoints. In the context of theater, the space defined by the stage can radically alter the way the audience perceives the performance and attributes meaning to movement. Some modern performances contextualise the space by removing the division between audience and actors. The performance leaves it to the audience to segregate the space in meaningful ways. Eventually, this provokes spectators to creatively undertake new interpretations of the script [103]. In this thesis, the notion of space refers to move-
ment in embodied performances, i.e. locomotive movement and the gestures of embodied performers. In general, these movements can be found in any kind of performance and on any kind of stage. The stage might depend on the type of performance or vice versa, but the movement studies presented here are essential for any form of embodied performance and are consequently a focal point of this thesis.

Locomotive spatial movement (hereafter just spatial movement) has very different properties to gestural movements of a body rooted to the spot. Spatial movement is not *iconic* and is much less diverse, yet it can convey an equally important amount of meaning in performances. Gestures can have different classes (e.g. *Iconic*, *Beats*, *Deictic* etc. [88]) with various different forms of iconicity. In contrast, locomotive spatial movement, in its basic conception, has only a few parameters such as direction, velocity and rotation. Moreover, the meaning of an iconic gesture has a contextual and a cultural dependency, where spatial movement has distinct properties of proximity and an (image) schematicity with fewer dependencies on context or culture. Yet, an embodied performance includes mixtures of movements and gestures, not all of which carry meaning. For example, some gestures are mere movements that indicate word search and help the speaker to communicate rather than the listener to understand. In contrast, moving towards someone while gesturing can carry very different meaning than gesturing while moving away. For example, waving while approaching is regarded as *welcoming*, whereas waving while departing is regarded as *saying goodbye*. Similarly, pointing gestures can drastically change their meaning if they are combined with rotational body movement, since both types of movement are referential.

All of these considerations, including the definitions and properties of gestures and locomotive spatial movement, are outlined and discussed in this thesis. As such, the work contributes to multiple areas of interdisciplinary research: not just automated storytelling (as built on automated story-generation) and embodied performance using robots, but the study of expressive gestures and physical meaning-making more generally, across a diversity of settings. The taxonomy and terminology that will be defined aims to foster further interdisciplinary research in the areas that contribute to them.

An automated system that aims to evoke an idea through a story can do so through the various layers presented here. A commonality of these layers is their shared image schematic basis. Research on automated storytelling has mostly been done at the layer of symbols and words [50, 104, 105]. But storytelling is more than the presentation of text and words. Even so, texts are an important window into language in the fields of automated story-generation and computational creativity. The latter typically sees story-generation as a process of text generation [49, 105, 106, 107, 108, 109], but what is
often overlooked in this view is a quality that cannot be found in the text itself, and which only emerges when the text is brought to life through performance. An engaging story-teller uses not just speech, but gesture, space (both movement and orientation within it) and timing to enhance the drama of the tale. Real story-tellers show as well as tell. They use their bodies to tell the tale, and machines should do likewise. Robots can make use of a variety of different modalities, each of which has been studied in different contexts: gaze [110, 111], facial expression [112, 113], voice [114], gesture [4, 13], movement [81] or even dance [27, 28]. Combinations of different modalities have also been investigated; for example with embodied performances that focused on speech, embodiment and agents [112, 115], gestures and movements in automated storytelling [4, 11, 116] and theater/improvisation [25, 26]. Different kinds of performances will vary in their emphases on different layers, while in the context of creativity studies embodiment is receiving more scientific attention [117, 118].

### 1.3.5 Embodied Creativity

The computational storytelling, story-generation and aspects of co-creative human-robot interaction presented in this work focus on embodiment, an aspect of computational creativity (CC) that is often overlooked or unspecified. This is partially due to the ambiguity of the term “embodiment” as explained in Section 1.3.1 and in part because the term “creativity” can also be defined in many different ways. On the entire continuum of CC research [119] there are more or less autonomously creative systems [120], non-embodied systems [35, 50] and inherently embodied systems [1, 121], all of which are used to understand, model or simulate the creativity of living beings [122]. Most of these attempts adopt a definition of creativity as an act or process which produces something of novelty and value [122]. Notably, this definition does not necessitate any form of embodiment. Meanwhile, other researchers have argued that an autonomous creative system can only be truly creative if it can fully interact within its creative environment [123] or has some form of perception of it [17]. Guckelsberger et al. [118] strongly underline the necessity for embodiment in creativity research. They argue that an agent’s physical capabilities (e.g. sensors, shape and means of movement) provide its potential for interaction with its environment, which structures its creativity. Guckelsberger et al. [118] reason that a creative act is necessarily situated, i.e. a creative act always arises in a social, cultural, historical and personal context. Furthermore, those authors argue that creative systems which rely on social interaction are even more dependent on some form of embodiment. Those systems must provide a plurality of perspectives through their agents in order to construct meaningful artifacts. Consequently, creativity emerges through the in-
teraction of these agents in and with their environment.

The present thesis embeds itself in the field of CC through various additions in different ways: First, the perception of authenticity in CC systems is increased by using robotic gestures which provide a storytelling system with some form of physical grounding. As such, it is comparable to the Marimba playing robot Shimom [121], whose creators aim to ground musical creation in physical movement. Second, this increased authenticity is an important factor for CC systems, and is represented here by an implementation of a question-guided story-generation system [125]. Third, the notion of movement and its implications for creativity are presented [2, 126]. This relates to CC research by Gemeinboeck and Saunders [127], who argue that Movement Matters in the meaning making process of co-creativity. Those authors present a performative approach in which humans and machines collaboratively explore a creative space of movements. They argue that this collaborative connection is knowledge-generating and enriches the social encounter between creative machine and human. To this extent, an interactive version of an embodied storytelling system is presented in this thesis, which focuses on the emotional and non-verbal aspects of a social encounter [21].

1.4 Research Questions

Symbols, movement, embodiment and gestures: each of those layers can connect a story with the audience. Most have been studied in their respective fields, but considering automated performances, how do these layers interact? The literature suggests questions about the individual contributions of the different modalities, but how do they compare to each other? Where do these layers overlap and how do they support each other? Answers to those questions are provided by this thesis. The study of linguistics and gestures acknowledges the role of cognitive, spatial schemas where the literature on automated storytelling largely neglects them, and this thesis will fill the gaps. The filler is one part literature review, one part theoretical exploration and one part empirical study that investigates the research questions to follow. The greatest gap in automated storytelling is the connection with embodied performances (gestures and space), which constitutes the core of this thesis.

First, a literature review of past and present work in the areas of automated storytelling and enactment will structure the discussion of the layers of performance. Second, this thesis expands automated storytelling from mere text-generation to the physical realm, and closes the gap between individual research that separately investigates these modalities. Third, this bridge is built by connecting gestures with
symbols, the most concrete units of meaning with the most abstract. Fourth, the spatial schemas that underpin embodied storytelling will provide a rigorous framework for embodied performances. Consequently, this thesis answers the following research questions:

- **RQ1**: What layers of performance can be combined for an automated performance?
- **RQ2**: What can embodiment add to a performance?
  - **RQ2.1**: What role do gestures play in performances?
  - **RQ2.2**: What role does the use of space play in performances?
- **RQ3**: What connects the layers of performance?
  - **RQ3.1**: How do spatial schemas support performances?
- **RQ4**: How do these layers interact with each other and the audience?
  - **RQ4.1**: How can image schemas help to connect performers with the audience?

The first question will be answered through the various empirical and theoretical results of this thesis. The layers of performance presented in Fig. 1 will be defined and examined through reflections on related works, which ultimately justify their inclusion as modular building blocks in storytelling performances. Specifically, the use of gestures and spatial locomotion are the missing pieces in the literature that this thesis contributes. Hence, **RQ2.1** and **RQ2.2** target the use of gestures and spatial movements in embodied performances, respectively. To answer **RQ2.1** and **RQ2.2**, a taxonomy of gestures and space for robotic performances is introduced. Here, we adopt the stance that gestures are more than mere epiphenomena of speech, but meaningful tools for communication, that in combination with spatial movement, provide a core component of embodied performances. Naturally, this leads to **RQ3**, which concerns the connection between the layers of performance. This connects the aforementioned image schematic nature of gestures with space, symbols, narrative and embodiment, thus raising **RQ3.1**. The third research question will be answered through a series of empirical studies. Those studies will show the usefulness of schematic mechanisms in embodied storytelling and provide evidence that the schematic use of space is appreciated by audiences over purely non-schematic, dramatic movement. Lastly, **RQ4** concerns how gestures and the schematic use of space can support interaction with the audience by introducing a source of emotion into embodied performances.
1.5 Structure and Contribution

Answers to each of these research questions can only be found at the intersection of multiple research domains. This makes the presented studies and theories highly interdisciplinary since they draw from multiple disciplines, e.g. gesture studies, automated storytelling, robotics and cognitive linguistics. The answers presented in this thesis provide useful insights for the same variety of fields. This relevance is partially reflected in the diversity of venues this research has been presented at, e.g. the International Conference on Computational Creativity (ICCC), the International Conference on Human-Robot Interaction (HRI) and a workshop on "Embodied AI - Exbodied Mind (Gesture-Perception-Robotics)".

In order to answer these interdisciplinary research questions, a framework is implemented that supports a modular approach to embodied storytelling. This is modular in the sense that the layers of performance (e.g. agents, movements, gestures) can be reorganized (added, modified or deleted) independently of other layers in the framework. The resulting framework is implemented as an AI system, called Scéalability. This system is modular and extensible, and offers a degree of explainability for its decisions. Through its underlying symbolic story-generation system, called Scéalextric [54], it is able to transform a purely textual narrative into a physical performance with embodied and non-embodied actors. The framework has been used to implement different storytelling modes in order to assess the practicability of interactive, embodied storytelling [125, 128]. It has also been used for a series of empirical studies assessing audience appreciation of an embodied storytelling performance under different conditions [2]. The modular nature of Scéalability allows these conditions to highlight the different layers of performance in order to address the central research questions.

It is a non-trivial task to quantify the contribution of embodiment to a performance. Again, the answer to RQ2 has both theoretical and empirical constituents. In a Double Act [1] we present embodied and non-embodied actors to audiences in order to explore RQ2 and show that embodied actors are more appreciated than their non-embodied counterparts. RQ3 is addressed by studies contrasting locomotive spatial movement with iconic gestures, which show that audiences are sensitive to the coherent use of space in embodied storytelling, and appreciate the schematic use of spatial movements as much as more culturally-specific pantomime gestures [2]. While most of the empirical studies have been conducted using crowd-sourced ratings of videos of the performances, some studies include evaluations from live performances on campus. Lastly, RQ4 is answered with the pre-
sentation of a system which packages insights from the gestural and spatial domain and links schematic intuitions with interactive elements. In this system, the user can interact and change the course of a story using non-verbal cues (i.e. gestures and facial expressions) in order to influence the story being enacted by the robotic actors. At this point, RQ1 can be adequately addressed, since the element of emotion is used to combine all the different layers of performance into one automated storytelling performance. In this combination, interactive storytelling is enabled through the use of gestures and movements by both actor and audience (Fig. 1). Even on the level of symbols and words, this system can include the element of emotion as a trigger for metaphoric exaggeration and irony [129].

The narrative of this thesis has begun with this introduction. Chapter 2 is dedicated to the Background Research. This literature review will present the different layers of performance with an emphasis on robotic storytelling, gestures and space. Since the origin of the performance is an automatically generated story, the review begins with a brief summary of automated storytelling systems, before branching into the different modalities of robotic storytelling. Additionally, it will survey research on gesture and space to provide the background relevant for the following Chapters. Collecting all of the pieces from previous and contemporary research will eventually bring to light what pieces are missing, thus motivating our research questions. And those questions will be addressed both theoretically and empirically.

In Chapter 3, the different types and properties of gestures are discussed with regard to how they are meaningful for embodied performances. The role of space in embodied performances is then discussed. Specifically, the image schematic underpinnings of spatial movement are highlighted as meaning bearing tools for storytelling. These theoretical considerations conclude with a taxonomy that allows us to understand how gestures and space can be productively combined in ways that respect their distinctive properties and synergies.

Chapter 4 outlines the practical implementation of the work. This includes a description of the additions and improvements made to the story-generation system Scéalextric [54], and a description of the Scéalability system and the various technical implementations necessary for a robotic storytelling performance with gestures, space and emotion. Since the challenge of combining and evaluating these different modalities is also a technical one, this chapter provides a detailed account of how those modalities can be implemented with embodied actors.

Chapter 5 then presents the empirical studies. These studies first focus on embodiment vs. non-embodiment in order to respond to RQ2. Subsequent experiments focus on Space & Gesture in order to respond
to RQ3 & RQ4. Finally, the thesis concludes with a discussion of the research questions and a preview of the future work this thesis has inspired.

The contribution of this thesis can be summarized as follows:

• It provides an interdisciplinary overview of works related to computational storytelling and embodied performances as well as movement in space and gestures.

• It describes a conceptual framework for embodied performance with robotic actors and its implications for the field.

• It outlines the implementation of various multi-agent, multi-modal storytelling systems.

• It conducts an empirical evaluation of different movement strategies, which inform the field of embodied performances about effective schematic movement.
2.1 STORYTELLING

The interdisciplinary nature of robotic storytelling can draw upon a plethora of related works. Since this work originates from the field of automated storytelling (with and without embodiment) its related work will be our starting point and we define some terminology.

Coming up with a good story and telling this story in an entertaining way are two different tasks that require very different skills, e.g. a good writer is not automatically a good narrator or actor. The same difference can be observed when we look at the domain of computational story generation. Story writing systems are not always equatable with storytellers. The former generates the fabula ("what is told") and the latter also requires the discourse ("how it is told") [52]. This distinction is important, because storytelling is a performative act, whereas story generation is not. Story generators may or may not provide the deep structures of the fabula, since they can also just provide the surface form for the discourse. A storyteller needs access to both: The deep structures allow a storyteller to span a complex performance with mixed modalities of space and gesture, whereas a surface form merely provides the script (i.e. the written text of a play or performance) to narrate the story. The notion of a script is therefore ambiguous. It can either refer to a list of instructions in a computational sense or it can refer to the written text of a play, which contains the dialogue (in direct speech), stage directions and instructions for the actors. The latter is a script in the manuscript sense and is the notion we employ here, unless otherwise noted.

Let us look at this distinction more closely: a story generator can follow its list of instructions and rules in order to create a plot. A plot is a sequence of events where each causes the next one [130]. On a plot-level (the deeper structure of the fabula), a characters’ fate is planned and constructed. Here, the fabula lays out the predetermination of each character [51]. If the story generator also generates a surface form, this means that the system creates the story as a narrative, i.e. telling of the plot through verbal, visual or written means. As a result, the actor might now access fabula and discourse for its performance.

The access to deep structure and surface form is important in the discussion of story generators, because the performance of the storyteller depends on this access. The storytelling performance we are trying to model requires the actors to turn a plot into physical ac-
tions. More precisely, spatial and temporal relations between characters, objects, causal relations and more, need to be expressed through physical action. Consequently, the literature on computational story-generators will be reviewed with a particular focus on whether or not they generate both *fabula* and *discourse* components in an accessible form.

2.1.1 Computational Story-Generation

2.1.1.1 Knowledge-Based Approaches

Computational story-generation has a long provenance as a branch of AI research. While it is still somewhat unclear what the very first generated story was, it is suggested that the first story generator was actually just a tool for linguistics in the field \[131\]. The linguist Joseph E. Grimes developed a basic grammar-based story generator in the early 1960s \[132\] in order to generate stories and consequently collect the reactions of listeners. Ultimately, Grimes was trying to investigate linguistic observations on a number of generated stories, for different native speakers. His story generator incorporated some of Propp’s *Morphology of the Folktale* \[133\], but despite using Propp’s folktale elements, Grimes considered the output to be too dull to continue with this approach.

Nonetheless, Propp’s morphology remained influential until modern times. For example, the *PropperWryter* system of Gervás \[134\], uses Propp’s morphology as a grammar for generating stories. Almost a century after Propp’s *Morphology of the Folktale* had been published, Gervás \[135\] reviews this methodology in the light of modern computational creativity and suggests a number of possible computational implementations. Propp analyzed many Russian folk tales and identified their recurring themes and structural elements. Among those elements, he identified a set of roles for characters in the narrative (dramatis personae) and a set of character functions. The former includes roles such as the hero or the villain. The set of character functions define the hero’s journey by providing elements that detail the involvement of the villain, a resolution, elements that describe the dispatching of the hero and more. Here, Gervás \[134\] explains *trebling*, the threefold repetition of character functions, with the example of *Snow White’s* stepmother trying to kill her in three different ways, only for the last one to succeed. Character functions are linked to one another by long-range dependencies \[136\]. For example, if the hero’s friend has been cursed at the beginning of the tale, it is the friend that needs to be freed from their spell at the end.

Gervás \[134\] highlights how Propp already addresses long range dependencies as one of the biggest obstacles for story generation. As a possible response, Gervás suggests a *plot driver generator* which pro-
vides an algorithmic procedure to generate valid sequences of character functions. The plot driver then is passed to a fabula generator, which maps story actions for the provided character functions. The mapping is based on pre- and postconditions of the story actions, which can carry over to the rest of the story. This mechanism enables the system to model long range dependencies between the character functions. On the discourse level, Gervás [137] shows how the vocabulary of his PropperWryter system can be expanded in order to “dress up” the system with fictional renderings, e.g. the basic description of moves on a chessboard becomes the story of kings and queens in a battle. This allows PropperWryter to produce fabula and discourse.

Another early story-generation system is the Novel Writer, published by Klein et al. [30] in 1973. The Novel Writer was the first system able to generate murder-stories. The roles of murderer and victim are inserted by the user and pre-programmed scripts take the input to spin a murder-mystery. The mystery theme is set, but the rules of the underlying simulation are stochastic and capable of random outputs. The authors encode their semantic deep structure as a directed graph with nodes as objects (e.g. man, window, hammer) and with edges as their relations (e.g. break, with, in). Thus, each entry in the network is a semantic triplet (e.g. man-break-window) connecting nodes with edges. The resulting network allows the system to generate stories through a set of grammar rules and a surface-level rendering of the semantic units in English. Like Novel Writer, the story generator in this current work (Sec. 2.1.2) is also based on semantic triplets. The Novel Writer relied on a set of rules and some simulation, but lacked a model of the characters’ or writer’s intentions and this has become the focal point of subsequent story generators.

Three years after Novel Writer, the TALE-SPIN system was able to generate a fictional world of inter-related characters [49]. In this problem-solving system, it is the protagonist of the simulation who needs to solve a plot-driving problem. The narrative is composed of the events which occur during the problem-solving. Hence, the internal mechanism plans how the protagonist can overcome the problem, whereas the surface structure tells the protagonist’s story as it does so. The system’s deep structures are based on Conceptual Dependency (CD) representations [59], which provides a canonical knowledge representation based on a few semantic primitives. TALE-SPIN represents planning structures in CD as cause-and-effect chains. Those planning structures include the goal-state that the character can achieve to solve its problem. An example structure looks like this:

dPROX (PTRANS), dCONTROL (ATRANS), dKNOW (MTRANS).

with dPROX(X,Y) - X wishes to be near Y
Planbox 0: if X is already near y, succeed.

adopted from Meehan [49].
The terms PTRANS and ATRANS are CD representations for physical transfer and transfer of abstract relation respectively. Due to its CD representations, the system’s deeper conceptual structures rely on only a few schemas, which - as this thesis postulates - is a property that can be exploited on all layers of a storytelling performance. On the resulting discourse level, the system creates fables with a given moral, e.g. “Once upon a time there was a dishonest fox and a vain crow” and the moral is “never trust flatterers”[49]. The surface level even includes some dialogue pieces for the characters, but no dedicated manuscript of how the story or dialogue should be performed.

Dehn [106] summarizes three important simulation levels of TALE-SPIN as character intentionality, social constraints and physical causality. In contrast, Dehn’s 1981 AUTHOR system proposed a model of the writer’s mind as it incorporated authorial goals and their constraints on the story [106]. She argues that a system, like TALE-SPIN, which is based on character intentionality, will generate stories that “go well” for the character. In contrast, a story generator that is based on the writer’s intentionality will focus on writing a good story. AUTHOR does not create a surface form in the sense that a story is produced, but the system creates authorial episodes that can be fleshed out into textual narratives, e.g. “Shy person (protagonist) in need of employment gets job as salesman but is incapable of facing $DOOR-TO-DOOR”[106]. The variable $DOOR-TO-DOOR can be extended with a specific instance of door-to-door activity. The discourse of the system is a network of author goals, which serve as a skeleton for the story. As such, Dehn’s system does not provide a surface form useful for a storytelling performance, but a deep structure with a focus on author goals.

These systems chiefly differ with respect to the domain in which they tell their stories, the way the textual story is being constructed or the amount of influence that a user of the system can exert. For instance, Lebowitz’s 1985 system UNIVERSE generated scripts for soap-operas [138], while Turner’s 1993 MINSTREL system generated plots in the domain of courtly knights [139]. The MINSTREL system, like Dehn’s AUTHOR, considers author intentionality with author-schemas and deals with character intentions as manifest in story events via character-schemas. Examples of an author-schema include “the goal of developing a theme, and the goal of including suspense in a story” [140]. MINSTREL uses a set of Transformer-Recall-Adapt Methods to guide the generator’s choices. The first of these methods is a standard-problem-solver that retrieves schemas from memory in order to build a scene for the story and lay out the deeper conceptual structures for it. If this solver fails, other methods take over and perform small adjustments to the specifications of the schema until the standard-problem-solver is successful. This problem solving stage reuses knowledge from pre-
vious stories. The goals that dictate the fabula in MINSTREL are de-
defined on a character and author level. With this implementation, MIN-
STREL is the first computational storyteller to explicitly model the
creative writing process [140]. On the discourse level, the surface form
presents stories about King Arthur and the Knights of the Round Ta-
ble.

In a similar vein, the MEXICA system by Pérez y Pérez and Sharples
[109] generates stories about the Aztecs of ancient Mexico. The deeper
structures of MEXICA are constructed with three major mechanisms.
First, MEXICA employs a story grammar, which is defined by Man-
dler [141] as “a rule system devised for the purpose of describing the reg-
ularities found in one kind of text. The rules describe the units of which
stories are composed, that is, their constituent structure, and the ordering of
the units, [...]”[141](p.18).

Pérez y Pérez and Sharples place a strong focus on the study of
the creative process. MEXICA uses the E-R model of engagement and
reflection, which first engages with a developing story by adding new
elements to it, before the reflection considers the impact of the engage-
ment. MEXICA employs “emotional links” [140] between characters
that evolve with story actions. For example, Jaguar knight might fall in
love with Princess. If this emotional link is established, Jaguar knight
and Princess can later become a couple.

Working with story grammars, the E-R model and emotional links,
MEXICA provides extensive deep structures. In contrast, however, the
resulting surface form on the level of discourse is rather limited. A
fully rendered story is rather repetitive, prosaic and the deep struc-
ture becomes visible. Pérez y Pérez [142] provides the following ex-
cerpt:

\[
\text{Jaguar Knight was an inhabitant of the great Tenochtitlan. Princess was an inhabitant of the great Tenochtitlan. A bad spirit took Jaguar Knight’s soul provoking Jaguar Knight to become intensely jealous of Princess. Jaguar Knight tried to scare Princess by pretending that Jaguar Knight wanted to kill Princess with a lance. But instead, Jaguar Knight stumbled and wounded itself. [142]}
\]

The deep structure of A − action − B, B − action − C, C − action −
D can be observed in the surface form of the story. Moreover, the sto-
ries generated by MEXICA typically follow this strict Jaguar knight
and Princess story arc.

Ideally, the generated surface text, the generated story, will be suffi-
ciently distant from the generator’s internal representations. To achieve
such creative distance is a main goal of the BRUTUS system of Bringsjord
and Ferrucci [143], which creates short stories from a set of pre-defined
topics. This rule-based approach features three main procedures: First,
BRUTUS uses a thematic frame to encompass the characters, their goals and events. Second, the system uses a simulation process to develop the plot. Lastly, a story grammar is used to render the plot as a surface text. BRUTUS provides both *fabula* and *discourse*. The system exploits a large knowledge base of literary forms to create a rich surface form, as exemplified by the following example:

*Dave Striver loved the university. He loved its ivy-covered clock-towers, its ancient and sturdy brick, and its sun-splashed verdant greens and eager youth. He also loved the fact that the university is free of the stark unforgiving trials of the business world — only this isn’t a fact: academia has its own tests, and some are as merciless as any in the marketplace.* [143](p.5)

In this example, the surface form has a greater variety of words to hide its deeper structures than are evident in MEXICA’s stories. The surface form is detailed, novelistic and sometimes features lines of dialogue for its characters. Hence, it achieves the intended creative distance. Nonetheless, Bringsjord and Ferrucci [143] assert that their system does not use any creative mechanism to spin its stories, which instead originates from *reverse engineering*. By that, the authors mean that they read and formalize stories in order to formalize a system to write these stories for itself. In other words, the authors have formalized “a generative capacity sufficient to produce this and other stories”[143](p.6) and this formalization is at the core of BRUTUS’ generation process.

Riedl and Young [144] propose that stories can be crafted around intent-driven plans, where those plans can then unify overarching character goals. This work uses planning techniques in combination with techniques from the Believe-Desire-Intention (BDI) model [145] to tackle the problem of long range dependencies. Their implemented system, called *Fabulist*, plans actions for many characters by modeling causality and intentionalitity. To ensure that characters’ intentions appear on the surface level of the story, the system inserts character goals into the narrative plan at the deeper structural levels. As a consequence, long range dependencies are preserved over the course of the story. In [146], the authors provide empirical evidence for the effectiveness of their approach, in which they show increased audience comprehension of character intentions over a baseline condition with a conventional planning algorithm. The surface form of *Fabulist* is rather shallow and gives insights into the formulaic structure produced by its internal mechanisms. Here is an example of a *Fabulist* story:

*There is a woman named Jasmine. There is a king named Mamoud. This is a story about how King Mamoud becomes married to Jasmine. There is a magic genie. This is also a story about how the*
genie dies. There is a magic lamp. There is a dragon. The dragon has the magic lamp. The genie is confined within the magic lamp.

[147](p.20, Table 4)

A story like this is far away from a performative script that could provide instructions and dialogue for narrators and actors. In fact, this output is closer to the underlying plot derived by the internal structures of Fabulist than it is to a novelistic story. In this sense, Fabulist is more of a plot generator than a story generator.

Another notable example is the Curveship system by Montfort [148]. Curveship is not actually a story generator. Rather, the system takes the representation of a story and then works to manipulate it. Its emphasis is on the perspective and performance of the story being told, which is somewhat similar to the implementations in Chapter 4. As such, Curveship constructs interactive fiction with a modular approach, that allows it to pinpoint distinct responsibilities and dependencies. For example, Curveship has a language-independent simulator module that updates its world models module. This module is specific to one literary work and sends events to the narrator module (which is specific to a language) and then sends the resulting narration to the presenter module. This allows Curveship to fine-tune the impact and effects of each individual module. In this way, Curveship enables a presenter to modulate the rendering of the chosen narrative viewpoint.

Curveship manages world models of characters, locations, events and more, and it can also manipulate the narration and description of world models. All that is needed is a story generator that provides the stories for Curveship to modulate. A marriage between Curveship and MEXICA has been described by Montfort et al. [107]. In the resulting system Slant, the generated story is treated as an artifact that can be rendered for the audience in different ways and views. We aim for some of the high-level modularity and control of Curveship in this current work.

2.1.1.2 Data-Driven Approaches

Each of these approaches is rule-based and symbolic. However, the field of modern AI is now largely dominated by Machine Learning (ML) techniques, specifically neural network approaches. In ML, a system is presented with a large amount of data (training examples) in order to create a statistical model of a phenomenon, which is then able to identify or generate new instances of the phenomenon. For story-generation, a ML system can be presented with a large database of stories in order to create novel stories of its own. In the past, a significant challenge for these systems was to learn long-range semantic and pragmatic dependencies, which would allow it to generate longer, coherent stories. Large language models, such as BERT
[149] and GPT [3], can provide a solution to this problem. Their Transformer architecture allows us to train a model on long sequences of words with complex long-range dependencies. An attention mechanism [150] inside the multi-layered encoding structure of the Transformer allows the model to pick up subtle syntactical dependencies.

Additionally, this architecture learns faster than previous generative language models, and supports impressive one-shot and few-shot learning capabilities. Ultimately, these language models can be used for a variety of different tasks, e.g. question-answering, translation, summarizing and text-generation. For the latter, one can provide a start-word, start-sentence or start-paragraph and the language model can interactively suggest the next, most likely word, sentence or paragraph. In order to avoid generating the same stories on repeat, the model can either be fine-tuned (i.e. presented with a specific data-set to learn on top of what it has previously learned) or the selection of the most-likely prediction can include some randomness.

We conclude this section by linking past and present. In the quotation below, the first four lines (shown in blue) are taken from a story created by the Grimes story generator [131]. The rule-based structure of this short story is obvious from the repeated outputs of the system, which only change the types of characters and actions, and were considered “boring” by Grimes. However, these three lines are now passed as input to GPT-3, a transformer-based language model with 175 billion parameters [151]. Using an AI-writing assistant system built on GPT-3, lines five to thirteen below (shown in red) have been generated:

A lion has been in trouble for a long time. A dog steals something that belongs to the lion. The hero, lion, kills the villain, dog, without a fight. The hero, lion, thus is able to get his possession back.

This all takes place in the high country in Africa. Things get even worse for the lion when the "devil" comes. Bring your brain in to play. The devil is tricking the hero, the lion, and takes power away from the hero, lion, who can no longer use his power.

In the beginning of time, the lion had no natural power. The power was given to him, for good reasons, at the end of time, by the Lord God. Mankind’s power is not like this.

It is only a tool, made to serve mankind.

---

1 Training a machine learning algorithm with only one or a few training samples.
2 An example of this can be found at https://transformer.huggingface.co/.
3 The service platform shortlyai.com offers a free trial to generate a few writing examples based on user input. At the time of writing, GPT-3 is not publicly accessible.
The example from the language model is deliberately short, yet it is possible to create longer responses with a comparable amount of coherence. This coherence is reflected in the way that the system is successfully able to keep track of characters, their references and intents over the course of multiple paragraphs. In this instance, the system even mirrors the manner of referencing the hero as “lion” whilst not being bound to use this reference as strictly as Grime’s system.

We recall the distinction between fabula and discourse (Sec. 1.1) and note that Language Models (e.g. BERT, GPT) only create the latter. Their black-box nature bars the user from understanding the deep structure of how the story has been written and only presents what has been written - the discourse. Only if the language model is retrained with data that marks the deeper structures within the training data, can the system provide some semantic hooks in its generated artifacts. Additionally, dialogue for a performance might be generated by augmenting the training data with additional dialogue options. A fine-tuned version of GPT might then be able to produce some deeper structure along with the surface text that can include dialogue. So if we want the language model to produce stories that feature instructions for accompanying gestures, movement or dialogue, we would need to curate an appropriate data-set first. Conversely, we can instead build on a system that allows us to expand its deeper structure and surface form. This way, a performer can span a complex performance with mixed modalities of space and gesture.

2.1.2 The Scéalextric System

The story-generator Scéalextric [54], used for all of the generated stories in this work, is not a language model, but a knowledge-based system. Scéalextric is unlikely to compete with the qualitative impression that large language models can provide. However, using declarative knowledge structures, Scéalextric offers a degree of explainability for its choices, and symbolic hooks that allow a performer to span a complex performance with mixed modalities of space and gesture. This means, that Scéalextric provides direct access to the fabula, which is necessary to make semantic decisions in enactment and performance.

Scéalextric is inspired by Cook’s PLOTTO book [53] from 1928. Cook provided a large set of combinatorial plot elements, which underpins a moral coherence of their linkages. Following one plot element along a chosen set in linkages provides a full plot. Similarly, Veale [50] uses a set of action triples to generate stories with a moral. The schematic and symbolic underpinning of Scéalextric’s knowledge-base enables research beyond the layer of text and provides accessible conceptual and technical access points for meaningful enactment of its stories with gestures and space.
2.1.2.1 *Scéalextric’s Deep Structures*

*Scéalextric’s* rich database[^4] of symbolic representations allow actions to be bound together by causal connections and characters to be bound to actions on the basis of their established qualities. The system binds individual plot actions into plot segments (or arcs) with the following two-character triplet shape:

1. \( X \) *action* \( Y \)
2. \( Y \) *reaction* \( X \)
3. \( X \) *re-reaction* \( Y \)

Each action in the *Scéalextric* knowledge-base is causally connected with the next, and successive actions often shift the focus from character \( X \) to character \( Y \) and back again. The publicly available list of plot segments contains entries such as “\( X \) flatter \( Y \); \( Y \) promote \( X \); \( X \) disappoint \( Y \)”. The *Scéalextric* system provides over 3000 plot segments of this kind, made from causally-structured triplets that range over 800 different plot verbs [^54]. The resulting story space is modeled as a forest of trees in which each vertex is a plot verb, and in which every random walk results in a causally-coherent plot. Here is an example of a traversal through the forest:

\[
X \text{ learns from } Y \rightarrow X \text{ is inspired by } Y \rightarrow X \text{ falls in love with } Y \rightarrow X \text{ sleeps with } Y \rightarrow Y \text{ fails to impress } X \rightarrow X \text{ is disillusioned by } Y \rightarrow X \text{ breaks with } Y
\]

The roles of characters (\( X \) and \( Y \)) can be filled with different fictional or non-fictional characters. A study by Veale [^55] shows that generated stories are rated higher on dimensions such as entertainment, laughter and drama when they use characters from a list of well-known fictional and non-fictional characters. This Non-Official Characterization (NOC[^5]) list is a knowledge-base containing semantic triples about famous people, living and dead, fictional and real. Here as an example entry for the fictional character *Rick Sanchez*:

Character: Rick Sanchez  
Name: Rick Sanchez  
Gender: male  
Address 1, 2, 3: Old Town, Maine, USA  
Politics: libertarian  
Marital Status: widowed  
Opponent: Tammy Gueterman, Evil Morty, Zeep Xanflorp, Risotto Groupon  
Typical Activity: pioneering new technologies, inventing electrical marvels  
Vehicle of Choice: home-made rocket ship

[^4]: available here: https://github.com/prosecconetwork/Scealextric  
[^5]: available here: https://github.com/prosecconetwork/The-NOC-List
Weapon of Choice: home-made death ray
Clothing: white lab coat
Domains: Rick and Morty
Genres: science fiction, comedy fictional

The deeper structure of Scéalextric has many different access points for performative augmentation. The richest and most accessible semantic units of Scéalextric are its actions on the plot-level. Veale’s Flux Capacitor [152] system has been used to create opening and closing Bookends that anchor the character arc through apt stereotypical categories [153]. The knowledge-base provides 449 of those categories with examples such as the President category: the arc can start at election and end at being deposed. These actions can span over multiple links and follow the same symbolic representation (X action Y) found in the entire knowledge-base:

- **President**: are_elected_by, appoint, are_overthrown_by, are_toppled_by, are_usurped_by, rule_over, are_deposed_by
- **Novelist**: write_about, are_published_by, are_read_by
- **Scientist**: experiment_on, evaluate, are_published_by

The “are_” and “_by” pre- and post-fix markers signal the passive role of the subject for this action. For example, in “X elects Y” the protagonist role of X is active and in “X is elected by Y” the role of X is passive. Ultimately, the idiomatic rendering of Scéalextric provides a rich, extensible surface form. For example, consider at the idiomatic rendering for “elect” and “are_elected_by”:

- **elect**: X helped to elect Y to high office
- **are_elected_by**: Y helped X to be elected, Y voted for X multiple times

Lastly, the causal connections between plot actions are discourse linkages: “so”, “but” and “yet”. These linkages indicate an expected continuation (“so”) or can highlight an unexpected continuation (“but”, “yet”). Plots that only contain “so” will result in a linear narrative with no twists and turns, whereas a succession of “but” and “yet” linkages is likely to produce surprise.

### 2.1.2.2 Extensibility

**Adding Character(s)** Scéalextric stories are modelled around a simple two-character structure which follows a single, linear plot line from start to end, limiting the surface form of Scéalextric stories. Veale et al. [1] present a remedy for these issues.

First, the authors introduce greater flexibility into the story grammar by breaking open the rigid plot triples. Through “recursive descent from a single starting triplet”[1], more sub-plots can be woven into the
global plot structure and secondary characters can be introduced. A range of additional action triplets is added to Scéalextric’s knowledge-base to introduce recursive expansion points to the causal graph. At those expansion points, the system can add recursive subplots by inserting further triplets. The original Scéalextric generates a plot from left to right, the recursive variant can produce extensions in depth, from top to bottom, while preserving the overarching plot shape.

Second, these additional triplets can refer to additional temporary characters. For instance, if character X is fighting with a temporary character X-enemy or is kissing with a temporary character X-spouse, the system chooses an appropriate NOC character that can serve as an enemy to X or a lover to X respectively. For example, if Clark Kent is chosen as X, then X-enemy is chosen from the “Opponent” field in the NOC list to be e.g., Lex Luthor. Conversely, X-spouse will be chosen from the “Marital Status” field of X’s NOC entry, which suggests X-spouse = Lois Lane.

Adding Spatial Semantics Scéalextric’s knowledge-base allows us to add new semantic information where needed, by altering and extending its large collection of interlocking action triples and their logical connectives, as well as surface-level idiomatic renderings for each action. The actions in Scéalextric come with meaningful properties, relations and categories. Inserting new information, such as emotional valence or movement direction, is possible for any story action. For instance, we can map many of Scéalextric’s plot actions to appropriate spatial schemas, of the kind presented in more detail in Sec. 2.3.3. In short, positive actions tend to bring characters closer in the physical (and emotional) sense, while negative actions can lead to a greater physical (and emotional) distance. For example, the actions date, marry and hire are labelled with a positive valence, signalling a physical or emotional closeness. In contrast, the actions disgust, fear and fool are labelled with negative valence, signalling a physical and emotional distance. Overall, 62.87% of Scéalextric’s actions signal a physical closeness and 31.93% a physical distance. The remaining 5.20% do not suggest a clear spatial schema.

These labels provide more fine-grained distinctions into schematic classes such as UP/DOWN or NEAR/FAR with explanations of the actions in the context of the schema. For example, the action X condescends to Y can be metaphorically paraphrased as X looking down on Y. Thus, the action condescend to has the label DOWN. Conversely, the positive action (are inspired by) has the label UP, because if X is inspired by Y, X is looking up to Y. In the same vein, the label BACK is used for actions such as avoid or leave behind to signal backing off in the physical space. The most frequent label is UP/DOWN with 24% coverage of the actions, followed by IN/OUT with approximately 18%. From those schemas, we can choose how and when gestures
and movements are appropriate on the deeper conceptual level and the level of performance. The choice of schemas and further related definitions are discussed in Sec. 2.3.3.

2.1.2.3 Dialogue

Among the various systems surveyed here, some systems produce extensive deep structures, others produce vivid surface renderings, and some do produce both. A few even provide some form of dialogue within their surface form. If we want to model an embodied storytelling experience with actors on a stage, we need to consider the difference between the text of the story being told and the text of the story being enacted. When a storyteller speaks directly to the audience, we call this narration. But if we want artificial agents to theatrically enact the story, we need them to articulate speech acts between each other, we need a system that provides dialogue as well as narration. The narration should include the description of events or locations for a narrator to read out. Conversely, the dialogue includes everything that actors must directly say to each other.

Scéalextric’s Dialogue Module  The chief virtue of Scéalextric as a symbolic, knowledge-driven system is that it exposes its inner logic at every level of a story. An external client of the system can see past the surface text to the unambiguous structures within, to see e.g., the underlying fabula of actions and consequences. For any story action \( X < action-V_i > Y \), it can use \( V_i \) as a key with which to look up appropriate speech acts for the devices embodying \( X \) and \( Y \). A dialogue module, as described in [2], can be added to Scéalextric in order to assign Scéalextric text to the narrator \( N \), and assign associated speech acts to the agents playing \( X \) and \( Y \) to utter as they enact the narrated events. The following is typical of the resulting script:

\[
X=\text{The Joker}; \hspace{1em} Y=\text{Catwoman}; \hspace{1em} Y-enemy=\text{Batman}; \hspace{1em} N=\text{Narrator}
\]

\[
N: \text{What if The Joker fell in love with Catwoman?}
\]

Something clicked inside The Joker when sexy Catwoman came along.

\[
< X \text{ fall_in_love_with } Y >
\]

\[
X: \text{“You are as appealing as a monster truck.”} \hspace{1em} Y: \text{“Why thank you.”}
\]

\[
N: \text{So at first, it was said that Batman had never loved seductive Catwoman as much as Joker loved her now.}
\]

\[
< X \text{ study_under } Y >
\]

\[
X: \text{“Please take me as your student.”}
\]

\[
Y: \text{“I am happy to be your guide.”}
\]

\[
N: \text{So Joker became a student of Catwoman for stealing jewels.}
\]

... skipping ahead ...
N: Yet Batman kidnapped Catwoman and kept her hostage. Y: “For god’s sake, call the cops, Joker.”

N: So Joker followed the crumbs and tracked Bruce down. Y: “Don’t lose the scent whatever you do, Joker.”

N: But Joker was not the fool that Bruce imagined. X: “I fight best as the underdog, Catwoman.”

... skipping ahead ...

N: After this, Joker rescued Catwoman from an attack with a batarang by Batman. X: “It’s time we took our relationship to the next level.” Y: “You and me as one?”

N: So Catwoman went down on bended knee and proposed to Joker. Y: “I think we should get married.” X: “Two can live as cheaply as one.”

N: In the end, Joker married Catwoman in Arkham Asylum and they honeymooned in Gotham City. Thereafter Joker and Catwoman were utterly inseparable; wherever Catwoman went Joker was sure to follow.

The dialogue model is responsible for generating two types of speech acts. The first, exemplified by this narrative, captures what characters say as they perform a plot action. This kind is generated for any action that involves X and Y, since these roles are filled by embodied actors. The second kind concerns actions performed by ancillary characters, such as X-friend or Y-spouse. Since these roles are not embodied, there is no device to speak their parts. We see them only through the eyes, and the words, of X and Y. In this case, the dialogue model still generates speech acts for X and Y, in which they reflect on the actions of the unseen characters. For instance, if the action is < Y-spouse cheat_with X-lawyer >, the narrator will tell us as much, but X may also be given the line (spoken to Y): “Just wait until you get his bill.”

2.1.3 Embodied, Automated Storytelling

Generation of text is a key component of storytelling. Yet, long before computational text-generation was possible, telling a story had been transformed into embodied performances. Acting out a text on a stage goes back to the ancient Greek and Roman theaters. The school
of *Chironomia* was founded by ancient Greek and Roman orators in order to study the effective use of hands to supplement or even replace speech. *Chironomia* persisted until the 19th century with works such as [154] and [155]. For a modern, automated storyteller, the use of hands, gestures and space creates the need for embodiment. Likewise, using a virtual or physical embodiment to present stories is redundant when none of the modalities of embodiment (e.g. facial expressions, gestures, movement) are put to use. Consequently, most research with automated storytellers focuses on one or more of those embodied modalities. Those studies, which use either a virtual or a physical storyteller to investigate its embodied modalities, provide the following related works.

Most relevant is the approach of Pelachaud et al. [13] who design an expressive gesture model for a storytelling Nao robot [11, 82]. Their approach transforms and formalizes gestures previously used for a virtual avatar in a unified framework. This is achieved by rendering gestures in a Function Markup Language and a Behaviour Markup Language. The result is a reusable database of approximately 500 gestures with descriptive annotations. The authors present how a subset of these gestures can be used for a robot that reads stories to children. Le and Pelachaud [156] evaluate and confirm that these gestures are perceived as appropriate to their objectives, yet their findings indicate that the gestures score poorly for perceived naturalness. Ultimately, their approach tries to connect storytelling within a general framework that allows it to be applied for different types of robots, e.g. physical or virtual.

Augello et al. [157] present *NarRob*, a narrative robot. Their system unifies a database of 209 annotated tales with a set of gestures in an OWL ontology [158]. The relation between gesture and text is based on the automated analysis of emotional valence of the text. Their *NarRob* uses the annotated corpus of tales to associate gestures and emotional expressions to the telling of the tale. Although, the authors are able to have a completely automated approach that can easily incorporate new stories, their tales are not generated by the system itself and only the performance is generative.

In addition to gestural behaviour, Ham et al. [4] include gaze behaviour in order to study its influence through a storytelling Nao robot. For their study, the authors animated 21 gestures and 8 gazing behaviours designed in collaboration with a professional stage actor. The results of their study indicate a greater effectiveness when gesture and gaze are combined as opposed to gesture or gaze alone. The GENTORO system by Sugimoto et al. [159] has been developed to support children’s storytelling with a robot and a handheld projector. The system can enhance children’s embodied participation in the sto-
rytelling experience. In a similar classroom setting, Catala et al. \cite{160} present storytelling with an automated agent as a creative activity. The authors use a simple virtual bot that can be controlled through tablets and implements a distributed storytelling interface. Extending this with an embodied robot and emotion, Costa et al. \cite{161} present a series of empirical studies. Their most important findings indicate that an embodied robot creates more attention compared to a virtual storyteller. However, not all research with embodied storytellers focuses on gestural expression. Expressive facial cues for a robotic storyteller have been investigated by Striepe and Lugrin \cite{10}. Their storytelling robot is an expressive robotic face called Reeti. The authors present test subjects with stories delivered by Reeti, an audio book, and a neutral robot speaker. Despite the fact that Reeti is merely a head without a body, the authors are able to use it to convey ironic intent \cite{112}. Moreover, they show that coherent use of facial expressions yields a greater perceived story-consistency \cite{162}.

In the relevant literature on robotic storytelling, a variety of modalities is investigated and compared with a strong focus on the interactive effects of those modalities. While some attempts have been made to automate non-verbal support for storytelling \cite{13}, most approaches rely on a fixed (non-generated) corpus of stories \cite{10, 157, 158}. Moreover, none of these approaches considers the modality of spatial, locomotive movement. Furthermore, these studies disregard the commonalities of movement schemas in locomotion or gestures \cite{16, 80}.

2.2 Embodiment

Our assessment of space and gesture in a storytelling performance, as presented in this thesis, is deliberately focused on their adaptability to other forms of creative performances. Storytelling is chosen as the task for the experimental evaluation, yet at its core a story starts with a textual script. This original script can become a story, a dance, a theater play or a joke. Those different domains come with a variety of different affordances, yet ideally all of them make use of space and gesture to express the same underlying ideas. The frameworks presented here can be transferred to other tasks that can benefit from space and gesture. Hence, an overview of other related works on other forms of embodied robotic performances is warranted. Notably, the notion of Embodiment is the one introduced in Sec. 1.3.1.

2.2.1 Embodied Robotic Performances

Movement that concerns the whole body of a robotic agent has been studied for a variety of different tasks such as comedy \cite{163}, theater/improvisation \cite{25, 26}, and dance \cite{27, 28}. For example, Rond
et al. [164] present ImprovBot, which is a robot utilizing basic schematic movements (towards, away, sideways, spin and circle) in narrative scenes. The movements aim to further evoke creative interpretations of the narrative. Robots can foster enactment through very different means. ImprovBot does so by collaborating with humans in improv activities [164]. Another instructive example of robotic performance that relies on embodiment is Shimon. The robotic marimba player Shimon uses gestures as building blocks for musical creation. The robot transfers motion (gestures from the human) to music. For this, the system relies on the schematics of the gestures, which allow it to improvise the musical composition.

In the field of acting robots, the natural movement of the robots has been inspired by animation principles or Laban Movement Analysis (LMA) [165]. LMA combines insights from anatomy, kinesiology and psychology to describe, visualize and interpret human movement. It is often used by dancers, physiotherapists and athletes.

Animation principles and LMA are integrated by Bravo Sánchez et al. [166] to create plays with robot actors in an educational context. Notably, Bravo Sánchez et al. [166] outline several challenges when creating a play with robots for educational purposes. Some of these challenges are nonetheless relevant for performances with robots in general. First, they note that pre-assembled robots, like the Nao robot, cannot easily change their appearance to fit a character. Yet, there are only few low-cost robots available and, as we will argue, Nao’s versatile movement can add to the embodiment of a character through iconic gestures. Second, they note that robots that do not move on the stage are limited in their expression and low-cost strategies for navigation on stage are challenging - a challenge that will be addressed here with spatial schemas in the practical implementation (Chapter 4). Third, Bravo Sánchez et al. [166] note that a high-level description of the performance script should exclude any details of the performance that can be enriched by the autonomous agents themselves. Our implementation responds to this challenge with a dedicated Interpretation Framework (Sec. 3.2).

While not all of these challenges are relevant to storytelling robots, Bravo Sánchez et al. [166] provide a comprehensive overview of the challenges posed by interactive drama with robots. These challenges are also relevant for robotic actors in domains such as the visual arts [167], co-creative dance [168], dance creation [169] and in the context of art therapy [170]. For the latter, research by Cooney and Menezes [170] provides examples of how motion can be used to foster creativity in the physical world. Those authors explore emotional and creative painting with a robot. In their own implementation of a robotic painter, they define heuristics for valence and arousal to convey basic emotions. On this deeper level, their system is then built on basic shapes and movement directions that are expected to evoke an
emotional response. Their future work aims to discover how complex emotions can be evoked through simple shapes on a canvas, an idea which is closely related to the schemas of movement in our work.

Dialogue is another aspect of performance in which robotic speakers can make use of gestures. Wilcock [171] presents WikiTalk, a hybrid system which is half question-answering system and half spoken-dialogue system. It uses Wikipedia as a knowledge source and renders the retrieved content in a conversational manner through the use of gestures. This research uses a Nao robot that enables a detailed dialogue procedure with the embodied agent. For example, the authors explain how Nao’s speech recognition confidence scores can be used effectively in a dialogue, how awkward topic shifts can be remedied by spelling out the first letters of the new topic, and how the system can be gracefully interrupted during a conversation. For the latter, the robot acknowledges the interruption with an “Oh sorry!”. These considerations have proven useful in our own practical implementation and can guide a smooth interaction with dialogue systems in general.

Csapo et al. [172] discuss the different modalities of face detection, tactile sensors, non-verbal cues and gestures in WikiTalk. They use their robot’s inbuilt face recognition software, as well as sonar sensors and speech direction detection to initiate the conversation. Additionally, they use a set of six gestures to signal discourse-level details, to introduce hyperlinks, or to manage turn-taking with human interlocutors. Regarding speech and gesture synchronization, the authors report that their animation software did not accurately reflect the timing of gestures when performed by their physical robot. Since, the meaning of gestures is highly sensitive to the co-occurring speech act, the timing of the movement is essential for the gesture to bear the intended meaning [173]. Thus, Csapo et al. [172] provide useful evidence for the importance of synchronizing gesture with co-occurring speech in a robot interaction. These authors, as suggested by gesture scholars [173], split each gesture into preparation, stroke and retraction phases to align the pitch of the spoken sentence with the stroke of the gesture.

2.2.2 Meaning-Making in Movement

Embodied movement through spatial locomotion or gestures is an important facet of meaning-making, which has relevance in a variety of fields. When robots move on a stage, their performances should convey additional meaning that augments that of the spoken dialogue and narration. When we speak of the meaning-making process of movements and gestures, it is tempting to dissect this interpretation into discrete components, and ask: what components of movements and gestures convey specific aspects of meaning? While in sign lan-
guage, signs have a morphemic structure that can be dissected and analyzed [174], gestures are not signs in this Sign language sense, and cannot usually be dissected into smaller meaningful units. Indeed, just as with spoken language, signers can use gestures with sign language [175].

2.2.3 Multi-Agent Systems

Performances often consist of multiple actors collaboratively working together. In our case, the performance comprises a single system orchestrating multiple agents. In computer science, a system which is composed of multiple interacting agents is referred to as a multi-agent system (MAS). Ferber and Weiss [176] list the following use-cases of a MAS: problem solving (i.e. as opposed to solving a problem in a centralized way, multiple agents can provide distributed problem solving), multi-agent simulation (i.e. creating artificial environments to test behaviours), construction of synthetic worlds (with a focus on interaction mechanisms and their impact on the system) and collective robotics (in which each agent has their own goal and cumulatively accomplish an overall goal).

Weyns et al. [177] distinguish between direct and indirect messaging between multiple agents. An example of direct messaging is implemented by the peer-to-peer application architecture, which describes a decentralized network of nodes that share resources (memory, bandwidth or computing power) among each other [178]. For example, a scalable peer-to-peer file distribution system has been implemented by Ratnasamy et al. [179]. Their Content Addressable Network provides efficient scaleability through the use of hash table-like functionality.

Conversely, an architecture that supports indirect interaction is the Blackboard architecture [180]. As a data-structure, a blackboard is a flexible, segmented repository of heterogeneous information (which resembles more a school noticeboard than a school blackboard). Multiple agents do not interact directly with each other, but they can pose their requests and contributions on a single, shared blackboard data-structure. The blackboard control system coordinates all procedural strategies and integrates various scheduling rules and operations [180]. An embodied storytelling performance does not require any large scale distribution of computing power, and the individual agents can rely on this basic version of indirect communication. Hence, the multiple devices (our smart speakers and robotic actors) communicate via a blackboard architecture. MAS applications that successfully employ a blackboard architecture to manage backstage interaction between multiple actors and agents include [181], [182] and [183].
2.3 SPACE

2.3.1 Embodied Cognition

The use of space in robotic performances allows the enactment of meaning in the physical dimension. In fact, the physical space is the theatre of immersion where the actor’s body – and those of an audience willing to play-along – is used to create a performance. The shared space and mutual embodiment of performer and audience can establish an autopoietic feedback loop. Induced by the shared space, every behaviour of an actor triggers a specific behaviour in a physically present spectator, and vice versa, thereby influencing how the actors behave [184]. For example, unexpected applause or laughter following a subtle joke might lead an actor to reiterate the punchline. When a physical robot is able to interact with its environment and manipulate it, its own perception will be altered. Likewise, the enactment alters the perception of the audience. A performing physical robot thus blurs the boundaries between its internal computations and its external movement, achieving a basic implementation of embodied cognition [185]. Since the philosophical framework of embodied cognition is not only relevant for the embodiment of the robotic actors, but also for the schematic underpinnings of their movement, a brief overview of the topic is warranted.

One of the two prevalent streams of philosophy of mind is Eliminativism, the theory that common-sense conceptions about our mind are fundamentally wrong, and that most of what is defined as a mental state does not exist and therefore cannot be reduced to some biological, physical correlate [186]. The other stream is Representationalism, the theory that semantic (or intentional) states are essential in order to derive a theory of cognition [187]. James Gibson derived an eliminativist theory [188] in reaction to computational theories that reduce cognition to rule-based manipulations of representations. His theory views perception as direct, action-oriented and replete with affordances6. As such, the behaviour of humans or animals must be explained in relation to their environment, which is why this theory is often referred to as Gibson’s ecological theory. Combining the eliminativist approach of ecological psychology with a representational Computationalism gives rise to the discipline of Embodied Cognitive Science [186]. Most importantly for the frame of this study, embodied cognitive science has led to a rethinking of language in the context of the body, i.e. language formerly being studied within a computational framework now includes its embodied nature [8].

---

6 An affordance of the environment is any action that the environment allows the agent to perform. It is mutually dependent on the agent and its environment [188].
In experimental research, these considerations can be observed in the Language Games conducted by Steels and Belpaeme [190] and Spranger et al. [191]. In this series of studies, a group of robotic agents has been initiated with the task of perceptually grounding categories, asking each other to perform and perceive body motions and other coordination tasks with an artificial, dynamically-evolving language.

2.3.2 Proxemics and the Stage

Embodiment necessitates a space without prior assumptions of movement in this space. In contrast, performative embodiment should make assumptions about the movement performed in that space, especially since the context of space in a performance can provide a vastly different context for interpretation. For instance, our understanding of the movements of pedestrians on the street is quite different from our understanding of actors moving on a designated stage. An audience might observe the space from a viewing gallery, and a proscenium arch can frame the action in a way that encourages a different kind of dramatic interpretation. This framing of the stage can alter our construal of meaning within that space [103].

The embodied performances in this research reflect the study of how humans use space. These studies, of what is termed proxemics, have most prominently been shaped by Hall [14]. Figure 11 depicts the interpersonal distances defined by Hall [14] with the position of a human at the center point (Fig. 12 is an adaptation for a humanoid robot [2]). The Intimate Space expands from 0 to 1.5 feet, the Personal Space from 1.5 to 4 feet, the Social Space from 4 to 10 feet and anything further than 10 feet is considered Public Space. Walters et al. [192] propose an empirical framework for human-robot proxemics in which Hall’s zones have been tested in the context of Human-Robot-Interaction (HRI), over a range of distances for a variety of human-robot interactions such as verbal communication, giving an object or taking an object. In a study by Hüttenrauch et al. [193], the authors test Hall’s distances of personal space in a HRI task by evaluating relative orientation and spatial distance between a robot and a human. Their results indicate that a simplistic parameterization of Hall’s distances may be insufficient to capture the dynamic character of spatial movement. Hence, they conclude that its adoption might not be straightforwardly beneficial in the design of socially appropriate robots.

Lu and Smart [194] propose theatre as a basis for modeling and evaluating robot interactions. These authors are less interested in the use of the theatrical stage than in the benefits to HRI research that a theatre context may provide over the traditional social psychology settings. First, they argue that modelling robot movement with the help of skilled actors allows one to “get a limited amount of motion to be
Figure 11: Interpersonal distances, after Hall [14]. Proxemics, the study of how humans use space, indicates how embodied performances can make appropriate use of spatial zones.

Figure 12: Each ring is one robot-step wide. The expansiveness of a robot’s gesture is tailored to its ring.

subtly nuanced, intended to convey the actor’s intent.”[194]. Second, the feedback of theatre professionals can help to improve robotic motion, because roboticists are usually not schooled in movement on a stage. Third, Lu and Smart [194] argue that a potentially ambiguous movement of a robot can be made explicit and unambiguous when it is set into the larger context of a performance.

2.3.3 Image Schematic Uses of Space

The basic back and forth of an embodied robot on a stage requires an understanding of space, but so does movement that is not bound by a stage, e.g. the flight of drones is not bound to the horizontal plane. Here, Bevins and Duncan [195] study the perception of drone movement and evaluate how observers respond to a set of schematic flight paths, e.g. Up-Down and Left-Right. Doing so, the authors present a mapping of flight motion to meaning. The sixteen flight motions include participant-generated flight paths and their own proposed motions. When participants were asked “If you were in the room with the robot, how would you respond immediately following the robot’s action?”, “Moving Away” was the most common response for various flight paths. Flying from front to back is attributed with “Follow it”. The authors conclude that most complex movements likely result in participants moving away from the drone. They suggest that subtle and simple movements draw the viewer closer. It is known from the early studies of Heider and Simmel [6] that even simple movement in space can lead an audience to project intentional behaviour onto inanimate objects, to perceive emotion where there is only motion (see Sec. 1.2.1).
<table>
<thead>
<tr>
<th>Schema</th>
<th>Movement</th>
<th>Metaphorical Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEAR</td>
<td>The robot is moving near another robot or object</td>
<td>There is an interest or sympathy</td>
</tr>
<tr>
<td>FAR</td>
<td>The robot is moving further away from the robot/object</td>
<td>There is a growing disinterest or disliking</td>
</tr>
<tr>
<td>FRONT</td>
<td>The robot is moving or turning in front of itself or robot/object</td>
<td>The robot is actively engaging</td>
</tr>
<tr>
<td>BACK</td>
<td>The robot is moving or turning to the back of itself or robot/object</td>
<td>The robot is actively disengaging</td>
</tr>
<tr>
<td>UP</td>
<td>The robot is moving upwards</td>
<td>The robot is displaying some superiority</td>
</tr>
<tr>
<td>DOWN</td>
<td>The robot is moving down or downwards</td>
<td>The robot is displaying some inferiority</td>
</tr>
</tbody>
</table>

Table 1: A listing of spatial schemas with their robotic realizations, indicating an additional potential for metaphorical meaning. Each row contains a schema that has an inverse motion in another row.

A study by Nakanishi et al. [196] shows that even minimal movement on one axis of a robot-mounted camera increases one’s sense of social telepresence. Implementing rotation and directional movement in a museum robot, Kuzuoka et al. [197] show that a robot’s rotation can influence the position of a visitor, and that full body rotation is more effective than partial, upper-body rotation. Nakauchi and Simmons [198] have investigated spatial movement in the context of queuing in line, and consider relative positioning in line as a parameter for achieving optimal, socially-accepted movement. Notably, in embodied performances the movement in space is especially useful because of its metaphorical potential. Table 1 provides examples of how schematic constructions, implemented simply with robots, can convey intention and emotion. However, even pervasive everyday metaphors [48] can have their exceptions. For example, UP may generally signify good, and DOWN bad, but we want a fever to go down, and do not want costs to go up. This observation also applies to the schemas presented in Table 1. There are some situations where moving away increases emotional closeness, and moving closer decreases it, as when e.g., the former signifies awe and great respect, and the latter signifies contemptuous familiarity. As with all powerful schemas, the benefits of generalization outweigh the occasional exceptions. While these and more general cognitive schemas are discussed in Sec. 1.2.2, they can also be found in the second class of robotic movements discussed here: Gesture.
2.4 Gesture

Embodied performances often feature acts of communication which do not solely rely on speech acts. Conversely, speech acts are often accompanied by gestures. The role of the body in communication had long been disregarded, but empirical work in cognitive science by McNeill [199], Bergen et al. [200] and Hauk et al. [77] has shown that the body is an important instrument for human language and communication. A study of Embodied Sociolinguistics by Bucholtz and Hall [79] 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. [201] conduct an extensive investigation into the evolution of speech originating from the body. The authors 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 physical 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 internal conceptualization of ideas and their external expression. McNeill describes them as fundamental assets of linguistics for our conceptualizing capacities [88].

As semiotic objects, the gestures understood as metaphors act as a cross-domain mapping to express internal feelings, concepts and thoughts in concrete terms [202]. As embodied signs, gestures do not only speak for themselves, they serve as a context for speech, while speech also serves as a context for gestures when both are integrated successfully. This contribution of additional meaning to the communicative act has been empirically verified in a number of experiments [203, 204].

2.4.1 Types of Gestures

Despite technological progress in the videotaping and analysis of gestures and body language, there is still no unified methodology to annotate and classify gestures [205]. Nonetheless, a range of studies, such as those in gesture recognition [206, 207], consider Kendon’s separation [173] of a preparation, stroke and retraction phase for the structure of a single gesture. For an overall classification, most studies refer to McNeill’s [88] classification of gestures into iconic, metaphoric, deictic, cohesive and beats:

- **Iconic**: A gesture resembles what it denotes. Example: Shadow boxing when talking about a fight.
• **Deictics:** A pointing gesture may refer to another object. Example: Pointing at another actor on stage.

• **Metaphoric:** A figurative gesture should not be taken literally, yet it communicates a truth about the situation. Example: Showing a trajectory with the hand when talking about a trip.

• **Cohesives:** A cohesive gesture binds two temporally distant but related parts of a narrative. Example: Making the same hand movement whenever the same character appears.

• **Beats:** A gesture marks narrative time. Example: A rhythmic arm movement indicates time passing.

The class of *Iconic* gestures requires that audiences recognize the iconicity of a gesture when it is performed by a robot. A study by Bremner and Leonards [208] shows that iconic gestures performed by a robot can be understood by humans almost as well as those performed by humans. Another study, conducted by Salem et al. [209], suggests that human evaluation of a robot is more positive when it uses iconic, referential and spatial gestures in addition to speech. Regarding spatial and referential gestures, it has been argued that gestures are primarily used to augment non-visuospatial speech communication with visuospatial information [88].

In the five classes of gesture outlined above, most can convey some visuospatial information, but *Deictic* gestures do so by definition. *Deictics* play a crucial role in human to human communication by supporting direct reference to visual and non-visual objects [210]. It has also been shown that robotic deictic gestures can shift our attention in much the same way as human uses of these gestures [211]. The level of abstraction in *Metaphoric* gestures is generally higher than that of *Iconic* and *Deictic* gestures, and there is evidence to suggest that distinct integration processes apply to these different classes of gesture in the human brain [212].

Metaphors exploit familiar source domains, so the same gestural movement can be metaphorical in one speech context and iconic in another. For example, the gesture "raising one arm above the head with a horizontal, open hand" is iconic when it accompanies the sentence "The plane flew way above the clouds", and metaphorical when it accompanies "She is way out of your league." See Figure 13 for a visualization of a Nao robot performing this gesture. A study by Huang and Mutlu [213] investigated four of McNeill’s gesture classes (all but *Cohesives*) as used by interacting humanoid robots in a narrative context. Those authors evaluate each gesture type on several fronts: information recall, perceived performance, affective evaluation, and narration behaviour. In their study, *Deictics* are shown to improve information recall relative to other gestures, while *Beats* lead to improvements in effectiveness.
Figure 13: Example of the ambiguity of gestures. The raised arm above head can vary in its interpretation, when presented with a different accompanying speech act. On the left, the gesture is iconic and the hand signifies the plane flying above the clouds. On the right, the gesture is a metaphoric signal of superiority.

There are observable overlaps between the reference framework used within spoken language and the reference framework used with gestures [214]. For example, if an event occurs to the left of a person, that person is more inclined to gesture to their left when retelling the event [88]. While this appears to hold for most Indo-European languages, there are some cultural dependencies. Speakers of the Mayan language Tzeltal use an absolute spatial reference framework for both speech and gesture, so if an event occurs to the west of a Tzeltal speaker, they are inclined to point west when they later tell of it [215]. Another example of cultural diversity is found in the Aymaran language. The Aymara people of the Bolivian Andes refer to future events by pointing behind rather than ahead of themselves [216]. However, some gestures appear relatively stable across cultures when there is a consistent, well-established link from form to meaning [93]. These recurrent gestures often serve a performative role, and fulfill a pragmatic function when they work on the level of speech [217].

2.4.2 A Computational Look at Gesture Generation

In order to utilize the expressive value of an automated actor through movement and gesture, it is necessary to develop a computational model. Such a model must be sufficiently close to human movement,
if it aims to elicit reliable, meaningful motion for a specific task. At the same time, the model must be able to (a) generalize for a variety of different embodied robots, (b) provide different moves within the same performance and (c) incorporate new movements. The challenge posed by the first requirement can be met by implementing a model which works on the level of schematic movements for virtual, embodied agents that do not depend on a specific physical implementation [218, 219]. However, these schematic approaches often fall short on (b) and (c), but machine learning can provide models with an ability to generalize. Chiu and Marsella [220] present a machine learning approach in which the task of non-verbal communication is divided into two sub-tasks. Unlike their previous approach, which models a direct mapping between two time series of speech and gestures [221], their improved model first maps speech to gestural annotations and then maps gestural annotations to gesture motions. For the first step, speech features are mapped to annotations using conditional random fields. The second step, synthesizing gestures, is achieved by learning representations of gesture motions using Gaussian process latent variable models. Finally, an interpolation algorithm determines a feasible trajectory which combines motion segments.

Data-driven gesture generation has been applied to different modalities, e.g. audio and text. Sadoughi and Busso [222] present a probabilistic graphical model based on speech audio input. Their model can be constrained to generate a discrete set of prototypical gestures, which allows it to be combined with rule-based systems. Similar audio-driven gesture generators have also been proposed by Ginosar et al. [223] and Ferstl et al. [224]. The former have trained convolutional neural networks on spectrogram features in order to generate 2D poses. The training included an adversarial mechanism, which has been adopted by the latter authors to work within a recurrent neural network (RNN) model. Ferstl et al. [224] combine this RNN with a gesture phase classifier to account for realistic gesture phasing.

Other work has focused on text-transcription-driven gesture generation. For example, another RNN approach learns a mapping of speech text to gestures [225]. Yoon et al. [225] train their Encoder-Decoder model with TED Talk video transcripts and pose estimations. The model learns the motion trajectories, which are then aligned with audio in a post-processing step for a social robot.

A multi-modal approach has been suggested by Kucherenko et al. [226]. Their Gesticulator gesture model "takes both acoustic and semantic representations of speech as input, and generates gestures as a sequence of joint angle rotations as output" [226]. For each frame of the video input, text and audio features are encoded, concatenated and fed into a feed-forward neural network. At the end of several fully-connected layers, the output pose is sent back into the model for autoregressive
training. Although the authors show that the combination of both modalities (audio and text) can improve gesture-generation models, they note that their system relies on a manually annotated data-set.

This current work does not include a computational, generative gesture model, but it does exploit the underlying patterns which can be found in human gestures. The evidence for those schematic patterns and their connection with gestures studies will now be surveyed.

2.5 Image Schemas

In our embodied storytelling experience, we want the motions and movements of the robots to be meaningful and not merely ornamental. Ideally, the movement will be coherent with the accompanying speech acts and will help to convey the ideas of the story. Ultimately, we are trying to map complex plot developments onto simple physical movements. The example case for our studies in Chapters 4 & 5 maps the mental distance between characters in the story onto physical distance of the actors on the stage. In essence, we are exploiting a conceptual metaphor.

A conceptual metaphor refers to the understanding of one conceptual domain, the target, in terms of another, the source \[48\]. For example, we can understand time in terms of money (e.g. “Stop wasting my time.”) or love in terms of a journey (e.g. “They parted ways.”). In case of love as a journey, the journey is the source domain and love is the target domain of the mapping.

Ideally, the gestures and movements in space will exploit simple spatial schemata. For this, we turn to image schemas. Image schemas are recurring cognitive structures shaped by physical interaction with the environment \[8, 67\], and can be observed not just in verbal but in non-verbal communication \[16, 92\]. As a distillation of embodied cognition, image schemas emerge from bodily experience and motivate subsequent conceptual metaphor mappings. They have been extensively studied as source domains for conceptual metaphors \[227, 228\]. While image schemas are not conceptual metaphors themselves, their role extends further than their use as possible source domains. Man- dler \[66\] argues they are involved in concept formation at a fundamental level.

A practical example of an image schema is CONTAINMENT\[7\]. We often experience containment in our physical world, whether by entering a room or pouring water into a glass (see Fig. 14 for further examples). Whether we are contained by the walls of the room or the water is contained by the glass, CONTAINMENT is a recurring, pattern of our interactions. Moreover, the CONTAINMENT schema

\[7\] As per convention, image schemas are written in uppercase letters.
Figure 14: Depictions of the image schemas CONTAINMENT, IN and OUT adapted from Johnson [8](p.23, p.32 and p.32 respectively). These present only one possible instantiation of each schema.

is often co-activated with the schemas FULL and EMPTY, because the fluid fills the container or the container is empty. These schemas can motivate a conceptual metaphor for which Johnson provides the example “Let out your anger” [8]. Here, anger, a metaphorical “fluid” housed in the body, is said to be released from its container.

Image schemas are, as this work will demonstrate, a suitable prelinguistic underpinning that can map complex ideas of the plot onto schematic physical movements. We will review the related literature on image schemas to answer a few important questions: What constitutes an image schema? How many image schemas are there and how are they categorized? What are spatial image schemas? How are image schemas useful and which are most relevant to this work?

2.5.1 Properties of Image Schemas

The term image schema was first introduced by Johnson [8] and mentioned in a case study by Lakoff [67]. Therefore, we rely on Johnson’s definition for what constitutes an image schema, examples of which are listed in Table 2:

An image schema is a recurring, dynamic pattern of our perceptual interactions and motor programs that gives coherence and structure to our experience [8]

Cienki [23] provides an overview of some shared properties and groupings of image schemas. Some fundamental properties are shared by all (or most) image schemas, while others allow us to group them. Some general properties shared by most image schemas are, according to Cienki [23]:

- Plus-minus parameter: This property, as described by Krzeszowski [229], is a bipolar tendency of most image schemas to be intuitively positive or negative. The intuition about the polarity is a
Table 2: Relevant image schemas from Johnson [8] and adapted from Cienki [23]. This list is not exhaustive.

connotation primarily derived from our bodily experiences. For example, CENTER is often understood to be more positive than PERIPHERY. Likewise, the WHOLE is often more positive than only a PART.

- Static versus dynamic nature of image schemas: Almost all image schemas can be understood as a state or as a process. In most situations CONTAINMENT is a state, e.g. water is contained by a glass and there is no active or dynamic process in which the glass is acting to contain the water. Conversely, the COMPULSION schema typically implies an active process in which a dynamic movement can occur. Cienki [23] notes however, that this situation can be inverted: CONTAINMENT is a dynamic process when a sheepdog is trying to contain a herd of sheep [23].

- Superimposition: When FULL or EMPTY often occur with CONTAINMENT, Johnson [8] speaks of a superimposition of schemas that overlap in their meaning and experience. Indeed, overlapping schemas can be hard to distinguish and individuate as separate schemas. As a consequence, this property groups certain image schemas based on their co-occurrence.

2.5.2 Quantity and Category of Image Schemas

The list of 27 image schemas in Table 2 is not exhaustive. To date, there is no closed set of image schemas, but there is a set of reliably recurring schemas that have been studied in cognitive linguistics: CONTAINER, OBJECT, PATH, PART-WHOLE and UP-DOWN [8, 23, 67, 227]. Moreover, an exact categorization of image schemas is problematic, because the superimposition property makes some image schemas more primary than others [230]. For example, CONTAINMENT may be considered more primary than FULL-EMPTY, or the
2.5 IMAGE SCHEMAS

CYCLE schema may be viewed as a specific case of the PATH schema (i.e. a circular path). This also allows more abstract schemas to be composed via the combination of more basic schemas, a property we will shortly review.

One further relevant distinction comes from Mandler, who looks at the formation and development of image schemas. She and her colleagues distinguish between spatial primitives, image schemas and schematic integrations. For Mandler and Cánovas [231] the set of primary image schemas (e.g. PATH, CONTAINER, THING, CONTACT, etc.) consists of the very first building blocks, which enable infants to make sense of their perceptions. Mandler and Cánovas [231] call these schemas the spatial primitives, and the authors argue that all image schemas are formed from these spatial primitives. Those primitives are considered the first conceptual building blocks formed in infancy through observation and interaction with the environment. For example, the observation of contact between objects, or lack of it, provides crucial conceptual information to distinguish inanimate from animate objects [232]. Inanimate objects only move when they forcefully connect with something, whereas animate objects do not need to connect with anything in order to move. For Mandler, image schemas are our representations of simple spatial events, as composed from spatial primitives.

Mandler [233] argues that the evolving conceptual system of infants starts out as a spatial system, built from those spatial primitives. She argues that space is highly salient to a human being, and spatial information can be easily understood. Mandler shows that spatial representations can form a “viable conceptual system that enables interpretations of objects and events without involving other kinds of information.”[233]. In [234], she provides evidence that a conceptual system built from spatial schemas enables early word understanding, because all notions in early language acquisition are describable with image schemas.

Mandler argues that schematic integrations describe the final image schematic stage in which conceptual representations begin to include non-spatial elements. Mandler [235] tells of an observation by Piaget [236], who describes how his 11-month-old daughter imitates his blinking, by first opening and closing her hands, then doing the same with her mouth, and later covering and uncovering her face. Mandler offers this as an example of how a familiar schema is mapped onto something else in order to understand it. In these instances of schematic integration, bodily action is blended with spatial understanding, and the spatial schema becomes the source domain for a conceptual metaphor.
An alternative grouping of image schemas is proposed by Cienki [23]. This grouping is composed of more general schemas, which subsume sets of more specific schemas:

- **PROCESS**: MATCHING, MERGING, CONTACT, LINK, SPLITTING
- **PATH**: STRAIGHT, SCALE, ITERATION, CYCLE
- **OBJECT**: PART-WHOLE, CENTER-PERIPHERY, SURFACE, MASS-COUNT, COLLECTION
- **CONTAINER**: FULL-EMPTY, SURFACE, CENTER-PERIPHERY

### 2.5.2.1 Spatial Schemas and Force Dynamics

If all image schemas are built from spatial primitives, then all image schemas can be considered spatial schemas. However, Cienki [23] argues that some image schemas are more obviously force-dynamical than spatial. The term *force dynamics* refers to how entities interact with respect to force [237], and in the context of image schemas, it refers to those image schemas that imply an interaction of forces, e.g. BLOCKAGE, RESTRAINT, COMPULSION, COUNTERFORCE, ATTRACTION, etc. Cienki [23] thus suggests this addition to his main groupings above:

- **FORCE**: ATTRACTION, ENABLEMENT, COMPULSION, COUNTERFORCE, BLOCKAGE, RESTRAINT

### 2.5.3 Application of Image Schemas

These theoretical considerations find their practical application in different related works. Not only does the theory of image schemas provide a *Conceptual Scaffolding* [9] for narrative processes, it also provides an algebraic basis for modeling complex processes and situations [69]. In this way, simple schemas can be used as primitive building blocks of larger, more complex structures [9, 70].

In artificial intelligence, image schemas are helpful with natural language comprehension or task planning in cognitive robotics. Imagine writing precise instructions for a robot that is supposed to crack an egg. Questions arise about the level of detail sufficient for the instructions to be executed, e.g. does every egg need to be grabbed in the same way and what is the exact force needed to crack it? This egg example is a frequent exercise in commonsense reasoning [238]. Spatial primitives can be combined to propose a solution for this problem [239]. For this, Hedblom et al. [239] propose a systematic combination of image schemas in order to capture concepts and events that show far greater complexity than a single image schema alone can provide. Hedblom [69] presents a formal combination of image schemas and demonstrates how her methodology can be applied to the *egg cracking problem*. First, the author formalizes image schemas, which
is a non-trivial task in and of itself, because of their superimposition property and, as Hedblom et al. [240] state, it “is sometimes not obvious whether two conceptual structures are just variants of the same image schema or whether they are different image schemas.” Consequently, the authors propose an Image Schema Logic (ISL).

ISL is a combination of various different formal calculi and first-order logic. ISL forms an “expressive multi-modal logic” [239] whose hierarchical structures represent image schemas at different levels of specificity. Each sub-logic in ISL can represent a different modality of an event or object. ISL has a topology of regions (e.g. whether two objects overlap or touch), cardinal directions (e.g. LEFT, RIGHT, BEHIND, ABOVE, etc.), movement dimension (e.g. whether an object moves towards or away from another object), temporal dimension (e.g. at some time in the past, at some time in the future, at all times, etc.), where each has its own syntax. By using image schemas as building blocks within a coherent logic, Hedblom et al. can use image schemas for concept invention [15], conceptual blending [240] and the modeling of commonsense reasoning [239]. An example of how these authors combine and hierarchically arrange image schemas is shown in Figure 15.
Richardson et al. [241] provide empirical evidence that supports the generality of image schemas in spatial reasoning, and show how spatial representations are evident in real-time language processing. The authors conduct an experiment in which they test the image-schematic intuitions of subjects for concrete and abstract verbs. Highly concrete verbs such as *fled, pulled, pushed* and *walked* are assumed to operate on a horizontal axis, which reflects the spatial representation of the LEFT/RIGHT image schema. Concrete verbs such as *flew, floated, sank* and *lifted* are assumed to operate on a vertical axis, which reflects the UP/DOWN image schema. Analogously, the highly abstract verbs *argue with, offended, rushed* and *warned* are also assumed to reflect the LEFT/RIGHT schema, while *hoped, increased, obeyed* and *respected* are considered more vertical (UP/DOWN). Each subject was presented with a list of verbs and four spatial representations of the image schemas UP, DOWN, LEFT and RIGHT. For example, the schema LEFT was presented as:

\[
\begin{array}{c}
\circ \\
\leftarrow
\end{array}
\]

along with verbs in the format: \( \circ \) offended \( \square \), \( \circ \) lifted \( \square \). The results from the study indicate that there is a shared agreement among the participants as to the spatial representations that best capture each verb. Moreover, this agreement confirms the expected axis for horizontal and vertical verbs, which implied the LEFT/RIGHT or UP/DOWN schemas respectively. With this and further evidence, the authors claim that *Language is Spatial* [241] and support a diagrammatic approach to language representation of the kind prominently described by Langacker [242] as *Cognitive Grammar*. In this view, grammar itself has an imagistic character.

### 2.5.4 Image Schemas in Gestures

Cienki [214] argues that gestures ground the cognitive model of situated speakers in their physical environments. Since the schematic nature of certain movements across different gesture types has been related to deep-seated image-schematic structures, Cienki presents an empirical study which assesses the reliability of image schemas in the characterisation of spontaneous gestures co-occurring with speech. In a dyadic setup, one set of gestures is produced in reference to abstract ideas (\( A \) gestures) while another (\( O \) gestures) serves a discourse-structuring or performative function. The classification of \( A \) gestures and \( O \) gestures for this study has been adapted from Müller [243]. \( A \) gestures denote abstract references such as abstract entities, properties, behaviours and actions or relative location and relative time. For example, relative time is denoted with a gesture relating to the back (past) or front (future) of the body with only a few known cultural exceptions [216]. In contrast, \( O \) gestures denote other discourse-structuring or performative functions such as emphasis, enumerating,
presenting, dismissing or requesting. Four conditions presented the same phrases: one with gestures and a transcript of the accompanying speech, one without sound, one without video but with spoken audio and written transcript and lastly, one with only a spoken transcript.

Participants were asked to provide the most appropriate descriptor for what they were given from a list of image schemas: CONTAINER, CYCLE, FORCE, OBJECT, PATH or OTHER if none of the image schemas seemed appropriate. The results suggest that gestures provide accessible manifestations of image schemas readily available in gestural forms. Additionally, Cienki [92] reports that gestures evoke different image schemas than speech alone would evoke. He concludes that gestures manifest additional information in the communicative act. With respect to A gestures and O gestures, Cienki reports that A gesture categorization shows a higher inter-annotator agreement than O gestures. He suggests that this effect may arise from A gestures having a more imaginable schematic form. The notion that abstract ideas are being more accessible through schematic motions fits well with our current understanding of image schemas.

Mittelberg [16] reports that data from motion capture experiments offer similar evidence for the link between image schemas and gestures. In her approach, she draws from Gibbs’ dynamic systems account of image schemas [244]. Gibbs’ understanding of embodied cognition relates to radical embodied cognitive science [189] in which there are no mental representations, but only immediate bodily interactions with the environment best described by differential equations or dynamical systems. Gibbs views image schemas as attractors within human self-organizing systems [244]. Pointing to Johnson [245], who emphasizes the role of image schemas as recurring patterns of bodily interaction with the physical environment, Mittelberg lists some key properties of image schemas and argues that gestures exhibit very similar properties: Both support embodied simulation, create temporary conceptualizations of concrete and abstract events, connect embodied action with linguistic action and can each be understood in terms of dynamical systems. She suggests examples of attractors within these dynamical systems, such as BALANCE, SOURCE-PATH-GOAL, RESISTANCE and VERTICALITY, and shows how these schemas have also been identified in gestures [246]. The most compelling evidence comes from spontaneous gestural enactment in a multi-modal discourse sequence that is tracked with motion-capturing. Figure 16 shows the original trace (blue) of the motion-capturing from Mittelberg’s study [16], as superimposed on an abstract depiction of the speaker.

Mittelberg identifies several image schemas based on motion-capture: PATH, EXTENSION, STRAIGHT, BOUNDEDNESS, CONTAINER, HOR-
IZONTAL, and LEFT-RIGHT all co-occur with the speech act "from where I was till like the end of the season". She takes this evidence as supportive of her claim of a body-centered, dynamic systems approach to image schemas operating in gesture.

2.5.5 Embodied Storytelling and Image Schemas

For the embodied storytelling in this work, we consider the image schematic underpinnings of gesture and space. When it comes to gestures, the research of Ravenet et al. [219] suggests a highly relevant approach. The authors propose a rule-based model to produce gestures for a virtual avatar. The rules connect an arbitrary text with gestures, where the connection is established on the basis of the image-schematic properties of the words in the text. For each word, their algorithm follows the path of hypernyms up in the lexical dictionary WordNet [247] until it reaches an image schema entry (UP, DOWN, FRONT, BACK and so forth). Those image schemas are in turn associated with gestures. The authors did not implement their system in a physically embodied agent, nor did they evaluate their approach. Although their schematic approach to gestures for automated storytelling has clear overlaps with this work, the focus here is on the schematic use of space and gesture that goes beyond a fixed set of gestures.

A recent paper by Spitale and Matarić [218] proposes an automatic gesture generator for robots based on image schemas associated with a given text. Their Image Schema Mapper uses a set of common Natural Language Processing (NLP) libraries to retrieve the most semantically similar image schemas given a word in a sentence. The authors show how the sentence “Nice to meet you” is mapped to the JOIN image schema based on the similarity of “meet” with “join” (M. Spitale, personal communication, April 7, 2021). Much like the work by Ravenet et al. [219], this system is limited to the modality of gestures but links this to an evaluation of the affective dimensions of the given text. Two approaches aim to use general schemas instead of a specific set
of gestures to support embodied storytelling. The current thesis provides empirical and theoretical evidence as to how and why such an approach is effective and useful, not only for gestures, but for movement in embodied, automated performances.

Moreover, we expand on these approaches by not only looking at gesture, but at the schematic use of space. For this, we can use image schemas as the source domain for a conceptual metaphor. Scéalextric constructs stories that develop observable relationships between the primary characters of the story. We can now map the character relation (target domain) to the spatial relation (source domain) and utilize the image schemas NEAR/FAR. The empirical evaluation of Chapter 5 will show that the coherent use of space, with respect to its underlying spatial schemas, is more appreciated by an audience than an incoherent use. In some instances of the studies, we create specific image schematic gestures for specific actions in the repertoire of Scéalextric, since its approach is explicitly action-centred. In these cases, we will use image schemas as the motivating source domain to tie gestures to story actions when no pre-existing gestures are available. To choose a set of image schemas for this mapping, we turn to those suggested by Hurtienne and Blessing [68], who focus on image schemas that relate to space: UP-DOWN, LEFT-RIGHT, NEAR-FAR, FRONT-BACK, CENTER-PERIPHERY, CONTACT, PATH, SCALE. Those schemas are the basis for the empirical work and investigation of this thesis, and are depicted in Figure 4.

In conclusion, theory and empirical evaluation point to a strong connection between the recurring embodied patterns of human interactions and human gestural communication. The literature is clear in its support for the cognitive underpinnings of gestures. Thus, embodied performance must acknowledge the role of image schemas, especially when they manifest as gestures, or when abstract concepts are conveyed using spatial metaphors. The potential for creative interaction through grounded metaphors provides a solid basis for computational performances. More broadly, this suggests that creative performances and computational creativity can benefit from an embodied approach.

2.6 CREATIVITY AND EMBODIMENT

2.6.1 Computational Creativity

Embodied AI storytelling requires a text-based story generator that can be augmented with robots using gestures and movement to provide a physical enactment of the story. As surveyed in Section 2.1.1, we can choose from a range of text-based story generation systems in
the domain of Computational Creativity (CC). Other embodied storytellers have also been presented, which are all examples of embodiment in creative installations. This work is not the first to postulate that the role of embodiment is important in creativity research. Other researchers arguing for the importance of creative embodiment generally adopt a shared notion of creativity, irrespective of their various definitions of embodiment.

This shared notion of creativity views a process as creative if its product exhibits **novelty** and **value** [122]. Abiding by these two requirements, Ventura [17] lays out the key steps that distinguish mere generation from true creativity, some of which are depicted in Figure 17. On a merely generative level, a system can produce something **novel** without the product having any value, e.g. a random number generator can create truly novel sequences of numbers without those sequences carrying any meaning, hence it is merely generative. Conversely, a system can produce something highly **valuable** by retrieving pre-existing products from a database of masterpieces. This is value without novelty. Stepping up from mere generation, Ventura describes different ways of mixing or extrapolating novelty and value to achieve a minimally creative artefact.

One way is to take various pre-existing, valuable elements and combine them to produce something novel. This is the **pastiche** approach
as described by Veale [248]. Veale’s Pastiche system is not strictly a generative system, but provides an implementation of blending theory [249]. Because the system’s value emerges from its combination of structured conceptual spaces, the approach is more than merely generative, and exhibits some evidence of creativity. More specifically, such an approach falls under Margaret Boden’s [250] definition of combinatorial creativity.

Boden [250] defines three forms of creativity useful for the study of creative systems: Combinatorial creativity, exploratory creativity and transformational creativity. Combinatorial creativity seeks out new combinations of preexisting ideas and concepts, e.g. pastiche. Exploratory creativity arises when structured conceptual spaces are explored and this exploration leads to novel results. Lastly, transformational creativity is the type of creativity at play when the conceptual space itself is transformed, such that entirely new dimensions can give rise to novel results. For example, a music mashup would be a combinatorial act of creativity, whereas Boden [250] refers to Jazz improvisation as an exploratory act of creativity. Redefining music itself is a transformational act of creativity, as when Schönberg’s atonal music defied the conventions of tonal hierarchies that once dominated classical European music.

Conceptual blending can move a system from mere generation towards actual creativity. Additionally, Ventura [17] suggests some further mechanisms to expand the creative potential of a system. He argues that a first possible step involves generalization by a generative model built from a set of inspiring training examples. This is a strategy adopted by most creative machine learning systems, i.e. feeding large amounts of training exemplars into a system so that it can learn to generalize and to produce novel instances of its own. But Ventura argues that the novelty of such systems can be inherently restricted by the “use of constructs found in the inspiring set” [17]. Ventura proposes that a process of filtration can choose a valuable candidate solution among the generated outputs. He further argues that a key addition is the ability of a system to perceive. To provide a system with perceptual abilities allows it to ground its results and receive feedback on its impact in the environment. Indeed, a feedback loop with perceptual grounding is a key element of any embodied system. Embodiment is not a hard requirement for creative systems in the CC community [117], just as arguments continue as to whether or not some form of consciousness is also a requirement for creativity [251].
2.6.2 Co-Creativity

Jordanous [252] describes co-creativity as the active collaboration of at least two participants in a creative process. In the context of CC, at least one of the participants must be computational. Lubart [253] distinguishes four types of co-creativity:

- “Computer as Nanny”: The system supports the creator by monitoring the working process, suggesting work breaks or deadlines. The computer can take over mundane tasks (e.g. auto-saving or reminding) in order to unburden the creator and thus help them to focus on the creative process.

- “Computer as Pen-Pal”: The focus of such a co-creative system is on the communication of creative ideas. For example, brainstorming software can allow creators to exchange ideas among themselves. A pen-pal system can help to record, distribute and communicate ideas in order to foster creativity.

- “Computer as Coach”: An expert system can recommend tutorials and exercises that foster creative thinking. These systems have access to information that can trigger creative processes.

- “Computer as Colleague”: The creator and the computer work as colleagues in order to collaboratively create. For example, this can be a back-and-forth process between machine and human in which both add their input to jointly create something of novelty and value.

For the “Computer as Colleague”, Lubart [253] argues that “Computers can probably better implement random searches than humans but humans are needed to select the best ideas and perhaps to fine hone these ideas, turning them into viable creative productions.” [253] This duet between human and machine connects to another aspect of co-creative systems that Veale and Pérez y Pérez [254] argue goes beyond “uninspired drudgework of creativity” [254]. They note that the strong suit of machines is the ability to explore a vast space of combinatorial possibilities, which - given a set of rules and creative methods - can be exploited to identify artifacts of novelty and value. In contrast, the human in the loop is the entity that embeds the creations in a cultural, historical and semantic context [254]. Each provides the qualities that they excel in, in order to create something that the other would not be able to create by themselves.

The interactive, embodied storytelling performances in this thesis (Chapter 4) are the products of co-creative systems. The performances connect an audience and a story generator through means of embodiment and interaction. The story generator Scéalextric, like most generators, is able to explore a vast space of plot combinations. Our
systems then use non-verbal means to communicate the story and to support interaction via non-verbal communication (e.g. gesture, facial expression). Consequently, the embodied system is co-creative in the manner of a colleague, because it jointly creates something of novelty and value by enhancing its own authenticity with its vivid use of embodiment [117]. The particular ways in which embodiment can be beneficial for a creative system are reviewed next.

2.6.3 Embodied Creativity Research

Embodied systems can be observed along the entire continuum of computational creativity [119]. Guckelsberger et al. [118] highlight how the situatedness of an embodied agent provides the crucial constraints and the intrinsic motivation to facilitate computational creativity. These authors argue that a physical environment comes with affordances and constraints that an agent can take advantage of, to e.g., foster more natural interactions between humans and machines. Moreover, the authors argue that situated agents come with greater behavioural complexity, which cannot solely be accounted for by programmed routines. As a consequence, a situated agent is more likely to be attributed with creative intent.

Other CC research explores embodiment because of the physical nature of the creative domain. For example, Schorlemmer et al. [255] apply their computational concept invention theory (COINVENT) in the music domain as they regard music creativity, especially musical performance, to be situated and embodied. Likewise, Schubert and Mombaur [256] model motion dynamics in the creative process of abstract paintings. The authors look at action paintings in the style of Jackson Pollock to understand and subsequently model aesthetic judgement in their creation process. With this study, Schubert and Mombaur investigate the embodied expression of the artists and transfer the model of human expressive motion to a robotic arm.

Guckelsberger et al. [117] have conducted a systematic review of ten years (2010-2020) of works submitted to the International Conference on Computational Creativity (ICCC). They consider any submission with mentions of any of the following terms: “embodiment”, “embodied”, “disembodiment”, “disembodied”, “embody” or “embodying”. For their analysis of 39 papers, the authors extend a topology first introduced by Ziemke [257] to describe different uses of embodiment in cognitive science. Guckelsberger et al. [117] define six different notions that can describe an embodied system:

- **structural coupling**: Relating to a notion by Varela et al. [73], a system can be perturbed (i.e. it can be affected in a minor or a major way) by the environment it is embedded in, and the environment can in turn be perturbated by the system.
• **historical embodiment**: For this type of embodiment, the notion of \textit{change over time} is important. Structural coupling is an antecedent of historical embodiment, since it views the present state of the system as being the result of repeated, past perturbations.

• **physical embodiment**: Any system with a physical body that has some form of interaction with its environment is physically embodied. This includes robots of all shapes and forms that can be subjected to physical force and those that are able to exercise force.

• **virtual embodiment**: This type of embodiment is a further addition by Guckelsberger et al. [117], who argue that advances in simulated systems allow virtual bodies to affect the environment. This type of embodiment is the virtual counterpart of physical embodiment.

• **organismoid embodiment**: Any virtual or physical embodied system can be organismoid if the system’s sensorimotor capabilities are based on those of living organisms. For example, the Nao robot is a humanoid robot with vision and tactile sensors, and thus provides organismoid embodiment.

• **organismic embodiment**: This describes a system that has self-sustaining capabilities, and which actively and autonomously tries to distinguish its internal structure from external structures (i.e. its environment). For example, a system that is able to fend off external perturbations has some form of organismoid embodiment.

Those authors find that 28 out of 39 ICCC papers describe systems with structural coupling, of which 5 only possess a low degree of structural coupling, and of which 13 are manifestations of historical embodiment. While 9 systems implement virtual embodiment to some degree, most of the embodied systems are largely physical (19 systems). Only 4 papers have been identified to present the organismic embodiment type, one of which only minimally.

The authors outline issues that emerge in CC research with regard to embodiment. First, the terminology is often too fuzzy and unspecific to support a proper classification or assessment of the benefits of the system’s embodiment. Second, the authors observe a bias against machine-embodiment, with some papers ascribing embodiment exclusively to humans. Third, some systems possess some human-like features but not actually use them. Guckelsberger et al. [117] refer to this as “under-attribution”.

While this thesis does not fully focus on the exact relationship of embodiment and creativity, it does present different kinds of embod-
iments in the context of a co-creative storytelling performance. Crucially, one of the systems that Guckelsberger et al. [117] survey is the system and its modes presented in Chapter 4 (identified embodiment types in bold-font):

- **Therapist Storyteller** (Sec. 4.3) is an interactive mode within our system. In this mode, a humanoid Nao robot asks questions to an audience in order to collaboratively create a story (see also [125, 128]). The interactivity allows the story to be altered by audience interaction. The users can perturbate the system in order to change the storytelling experience, thus the system exhibits structural coupling. Moreover, the humanoid Nao is a physical, organismoid robot. As the artificial storyteller is able to learn from the knowledge that it gains from the audience, it manifests historic embodiment to some extent. To conclude, the system exhibits **structural coupling**, **physical**, **organismoid embodiment** and some **historical embodiment**.

- **Double Act** (Sec. 4.4) is a mode presented in [1]. In this mode, a Nao robot is coupled with an Amazon Alexa speaker to present a non-interactive performance that can contrast embodied (Nao) versus non-embodied (Alexa) story presentation. The two devices perturbate each other, which provides a manifestation of **structural coupling**. In addition to **physical embodiment**, the Nao employs human-like gestures, which underpins **organismoid embodiment**. However, this mode does not exhibit historical embodiment, i.e. the systems do not incorporate knowledge over time.

- **Story Enactment** (Sec. 4.5) is a mode in which two Nao robots and one Alexa device present a storytelling performance using gestures and space without audience interaction (see also [2, 258]). Since the robots are mostly executing the performance script, only a **low degree of structural coupling** is exhibited. However, the humanoid robots manifest **physical** and **organismoid embodiment**.

- **Interactive Storytelling** (Sec. 4.6) is not included in the review by Guckelsberger et al. [117], but is the fourth mode we can classify according to their taxonomy. This mode features two humanoid Nao robots and one Alexa device. Additionally, the audience can use non-verbal signs (gestures and facial expressions) to alter the course of the presented story in real time [21]. The developing story takes previous decisions into account and therefore manifests **historical embodiment**. The interaction provides **structural coupling** and the agents provide **physical** and **organismoid embodiment**.
In light of this analysis, we can situate this thesis’ contribution in the field of Computational Creativity research with embodied systems. The implemented modes (Chapter 4) and the evaluated performances (Chapter 5) show all but one kind of embodiment, namely organismic embodiment. This label could only be attached three times in ten years of research, which makes it a rather rare and specific type of embodiment. Hence, our implemented systems extend across most of the embodiment spectrum. Just how well they do this will be evaluated in Chapter 5.

2.7 CONCLUSION: LITERATURE REVIEW

Let us now define what we mean by some key terms (bold font) in the context of the existing literature and how they will be used in the rest of the thesis. As noted in Sec. 2.1, the storytelling performance we are trying to model requires the actors to turn the plot into physical action. More precisely, spatial and temporal relations between characters, objects, causal relations and more, need to be expressed in a physical, performative act. Under these requirements, a performer’s goal is to express internal concepts which organize thoughts and perceptions about the story. Since the term conceptual scheme describes the system of concepts which organize thoughts and perceptions [259], we define a performance as a conceptual scheme turned into physical action.

When concepts become movements, movements suggest meanings and meanings evoke concepts in the minds of an audience. Since every link in this chain is under-determined, creativity can insinuate itself into every part of the meaning-making process. The physical actions of a performance suggest meanings, or reinforce what is also communicated with words, so the most effective actions tap into an audience’s sense of familiarity, obviousness and conceptual metaphor. In this work, we consider storytelling as a performance that combines linguistic (spoken) and physical (embodied) actions. For this, we use gestures and spatial schematic movement as embodied signs (Sec. 2.4).

The starting point is a story generator, which has to provide the three layers of fabula, discourse and dialogue. As outlined in Sec. 2.1, access to all of these layers is important and any story-generation system that exhibits all three qualities can be used as the basis for our performative system. The deep structure allows a performer to span a complex performance with mixed modalities of space and gesture, whereas a surface form provides the manuscript to narrate the story. Lastly, the performance should not just be a narration but an embodied enactment, hence we have distinguished between those two terms in Sec. 2.1.2.3. Having reviewed different story-generators
we can summarize this review with respect to our three requirements in Table 3.

Most of the systems in Table 3 provide access to both deep structures and surface forms. As discussed in Sec. 2.1.1.2, the data-driven approaches, of which GPT [3] is a representative example, do not provide access to the deeper structures. A neural language model’s internal structure is a complex vector space and extending the model to respond with gestures for certain action verbs may require the retraining of the entire model.

At the same time, very few systems provide rich or easily extensible surface structures, e.g. BRUTUS [143], PropperWryter [134] or GPT [3]. Lastly, only Scéalextric [50] provides a dedicated dialogue module, a rich extensible surface form and a semantic knowledge
network that best suit our needs (Sec. 2.1.2). Therefore, our research extends Scéalextric’s capabilities to include multiple performative domains such as gestures and movement in space.

The gestures in our embodied performance establish a communicative link to the audience (Sec. 2.4), because they are an important instrument of human language [77, 88, 200] with rich semiotic properties [90, 201, 202, 203] that are embedded in a social, cultural and ideological context [79]. Gestures can be differentiated according to their semiotic properties (Sec. 2.4.1) which make them more than just ornamental moves. Moreover, they exhibit the same image schematic underpinning [16, 80] which we can find in complex linguistic expressions, thoughts and spatial movement (Sec. 2.3.3). These image schemas are the recurring, dynamic pattern of perceptual interactions [8], which we use as source domain when mapping complex plot developments onto schematic physical movement (Sec. 2.5). The case study for this mapping will investigate the image schematic movement NEAR/FAR (Sec. 2.3.3). Here, we will map the relation of the characters from mental space to physical space and let the movement carry additional meaning.

The resulting performance arises from an embodied system (Sec. 2.6) with important features such as physical, organismoid, historical embodiment and structural coupling [117]. As such it situates itself on the CC continuum [142] among other embodied systems (Sec. 2.2.1).
A CONCEPTUAL FRAMEWORK FOR EMBODIED PERFORMANCE WITH ROBOTIC ACTORS

The theoretical contribution of this thesis is the description of a conceptual framework for embodied performance with robotic actors. The framework concerns itself with the semantics of gestures and locomotive movement and provides a taxonomy of robotic movement. Additionally, these theoretical concerns are implemented in a storytelling architecture, which complements the conceptual framework. As such, it connects the research field of gestures and movement studies with robotic performance. The conceptual framework will be presented in a tripartite structure (shown in Figure 18):

• **Performance Framework**: Provides a taxonomy of gestures and movement for performative action. The taxonomy and theoretical considerations are relevant for various types of embodied performances with robotic actors. This chapter lays down the theoretical foundations of the Performance Framework. Ultimately, the framework will be implemented for a storytelling performance (Chapter 4) and an empirical evaluation will assess its merits (Chapter 5).

• **Interpretation Framework**: As opposed to the Performance Framework, the Interpretation Framework goes beyond the connection of movement and meaning and looks at the possible construals of meaning that spatial metaphor and storytelling have to offer. Its further development is the starting point for the discussion of future work in Chapter 6.

• **Integrative Framework**: This last installment integrates the theoretical considerations in various different storytelling modes. Its implementation is described in Chapter 4. The purpose of this integration is the demonstration and evaluation of the theoretical considerations and research questions in an interactive human-robot setup. This framework introduces Scéalability, a scalable, modular architecture for robotic storytelling.

This chapter develops the theory behind each framework. It places the strongest emphasis on the Performance Framework as this guides the empirical evaluation and is the main source of answers to the research questions.
3.1 PERFORMANCE FRAMEWORK FOR ROBOTIC ACTORS

We humans use our bodies to tell jokes, engage in animated conversation, and communicate feelings in play and in dance. Starting from a narrative perspective, with a system designed to support the performance of computer-generated stories with computer-controlled robotic actors, we set out to generalize our approach and create a Performance Framework for embodied communication that can support multiple types of performance. Key to this approach is the meaning-making potential of physical (embodied) acts, which we ground in image-schematic models of language. As we will show here, storytelling provides an ideal basis for empirically responding to our research questions, but our aim is to broaden the framework to accommodate new possibilities and new kinds of performance.

We adopt a bottom-up approach to unifying theory and practice, in which an implemented AI system supports the empirical studies that motivate our hypotheses, before we generalize those hypotheses into a combinatorial framework for embodied meaning-making. We begin by defining a taxonomy that accommodates humanoid movements from walking to posing and gesturing. Although physical actions are not words, deliberate physical actions do have a semiotic component that we will analyse here. So, by exploring robotic enactment in a storytelling context, we can identify the semantic units of movement and their cognitive-linguistic underpinnings in image schemas and conceptual metaphors. Ultimately, our goal is to identify the points of contact between action and meaning where creativity – and in particular, machine creativity – can emerge.

The Performance Framework outlines which movements can be executed in parallel or in series, to convey meanings of their own or to
augment the meaning of the spoken dialogue (Sec. 2.1.2.3). In addition, it considers the properties of physical actions to identify those that are additive (when compounded movements achieve a cumulative effect), persistent (when a movement has a lasting effect on the physical relationship between actors), and summative (allowing an action to summarize what has already occurred). For example, the meaning conveyed by one actor stepping away from another intensifies with each additional step. The action and its meaning are also persistent, since, unlike gestures, stepping away does not necessitate a subsequent retraction of the action. After multiple steps, the resulting distance between actors (and characters) is the sum of all steps, and so conveys a global perspective.

We distinguish between locomotive spatial movements (henceforth just spatial movement) along a stage, postural reorganizations of the body, and gestures made with the hands, arms and upper-body to communicate specific intents. For gestures, we also discriminate pantomimic or iconic gestures (which play-act a meaning, e.g. using an invisible steering wheel to signify driving) from more arbitrary actions (which may use metonymy to depict culturally-specific actions, such as bending the knee to propose), from those that instantiate a conceptual metaphor to achieve their communicative intent. The framework formally integrates each of these forms of physical meaning-making, and constrains how they work with each other in the realization of a coherent performance.

3.1.1 Taxonomy of Robotic Movement

Notwithstanding the range of studies in the related works (Section 2) that feature similar notions of robotic gesture and movement, their methodology does not rely on a shared taxonomy. Based on the literature review, we propose the taxonomy depicted in Figure 19. This taxonomy is the conceptual basis for the Performance Framework and the attributions and labels in Fig. 19 will be explained within this section.

The Performance Framework is applicable to a variety of embodied performance types that include robots, such as dancing, storytelling, joke telling and conversation. While those tasks impose unique requirements for hardware and software, the framework provides a unified conceptual perspective. The next sections present the framework, and explain its terminology and its syntax for describing movement. We start with a technical description of the specific movements that can be derived from the conceptual organization of movement types.

3.1.1.1 Gestures

Where the movement in related works lacks a shared taxonomy, the field of gestures (as reviewed in Sec. 2.4.1) mostly refers to a classifica-
Figure 19: A taxonomy of robotic movements and their properties. Nodes to the left and right depict broad types of robotic movement and their specific sub-types (rotational, spatial, iconic gestures, narrative beats, etc). Shared properties are indicated with horizontal connections (dashed lines).

McNeill [88]. McNeill classifies gestures into five types based on their semiotic properties. As embodied signs (Sec. 2.4), gestures can exhibit a range of different properties. These properties are listed in Table 4, along with illustrative examples of each gesture. We illustrate each gesture in its most iconic form, with the exception of the Beats and Cohesive gestures [88], since these are always specific to the temporal context in which they appear. As noted earlier (Fig. 13), the Iconic and Metaphoric gestures can use the same movements to convey different meanings in different contexts [88].

- **Iconic**: An iconic gesture has an obvious meaning, since an icon can clearly substitute for what it is supposed to represent (see [260] and [261]; the latter provides a thorough linguistic discussion of signs, icons and gestures). These icons of physical actions are schematic by nature, insofar as they enact patterns of embodied experience [261]. We therefore attribute the property of Obviousness to the Iconic gestures, since they make meanings more explicit, and leave little room for alternate interpretations. Table 4 presents an example of a robot steering an invisible vehicle to iconically depict the act of driving. Despite their obvious iconicity, many iconic gestures can be culturally-specific. As discussed in Section 2.4, gestures that are obvious to the speakers of one language may be confusing, misleading and far from obvious to members of a different cultural or linguistic grouping.

- **Deictic**: Since pointing gestures refer to spatial/physical landmarks, we ascribe to Deictic gestures the property Referential. The technical implementation of such a gesture requires a limb, ideally an arm, that can point at the target reference. It is also
beneficial if the pointing is further supported by the head or gaze direction of the robot [262]. As with Iconic gestures, Deictic gestures also overlap with metaphorical gestures in different contexts (e.g. pointing ahead of oneself to signal a future event). Table 4 shows the example of a robot pointing ahead with its arm.

- **Metaphoric:** Metaphoric gestures, labelled Metaphorical in Fig 19, are the most challenging to implement since their intent must be discerned via a mapping from literal to non-literal meanings. Yet, as a consequence of this mapping, metaphorical gestures also open new possibilities for creativity within the system. An example of the creative use of metaphorical gestures has been provided in Figure 13.

- **Cohesives:** Cohesive gestures are dependent on their context of use, and require careful timing. Whether a shaking of the fist, a circling of the finger or a turning of the wrist, such movements only make sense in a narrative if they are used coherently. Coherent usage aids discourse comprehension and allows audiences to construct a spatial story representation [263]. Moreover, Cohesives can strengthen our grasp of the whole narrative if they are used recurrently to reinforce persistent or overarching aspects of the plot. We therefore ascribe the attribute Global to these gestures.

- **Beats:** Beats are just as context-dependent as Cohesives, but lack the latter’s global influence, as they are relevant to one-off events only. Since the movement itself is less relevant than its timing and its context of use, no concrete example is offered in Table 4. In opposition to the Cohesives, we label these gestures as Local. Note that a gesture that is considered a Beat in one task domain, such as story-telling, might serve a global role in the synchronization of movement in another, such as dance. In that case, the gesture would be labelled Global in the latter context.

Since the gesture types illustrated in Table 4 do not constitute an exhaustive list, some additional properties may need to be included in the future. For example, some gestures are performed with two hands and so can exhibit relational properties in the way that each hand, representing a character, relates to the other [264]. We can also consider the naturalness of the gesture, as this is an important property for HRI and a common basis for assessing any computational model that uses gestures [213, 265]. However, we might also view naturalness – as we do here – as an emergent property of the implementation, rather than as a constitutive property of the performance framework itself.
3.1 Performance Framework for Robotic Actors

<table>
<thead>
<tr>
<th>Gesture Type (Example)</th>
<th>Depiction (Example)</th>
<th>Req. Hardware</th>
<th>Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iconic (Drive)</td>
<td><img src="Image" alt="Iconic Gesture" /></td>
<td>Arms, hands</td>
<td>Obviousness</td>
</tr>
<tr>
<td>Deictic (Point)</td>
<td><img src="Image" alt="Deictic Gesture" /></td>
<td>Pointing limb</td>
<td>Relational</td>
</tr>
<tr>
<td>Metaphoric (PUOH)</td>
<td><img src="Image" alt="Metaphoric Gesture" /></td>
<td>Arms, hands</td>
<td>Metaphorical</td>
</tr>
<tr>
<td>Cohesive</td>
<td><img src="Image" alt="Cohesive Gesture" /></td>
<td>Limb</td>
<td>Global</td>
</tr>
<tr>
<td>Beats</td>
<td><img src="Image" alt="Beats Gesture" /></td>
<td>Limb</td>
<td>Local</td>
</tr>
</tbody>
</table>

Table 4: Examples of gestures: Iconic, Deictic, Metaphoric, Cohesive and Beats by a robot. Each gesture is ascribed a general property, along with its hardware requirements.

3.1.1.2 Body Movement

Spatial Movement describes a simple trajectory of an agent along one axis. This core movement requires the agent to possess a means of locomotion, such as wheels or legs. In its basic form, spatial movement in one direction is a transformation of the positional coordinates in one variable:

\[
\vec{x} = (\omega, 0, 0) \quad \text{with } \omega \in \mathbb{R} \tag{1}
\]

This movement is compatible with all other movement types. Combined with rotational movement, it covers all directions on the 2D plane. When the mode of locomotion allows for it, the vector can be positive or negative. This kind of movement can exhibit the following properties:

- **Global**: The moving body affects the relative proximity, shared references and spatial configuration of all agents in a performance, and so has implications for the performance of the narrative as a whole.
• **Relational:** The movement has implications for other agents on the stage since an absolute change in position for one actor also changes its position relative to others.

• **Summative:** The movement of an actor into its resulting position summarizes, in some general sense, the history of past actions up to this point.

• **Additive:** A movement compounds a previous action to achieve a perceptible cumulative effect.

• **Persistent:** A movement has a lasting effect on an actor or its physical relationship to others.

The property **Obviousness** is not attributed here, since actors (robotic or otherwise) can move in space without necessarily conveying meaning. Some movements help a speaker to communicate while being uncommunicative in themselves, as when an actor steps back to maintain balance, or moves their hands in time to their words as they speak. In contrast, it is hard to perceive a rotational movement as unintentional, since rotation carries such an obvious, iconic meaning. Thus, while the property of **Obviousness** is not wholly context-free, it is sufficiently robust across contexts to earn its keep in a performative HRI system.

**Rotational Movement** This movement requires an actor to possess some form of rotational joint, so that it can rotate around one axis. While a humanoid robot can simply turn its head or torso, a complete rotational movement requires full body rotation. In some cases, rotation is only possible in combination with spatial movement. For example, some bipedal robots cannot rotate on the spot, and need to walk in a curve to achieve full rotation (see Fig. 9 for all rotational joints of the Nao robot). This kind of movement can exhibit the following properties:

• **Obviousness:** When the movement achieves the iconic action of turning away from, or turning toward someone else, this iconicity deserves the label **Obviousness**.

• **Relational:** The rotation has implications for other agents on the stage since an absolute change in orientation for one actor changes its relative orientation to other agents.

• **Summative, Additive and Persistent:** These properties hold the same meanings for rotational movements as they do for spatial movements.
### 3.1 Performance Framework for Robotic Actors

#### Table 5: Depiction of the body movements Spatial and Rotational with their corresponding name, depiction with a physical robot, movement vector and properties in the respective columns.

<table>
<thead>
<tr>
<th>Movement</th>
<th>Depiction</th>
<th>Properties</th>
</tr>
</thead>
</table>
| Spatial    | ![Spatial Depiction](image) | • Global  
• Relational  
• Summative  
• Additive  
• Persistent |
| Rotational | ![Rotational Depiction](image) | • Relational  
• Obviousness  
• Summative  
• Additive  
• Persistent |

#### 3.1.2 Combinations of Robotic Movement

Defining the basic types of movement and their properties provides a foundational set of movements that can be implemented for different kinds of robots. Basic movements can be considered primitive actions in a performance system, whose possibility space is the space of their possible combinations. Gestures can be combined with whole body movements (spatial and rotational) to produce complex behaviours. The individual movements themselves are not creative – many are iconic, and highly familiar – but the mapping from narrative to physical action does allow for metaphor and for other creative choices [122]. The example combination provided in Figures 20 and 21 show a forward movement followed by a backward movement, paralleled by a rotational movement during the transition. The resulting performance (see Fig. 21) is the sum of its parts, and fosters audience interpretation of the performer’s behaviour. This is where the properties Summative, Additive and Persistent come into play.

An embodied performance can draw on all available movements and all possible combinations of such. Table 6 presents a combination matrix showing possible combinations, mutually exclusive movement types, and restricted combinations. The group that is least conducive to interaction with others is the Beats. Due to their local property, these are grounded in a specific narrative moment, which does
Figure 20: Combination of spatial and rotational movement in a temporal representation. Spatial movement can co-occur with rotational movement. Different directions can be achieved by combining rotations with positive or negative spatial moves.

Figure 21: The result of parallel spatial and rotational movement is a curved walk. Here $t_0$ is the robot’s position prior to movement. $t_1$ is the point when the robot shifts from forward movement to backward movement, whilst completing half of the 90 degree rotation.
Table 6: Possible combinations of movement types. comb. (green) are combinable movements, restr. (yellow) are combinations that are only possible to some restricted extent and excl. (red) are movements that are mutually exclusive.

not permit metaphorical, iconic or deictic displays. This momentary status also strongly prohibits combinations with Cohesives. Nonetheless, Beats can be performed during spatial or rotational movements, as this does not change their function. In fact, spatial and rotational movements can be combined with all other movement types, as well as with themselves. However, a spatial or rotational movement during an iconic or metaphoric gesture can cause positional changes that affect the gesture, while deictic gestures are also sensitive to any referential changes of position. For example, pointing while walking is a much more restrictive task than either alone, since the target of the reference might move behind the performer.

By definition, iconic and metaphoric gestures exclude each other. As with a change of context, a gesture’s obviousness can be exchanged for a metaphorical interpretation, but a combination of iconic and metaphoric gestures must be sequential, not parallel. Likewise, the Cohesives can be combined with any other movement type except for the Beats, since these groups have opposing global and local properties.

3.1.3 Performance Framework at Work

Figure 22 depicts four example tasks, the requirements of each, and the applicability of the framework to each instance of the task. The framework is designed to meet the demands of these different tasks. When the Software Modules for a task depend on the choice of performing agents (e.g., embodied/non-embodied, single/multiple), the properties needed to support an appropriate conceptual response are given.

Different performance types can place varying emphases on the meanings of any given movement. Dancing is an expressive act which aims to convey themes and emotions through the use of the entire body. While Fig. 22 lists only Body Movement as a necessary requirement for dance, dancing can have other requirements in context, e.g.
Figure 22: Diagram of four possible tasks that can be accounted for by the framework. From left to right, each task poses some prototypical requirements for the corresponding software module, which can be addressed by the performance framework properties.

Single or multiple bodies which can - but do not need to - move synchronously. Dance types can range from the highly coordinated to the highly improvised and relatively uncoordinated. As shown in Fig. 22, rhythm can, for example, be achieved with a repetition of movements. However, while rhythm and synchrony are listed as prototypical requirements of a dance task, these are neither necessary nor sufficient for dance, and this point applies more generally to all performance types, from dance to storytelling to joke-telling and casual conversations. In robotic dance, complex relational movements and motion dynamics are at play, which may or may not exhibit synchrony and rhythm [28, 266, 267]. Within the Performance Framework, additive, relational, local and global properties can be identified, and, in cases where it is required, rhythm can be achieved by adding repeated movements, just as synchronized movement can be realized in terms of global and relational additions. Ultimately, Body Movements is a flexible requirement which should always acknowledge the diversity of bodily capabilities across humans and across robots, making it all the more important that each possible requirement is appropriately integrated on the software level.

For storytelling, the empirical evidence in Chapter 5 will show that movement, gesture and relative positioning play an important role in enacting a story well. These requirements can each be met using movements with persistent, relational, global and additive/summative properties. When actors undergo changes in their physical spaces that mirror the changes undergone by characters in a narrative space, metaphorical schematic movements and gestures can depict plot progression and character interrelationships.

Certain performance types, such as stand-up comedy, place a greater emphasis on timing than others [268]. When timing is key, spatial movements may be subtle and minimal [269], making the properties obvious, relational and global all the more important. For example, the timing needed to land the punch line of a joke requires global and relational knowledge of the performance as a whole, while the use of
iconic gestures throughout can increase the effectiveness of the performance.

Lastly, conversational agents make use of various discourse strategies that can be enhanced by the use of iconic and deictic gestures. The latter are especially useful in maintaining shared attention and awareness, by mirroring movement in a topic space with movement in physical space during a conversation [270].

The Performance Framework has been implemented for storytelling robots in the Integrative Framework which includes a software architecture called Scéalability. Its realization is described in Section 4 Practical Implementation.

3.2 INTERPRETATION FRAMEWORK FOR STORYTELLING ROBOTS

Whether dance, storytelling, theater or comedy, connecting the underlying script with the performance requires an appropriate choice of the movements to be enacted. When a script dictates the actions, there is no space for choice. Likewise, when a script provides simple disjunctive choices – do this or that – it allows a performer to explore the space of possible stories without regard for the emotions of the characters. The actor’s capacity to interpret the script should consider these emotions and how they can shape the performance. By creating choices at the time of performance, an interpretation can look at the unfolding narrative so far and shape the course that the actor will take. Since these choices are guided by the character’s emotional valence at any given moment, we introduce an emotional layer between the level of scripted actions (what happens next) and the expressive level (physical movements and spoken words). This additional layer annotates the emotions implied for each action and movement, to inform the actors about the emotional resonances of their choices. By considering the influence of earlier actions in the plot, choices can be made in the moment, to reflect an interpretation of how characters should be feeling and acting. The Interpretation Framework provides the tools to a performer to make deliberate use of gesture and space for an emotionally-informed performance. While, the following theoretical evaluation is concerned with storytelling, the Interpretation Framework itself can be implemented for any script-based performance with disjunctive choices.

By meaningfully connecting plot actions to movements, the Performance Framework allows a performer to pick its movements to suit the action \( x \) at hand. More formally, \( C(x) \rightarrow E(x) \) denotes the mapping of an action \( x \) as it is represented in the conceptual domain \( C \) of stories to its expressive realization in the embodied, physical world \( E \). For example, the insult action can be expressed with an iconic gesture.
in which an actor “flips the finger” to another actor. That other actor may show that they feel disrespected by moving their head slightly backwards. In this case, $C(\text{insult}) \mapsto E(\text{insult})$ because the actors physically express the insult action that the plot calls for. However, each actor should take into account the current state of the story, and their residual feelings that carry over from earlier actions. If we denote this state of the story as $S$, then the performers consider the mapping $S(C(\text{insult})) \mapsto S(E(\text{insult}))$. In a story space with $x_N$ possible actions, the general form of this mapping is $S(C(x_N)) \mapsto S(E(x_N))$.

Skilled actors are able to interpret an action within the context of the unfolding story, so the complementary Interpretation Framework allows performers to interpret each action in context. Suppose character A has supported B in some way, or confided in B, or defended B, and B then responds by insulting A. Viewed in isolation, the insult should make A feel disrespected, and even a little attacked, so it would be appropriate to embody this event as $C(\text{insult}) \mapsto E(\text{insult})$. In this case $E(\text{insult})$ is the embodiment of an insult (an insulting gesture) performed by B and $E(\text{insulted})$ is the embodiment of the response to an insult (a shocked gesture) performed by A. However, given the earlier events which make this insult all the more shocking, it would be even more appropriate, from A’s perspective, to enact $C(\text{insulted}) \mapsto E(\text{attacked})$, since attacked carries more shock value than insulted. Each performer brings a different interpretation to bear on the same plot action. So while B interprets the insult action directly, A interprets it as attack. The result is a performative blend of the two enactments. B enacts its agent role in the insult while A enacts its patient role in an attack. That is, while B enacts the event via the mapping $S(C(\text{insult})) \mapsto S(E(\text{insult}))$, A uses the mapping $S(C(\text{insulted})) \mapsto S(E(\text{attacked}))$. The more general form of A’s interpretation is $C(x) \mapsto E(\bar{x})$. It is the task of the Interpretation Framework to provide the mapping mechanisms for interpretations such as these.

### 3.2.1 The Representation of Gestures within the Framework

As part of the Integrative Framework in Section 3.3 the work from [128] is presented in which we define one-to-many mappings from plot actions to gestures and movements, from which performers can choose an appropriate but context-free enactment at random. The purpose of the Interpretation Framework is to transform this choice from a purely disjunctive one to a choice based on interpretation in context. To this end, each gesture must be understood by the system as more than a black box motor script. So in addition to duration information, a schematic classification as given in Table 1, and the properties given in Table 4, we must give the framework an emotional basis for making its choices.
Our database of gesture representations is available from a public repository. More than 100 movements are currently stored and labelled for use in embodied storytelling. Each is assigned a unique name that describes the movement briefly. This name aims to be as telling as possible in just a few words, while a longer free-form description is as explicit and detailed as possible. For example, the movement named “strike down” has the more detailed description “right arm squared angle lifted above shoulder, quickly striking down with hand open.” This movement, which takes approximately $2\frac{1}{2}$ seconds to execute, is labelled as a schematic DOWN movement. This motion is not associated with rotational or spatial movement, and its possible uses as an iconic or metaphorical gesture depend on the narrative context in which it is performed.

As noted in [128], the existing disjunctive mapping from plot actions to physical actions is further labelled with an appropriateness label, since some gestures are more obviously suited to their associated plot actions than others. For example, the action disagree_with can map to either of the gestures “shaking the head” or “shaking the head, raising both arms and turning away.” In this case, both gestures are equally appropriate for the action. For another action, however, such as contradict or break_with, the latter is more appropriate than the former. Three distinct appropriateness levels – high, medium and low – are used to qualify the mappings of actions to gestures. This suggests a very practical motivation for metaphor within the system: the mapping $C(x) \mapsto E(\bar{x})$ is preferable to $C(x) \mapsto E(x)$ when $E(\bar{x})$ offers a more appropriate enactment for $x$ than $E(x)$. In general, metaphor will be motivated by a mix of concerns, from the practical (does this action have a vivid enactment that really suits it?) to the expressive (does this action adequately capture the feelings of the moment?). Notice that in each case, however, metaphor depends on questions of expressive adequacy, and the question of whether the systems knows of a better way to communicate what it wants to say.

### 3.2.2 Selective Projection for Creative Interpretation

An embodied performance of a story is a careful presentation of story elements – plot, character, emotion – in a physical space. As such, performers project elements from the story space, a space of words and concepts, into the presentation space, a space of gestures and movements and spoken dialogue. The performers are themselves, with their own physical affordances and limitations, and the characters they play. In the terminology of [249, 271], the performance is a conceptual blend.

Turner and Fauconnier’s Conceptual Blending Theory has previously been used to model stories in a computational setting [19]. The basic

---

1. https://osf.io/e5bn2/?view_only=2e30ee7e715342d59c371b5d30c014e0
3.2 Interpretation Framework for Storytelling Robots

Figure 23: Depending on the narrative context, the action $B$ insult $A$ can cause $A$ to feel "insulted" ($C(x) \rightarrow E(x)$) or to feel "attacked" ($C(x) \rightarrow E(\bar{x})$). $A$’s emotional response will then dictate the actor’s physical reaction on stage.

theory has been extended by Brandt and Brandt [18] to incorporate additional spaces that are especially relevant here, such as a reference space (for the underlying story), a context space (specifying situations within the story, and discourse elements relating to those situations), and a presentation space in which story elements are packaged and prepared for a performance.

Consider again the example story in which character $B$ insults $A$ after $A$ has shown favor to $B$, perhaps by praising, aiding or defending $B$. In the reference space this plot action is literally captured as $B$ insult $A$. As mediated by the context space, however, which brings both situational relevance and a discourse history to the interpretation of events, $A$ views this insult as an attack, and so the action is instead represented in the presentation space as $B$ attacks $A$. Since the performers take their stage directions from the presentation space, $A$ will move, gesture and speak as though the victim of an actual attack. So, when $B$ performs a “giving the finger” gesture to $A$, $A$ will do more than lean back in disappointment – the standard response to an insult – it will step away with its arms extended in a defensive posture. This construal of events by $A$ and $B$, the first of three scenarios unpacked below, is illustrated in Figure 23.

- **Scenario 1**: An insult delivered in some contexts can surprise more, and wound more, than in others. The standard response, which entails a literal mapping from the reference to the presentation space, is $S(C(\text{insult})) \rightarrow S(E(\text{insult}))$. However, in a story state $S$ that makes the insult seem all the more severe,
E(insulted) may feel the same as to be E(attacked), to produce the non-standard mapping $S(C(\text{insulted})) \mapsto E(\text{attacked})$. In that case, it is not the embodied response to an insult, $E(\text{insulted})$ that is enacted by the insult’s target (leaning back, with head down) but $E(\text{attacked})$ (stepping back, arms outstretched defensively).

• **Scenario 2**: A performer whose character, A, praises the work of another, B, might enact a show of “praise” with a clap of the hands or a nod of the head. This is the standard response in a story context where praise is literally interpreted as praise, that is, $S(C(\text{praise})) \mapsto S(E(\text{praise}))$. However, if the context indicates that A has strong grounds to respect and feel inspired by B – perhaps B rescued A in the previous action – then $S(E(\text{praise}))$ may be interpreted in this light to produce a stronger reaction than praise. As such, $S(E(\text{praise}))$ might rise to $E(\text{worship})$ and the performer playing A will bow accordingly.

• **Scenario 3**: A succession of actions that reinforce the same emotional response in a character can lead to a character feeling and expressing that emotion to a higher degree, shifting its embodied response from the standard interpretation to a heightened, metaphorical level. Suppose the story concerns character A treating character B as a lowly minion. A overworks and underpays B, taking advantage of B at every turn. If A should now scold B, B may interpret $S(E(\text{scold}))$ as $E(\text{whip})$, or interpret $S(E(\text{command}))$ as $E(\text{enslave})$, and finally interpret $S(E(\text{fire}))$ as $E(\text{release})$. Interpretative performance allows for a shadow narrative to play out in physical actions as the literal narrative is rendered in speech.

In each scenario, the situated actor uses contextual information to interpret the current plot action, in light of previous actions, and chooses to accept the scripted action ($x \mapsto x$) or to take a metaphorical perspective ($x \mapsto \bar{x}$) instead. An alternate construal, such as construing an insult as an attack, or an act of praise as an act of worship, changes the physical enactment of the action in the script. Notice that when an alternate enactment is chosen, the dialogue associated with the scripted action is still used. The combination of one action’s gestures with another action’s dialogue adds further variety to the blend, while also helping to foster understanding by the audience. Gestures are dramatic on a physical level, but dialogue carries a more explicit semiotic content. Even when the performers choose to be metaphorical, the performance remains grounded in some literal aspects of the script. This grounding is rooted in the assumption that audiences are capable of fully comprehending the narration and dialogue of the script. When this is not the case, gestures and other non-verbal cues become an even more important channel of communication.
Figure 24: Selective projection in a situation in which “being insulted” is presented as “being attacked”. Situational relevance is determined by the emotional valence of the current action. Figure based on context-dependent blending theory by Brandt and Brandt [18], and adapted from Li et al. [19].

A blending interpretation of Scenario 1 is illustrated in Fig. 24, further adapting the treatment of Brandt and Brandt [18] that is offered by Li et al. [19]. Notably, situational relevance is informed by the Emotional Valence of the situation, the calculation of which we consider next.

3.2.3 Emotional Valence in Story Progression

In Section 3.3 Integrative Framework (see also [1]) we show how the actions, characters and structural dynamics of a story can influence the performers’ reactions so as to elicit a comedic effect in a performance by robots. In that approach, the logical structure of the narrative – in particular, whether successive actions are linked by “but” or “then” or “so” – provides a reasonable substitute for an emotional interpretation of the action, so that performers know when to act surprised, or can infer when an audience might be getting bored (e.g., because the plot lacks “but” twists) or confused (e.g., because it has too many “but” twists). To go deeper, we must augment this structural perspective with an emotional perspective, so that performers can grasp why certain actions are linked by a “but” and not a “so.” We begin by situating each role (A and B) of every possible action in a plot (Scéalextric defines more than 800 different actions) on the following four scales:
disappointed ← (A or B) → inspired
repelled ← (A or B) → attracted
attacked ← (A or B) → supported
disrespected ← (A or B) → respected

These emotions are chosen to suit the action inventory of a story-generation system like Scèalextric [50]. Other emotional scales may be added as needed to suit other tasks, such as dance [272]. The storytelling system draws from a knowledge base of over 800 actions, which can be causally connected to create stories that exploit tropes and other common narrative structures. Each story revolves around two central characters and a variety of secondary figures (partners, spouses, friends, etc.). The most common themes elicit feelings of trust, respect, admiration and cooperation about and between those characters.

For example, the insult action associates a strong sense of being disrespected with the B role. When A insults B, we expect B to feel very disrespected (or negatively respected). Similarly, the worship action associates a strong sense of being inspired with the A role, and a strong sense of being respected with the B role. Conversely, the surrender to action associates a negative sense of being attacked (and so a positive sense of being supported) with the B role, because A is no longer an active threat to B. Viewed individually, each emotional setting can be compared to that of the previous action, to determine how much change has been elicited by the current plot turn. It is this change that explains why certain transitions warrant a “but” and others warrant a “so” or a “then.” It can also motivate why an insult can come as a surprise to a character, and feel more like an attack.

The four emotional scales can also be viewed in the aggregate, to determine an overall valence for the current action from a given role’s perspective, or to determine an overall shift in valence from one action to the next. We calculate the valence of a role in an action \( \alpha_i \) as the total valence across all emotional scales for that role in that action. See equations (2) and (3) for the valence of the A and B roles in an action denoted \( \alpha_i \). A positive valence for a role indicates that a character in that role should experience a positive feeling when the action is performed; conversely, a negative valence suggests a negative feeling for the character in that role.

\[
\text{valence}_A(\alpha_i) \leftarrow \text{inspiration}_A(\alpha_i) + \text{attraction}_A(\alpha_i) + \text{support}_A(\alpha_i) + \text{respect}_A(\alpha_i)
\]

(2)

\[
\text{valence}_B(\alpha_i) \leftarrow \text{inspiration}_B(\alpha_i) + \text{attraction}_B(\alpha_i) + \text{support}_B(\alpha_i) + \text{respect}_B(\alpha_i)
\]

(3)

A character is a persistent entity in a narrative, one that moves through the plot from one action to the next. The current valence of a
character is a function of the valence of the role it plays in the current action, and of the valence of its roles in previous actions, with the current action making the greatest contribution. Previous actions have an exponentially decaying effect based on their recency. If $0 < \beta < 1$ specifies the weight given to the current action, the contextual valence of the characters filling the A and B roles is given by equations (4) and (5) respectively. We assume a fixed decay rate, while acknowledging that certain events might have a stronger and more lasting impact on perceived valence than others. It remains to be seen in future work whether this simple one-size-fits-all approach needs to be replaced with a more variable, local solution.

$$\text{context}_A(\alpha_i) \leftarrow \beta \cdot \text{valence}_A(\alpha_i) + (1 - \beta) \cdot \text{context}_A(\alpha_{i-1})$$ (4)

$$\text{context}_B(\alpha_i) \leftarrow \beta \cdot \text{valence}_B(\alpha_i) + (1 - \beta) \cdot \text{context}_B(\alpha_{i-1})$$ (5)

Calculating aggregate valence levels for the characters in a story allows the interpretation framework to track their changing emotions to each other over time, at least on a gross level. Although it is highly reductive, this gross level allows performers to distil complex emotions into simple but expressive physical actions. Because they are calculated as a function of the valence of current and past actions, these levels are both summative and persistent, and thus well-suited to making decisions regarding summative and persistent physical actions in a performance. If a significant increase in positive valence for a character A is interpreted as a result of actions involving character B, then performer A can move a step closer to performer B in physical space. Conversely, a significant decrease can cause A to move a step away from B. This increase or decrease for A is given by equation (6). The same spatial/emotional calculus applies to B’s perspective, as given in equation (7). In each case, a significant increase or decrease is determined to be a positive or negative change that exceeds a fixed threshold $\tau$. In this way, the relative spatial movements of performers on stage are not explicitly indicated by the script (in the sense of a manuscript, see 2.1), or directly associated with the actions in the plot, but determined by each performer’s evolving interpretation of the narrative context.

$$\Delta_A(\alpha_i) \leftarrow \text{context}_A(\alpha_i) - \text{context}_A(\alpha_{i-1})$$ (6)

$$\Delta_B(\alpha_i) \leftarrow \text{context}_B(\alpha_i) - \text{context}_B(\alpha_{i-1})$$ (7)

The emotional valence of an action for a character, much like a character’s “inertial” contextual valence, is derived from four emotionally
charged scales that have been chosen to suit our system’s inventory of 800 plot actions. New parallel scales can be added, or existing ones removed or replaced, if this inventory were to change significantly. One obvious omission is an arousal scale \([273]\), to show the degree to which an action either calms or arouses a particular role. Arousal is not a charged dimension – for one can be as aroused by hate as by love – and so it does not contribute to our calculations of valence. Nonetheless, an arousal dimension is useful for indicating the scale of an actor’s response. A high-arousal action may demand a bigger and more dramatic physical response than a calming, low-arousal event. For that reason, it makes sense to add a new scale as follows:

\[
\text{calmed} \quad \leftarrow \quad (A \text{ or } B) \quad \rightarrow \quad \text{aroused}
\]

A state of high-arousal can be conveyed with a sweeping, high-energy gesture. while a calm state might be conveyed with a slow movement or a slight gesture. Of course, the robot platform may not support the distinction between high- and low-energy motions. The extent to which it does, or does not, indicates the extent to which an arousal dimension is worthwhile in a storytelling context. Still, we may find that arousal is wrapped up with the question of contextual valence and how quickly the influence of context should decay. If arousal can be shown to influence the rate of decay, it can be a valuable addition to the framework whatever robot platform is used.

3.2.4 Opportunities for Metaphor and Irony

Valence and arousal are key elements of suspense \([274]\). They enable us to keep listeners on the edge of their seats before arousing them with a sudden plot twist. At key points, the inertial valence of a character’s perspective can flip from positive to negative, or vice versa. When a substantial shift in inertial valence occurs at the current action, it will have been brewing for some time. When this shift exceeds a fixed threshold \(\Delta\), an actor might mark the shift with an exaggerated response that goes beyond the script. Fig. 25, which tracks the inertial valence of characters that fill the A (blue line) and B (orange line) roles in a Scéalextric script of 31 actions, highlights shifts that support metaphor and irony. The full script is available in Appendix C.

Some of these shifts are marked by dotted circles in Fig. 25. The first circle (I) at action #10 offers the opportunity for a metaphoric and ironic response, because the preceding positive plot elements are followed by a very negative plot development. A metaphorical response is appropriate when a character’s inertial valence is greater (by \(\Delta\) or more) than its specific valence in the current action, and there is a need to dramatize this lagging emotional load. This shift offers the opportunity for selective projection and the dramatization
Figure 25: Interpretative decision points in a script with 31 plot actions (all actions listed in Appendix C). Dotted circles mark dramatic shifts of inertial valence, which enable interpretative decision points. The blue line tracks the inertial valence of character A; the orange line tracks that of B.

is achieved by enacting a different but similar action with a valence (for that character) that is closer to its inertial valence in context. For instance, at plot point #10 in Fig. 25, an actor might react with gestures that suggest a character is despised rather than merely resented as the former gesture conveys a stronger feeling of disrespect than the latter.

Irony is a valid response when the fall in a character’s inertial valence exceeds the threshold $\Delta$, signifying a failure of expectations. Irony is also supported when an action is explicitly tagged with a failed expectation such as a decrease in respect when respect is warranted. A negative shift in respect is equivalent in scalar terms to a positive shift in disrespect, yet the former does more than convey disrespect; it also captures a failure to be as respected as one expects. An actor can now make the interpretative choice to convey the surprising lack of respect with a gesture that instead implies an increase of respect, while relying on narration and dialogue to make the actual disrespect clear. This is the example provided in situation I of Fig. 25: Plot point #10 signals the resentment of character B towards character A. Here, character B could enact a gesture of respect (e.g. bowing down towards character A) whilst verbally insulting character A. Similarly, situation II at plot element #14 allows for a selective projection in which underestimate can become laugh at. Before the valence reaches a new peak in plot element #20, situation III allows the action marry (#19) to be enacted with much greater attraction with a gesture of love. Lastly, situation IV supports the third scenario in Sec. 3.2.2 where inspiration can be exaggerated from a bowing gesture to a gesture of worship.

We can estimate the opportunities for the performers to make interpretative choices by analyzing 10,000 generated stories with a setting
of $\Delta = 2$ (so a character’s inertial valence must shift by 2 points or more to enable irony or metaphor). In this sample of 10,000 stories, 47,532 interpretative choice points are identified: 34,810 support metaphor, and 22,352 support irony. At this setting, a story has 4 to 5 opportunities for interpretative choice; 3 to 4 for metaphor and 1 or 2 for irony. We can adjust the $\Delta$ parameter to allow for more or less interpretation by the actors. Thus, when $\Delta = 1$ the average story provides 10 opportunities for interpretative choice, 8 to 9 for metaphor and 3 to 4 for irony (the two are not mutually exclusive). When $\Delta = 3$, the average story offers just 1 or 2 opportunities for interpretative departures from the script.

3.3 INTEGRATIVE FRAMEWORK FOR ROBOTIC PERFORMERS

The Performance Framework (Sec. 3.1) and the Interpretation Framework (Sec. 3.2) represent the two passive frameworks. In the sense that the performance and interpretation happen autonomously without interference from, or interaction with, an audience. Nonetheless, a robotic performance’s merits should be judged by its impact on the audience. To assess this, several different interactive modes have been implemented, which have all been incorporated in one Integrative Framework called Scéalability. Its structure will now be presented before its practical implementation is described in Chapter 4. The practical implementation will allow us to assess different aspects of the conceptual framework and thereby answer the research questions through a series of studies presented in the Empirical Evaluation (Chapter 5).

3.3.1 The Scéalability Architecture

Scéalability is a storytelling system that has the Scéalextric [50] story generator at its core. The latter generates the stories that the former’s cast of embodied agents will turn into a theatrical performance. Scéalability is both multi-modal and multi-agent. It treats storytelling as a performance to be shared by cooperating but otherwise independent agents, in the same space but perhaps in different modalities. As a scalable, modular architecture, it supports various configurations and scenarios, from a single robot (Nao) or a single smart speaker (Echo/Alexa) working alone, to a double-act of these devices, to an arbitrary combination of multiple robots and speakers. Backstage coordination amongst devices is provided by a blackboard structure [107, 180]. Although agents speak to be overheard as part of the performance, their real communication takes place invisibly, on the blackboard. This coordination allows devices to play different but complementary roles in a performance. Some narrate, others act, while others comment on this narration, or this acting, or indeed, on the story
Table 7: Configurations of Scéalability result in interactive/non-interactive, embodied/non-embodied and single/multiple agent implementations.

- **Therapist Storyteller**  
  [embodied \ interactive \ single agent]
  This first instantiation of Scéalability was presented in [125, 128]. It features a single, humanoid robot (Nao) with different settings for interactive storytelling. This mode presents a minimal way in which to combine a story generator with a humanoid robot that interacts with an audience to craft stories in a guided search. The practical implementation presents technical issues itself. Despite all this coordination, some residual friction between agents is desirable if it produces humour in the performance [1].

In effect, the system comprises two kinds of modules, or models: domain models and meta-models (see Sec. 4.4.3). The Scéalextric generator is the foundational domain model: it defines the semantics of story actions and the causal relationships to each other. Various meta-models harness this domain content for specific “acts”. For instance, a single-act with the Nao may engage in a dialogue with the audience [1, 125], since it has no other device to interact with. Double acts use their meta-model to synchronize their actions with each other [1], while a triple-act of two Naos and one Alexa needs a more complex meta-model still.

Each of the various meta-models comes with different interactive affordances. For example, a single embodied storyteller is more of a narrator who talks about or sometimes has to enact multiple characters. On the contrary, a group of embodied and non-embodied robots can leave the task of narration to non-embodied devices and the enactment of multiple characters can be taken on by multiple embodied robots. The research questions (Sec. 1.4) are addressed by the different configurations of Scéalability and subsequent empirical evaluation. These different configurations, called modes, aim to compare embodied vs. non-embodied, single vs. multiple, schematically moving vs. pantomimic moving robotic actors. An overview of the different modes and their contribution to the research questions is depicted in Table 7 and outlined below:

- **Therapist Storyteller**
  - [Related Work: [125, 128]]
  - Embodiment: humanoid robot (Nao)
  - No. Agents: 1
  - Results: Interactive Settings

<table>
<thead>
<tr>
<th>Name</th>
<th>Related Work</th>
<th>Embodiment</th>
<th>No. Agents</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Therapist Storyteller</td>
<td>[125, 128]</td>
<td>humanoid robot (Nao)</td>
<td>1</td>
<td>Interactive Settings</td>
</tr>
<tr>
<td>Double Act</td>
<td>[1]</td>
<td>humanoid robot (Nao), smart speaker (Alexa)</td>
<td>2</td>
<td>Empirical evaluation</td>
</tr>
<tr>
<td>Story Enactment</td>
<td>[2, 238, 275]</td>
<td>humanoid robot (Nao), smart speaker (Alexa)</td>
<td>3</td>
<td>Empirical evaluation</td>
</tr>
<tr>
<td>Interactive Storytelling</td>
<td>[21]</td>
<td>humanoid robot (Nao), smart speaker (Alexa), webcam</td>
<td>3</td>
<td>Practical implementation</td>
</tr>
</tbody>
</table>
and shows how a multi-modal approach to storytelling can incorporate different layers of performance, and thus provides preliminary answers to RQ1 and RQ2 (see 1.4).

- **Double Act**

  [embodied vs. non-embodied | non-interactive | multiple agents]

  The Double Act [1] considers the pairing of two creative systems in the same thematic area, a speech-based, non-embodied storyteller (with Alexa) and an embodied, gesticulating storyteller (using a Nao robot). Working together, these two compensate for each other’s weaknesses while creating something of comedic value that neither has on its own. This instantiation contrasts embodied vs. non-embodied performances and its empirical evaluation thus provides further answers to RQ2 (see 1.4).

- **Story Enactment**

  [embodied & non-embodied | non-interactive | multiple agents]

  This instantiation has been explored in detail in [2, 258, 275]. The Story Enactment features three agents, two Nao s and one Alexa smart speaker. The two embodied agents present a story narrated by the smart speaker in a non-interactive fashion. This version of the robotic performance emphasises the movement of the robots and their use of gestures. Investigating the spatial schematics and gestural appreciation in an empirical study, this configuration provides answers to RQ3 and RQ4.1 (see 1.4).

- **Interactive Storytelling**

  [embodied & non-embodied | interactive | multiple agents]

  Bringing everything together, multiple robots narrate and act out a machine-generated story whose plots can be dynamically altered in response to non-verbal audience feedback. The enactment and interaction focuses on gestures and facial expression, and draws on cognitive-linguistic insights to enrich the storytelling experience. With a focus on emotions and the non-verbal behaviour of the audience, this provides answers to RQ4.1 (see 1.4) and presents opportunities to expand Scéalability to include dynamic audience feedback.

The Integrative Framework produces internal and external interactions. Internal interactions are those between the modules of the Scéalability framework, which allows intra-modular coordination to manifest different types of performances. Those performances pose technical challenges when confronted with a practical implementation.
(Chapter 4). The external interactions are those between the performing agents and the audience. Both types of interactions can be observed more or less clearly in the resulting performances and as such can be judged and evaluated using HRI questionnaires. In return, these empirical evaluations (Sec. 5) can again inform the implementation and answer the research questions posed in Section 1.4. It is now time to consider the practical implementation of these issues in more depth.
PRACTICAL IMPLEMENTATION

4.1 IDEAS, DEPTH AND STRUCTURE

The principle purpose of the practical implementation is to ground the conceptual and theoretical work of this thesis in the physical world. First, the strong emphasis on embodied performances creates the need to test hypotheses and frameworks within an actual, physical embodiment (robots). Second, the affordances, limits and challenges of physical robots are different to those of purely theoretical concerns and can foster novel insights. Lastly, an argument about movement, gestures and space necessitates a practical, physical implementation if the research questions are to be addressed appropriately. Consequently, the following implementation provides all of the performances that are needed for the Empirical Evaluation in Chapter 5.

Our performing robots include two Nao robots from SoftBank with slightly varying model versions. Both run with the same NAOqi (Version 2.1.4.13) operating system. Yet, some functionalities are only available to the newer model with the latest NAOqi version. The NAOqi interfaces that control our robots require Python 2.7 as a programming language. At the same time, most of the integrated software, e.g. the blackboard, neural networks, speech-synthesis, etc., relies on code in Python 3.+. Moreover, the connection between all performing robots is dependent on an intranet connection specific to the experimental lab (at UCD’s Computer Science department) although this setup has been reproduced at conferences with live performances. Due to these complex dependencies between hardware and software, this Chapter places the focus on utility and usefulness rather than on code\(^1\), for the benefit of other researchers that may work with similar storytelling robots, gestures and movements. Algorithms are provided in an abstract form, as are any databases\(^2\) that can be used independently of the underlying hardware.

The structure of this Chapter moves from a basic implementation of a single storyteller towards a multi-agent, multi-modal model of robots enacting and narrating a story. Each addition will be justified with respect to the main research questions and its usefulness to the empirical evaluation to follow in Chapter 5. We will begin with the inner workings of the Scéalability implementation and then present

\(^1\) An excerpt of the code is publicly available at: https://github.com/PhilWicke/RobotStorytelling

\(^2\) Available at: https://osf.io/e5bn2/?view_only=2e30ee7e715342d59c371b5d30c014e0
In order for Scéalability to support a variety of storytelling modes, a modular approach has been chosen, which allows a variable interplay of gestures, dialogue, plotting and narration. A blueprint of the system’s architecture is presented in Fig. 26 (see also [2]). Both, the Dialogue Module and the Scéalextric story generator have been introduced in Sec. 2.1.2.3 and Sec. 2.1.2 respectively. Thus, the important additions are the Gesture Module and the Spatial Module. In Fig. 26, Scéalextric populates the blackboard with a story, characters and plot elements (actions). From those elements, the Dialogue Module creates a script for the performance. The script includes speech acts which are directly forwarded to the devices. Iconic or spatial movement and reaction values accompanying the dialogue are passed onto the Gesture and Spatial Module in order to select the movement procedures which will be passed on to the devices.

4.2.1 The Gesture Module (GM)

The Gesture Module takes care of the gestures of the actors in line with the conceptual framework. In particular, this module allows to create conditions which can be used to answer RQ2.1: What role do gestures play in performances?

We can think of gestures as embodied speech-acts, and so this module works in much the same way as the dialogue model. For any narrative action involving an embodied character, a corresponding gesture may be retrieved for the actor to perform. The model indi-
cates whether the gesture, if any, is to be executed before or after any speech act. For example, for the action “A propose to B”, the model suggests that the actor for A execute a bended_knee gesture before any proposal is uttered. Most Scéalextric actions are linked to one or more gestures, from either the Nao’s built-in repertoire or a set of 18 new additions that are added by Scéalability [125]. Two examples are shown in Fig. 10.

4.2.1.1 Incorporating Pre-Existing Gestures

Previous related research with storytelling Nao robots and gestures differ from the current implementation 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 Scéalability 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 and various spatial moves 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 own repertoire and associated each of these gestures with plot verbs from the Scéalextric system. Ultimately, however, 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 established strong, medium and weak associations to one or more Scéalextric verbs. Overall, 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 construct a plot, 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. 27.

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 Event-related Potential (ERP) study by Habets et al. [276] 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 Csapo et al. [172] uses a very small

3 The knowledge base of associations is available in the research repository at: https://osf.io/e5bn2/?view_only=2e30ee7e715342d59c371b5d30c014e0
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 embodies. In [88] 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 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.

Even when an action is not associated with a gesture, the model may still suggest one based on the choice of speech-act. If a speech-act ends with a question mark, the actor may execute a quizzical gesture (e.g. the robot is scratching its head); if it ends with an exclamation point, it may execute a gesture that indicates a show of...
resolve. Alternately, the affective nature of the action can be used to suggest an apt gesture. The affect of an action is derived from Scéalextric knowledge base as described in Sec. 2.1.2.2 and not related to the four emotional scales of Sec. 3.2.3. By using the positive/negative valence of the action to reason about the emotional state of the character, the gesture model can map that state onto a corresponding gesture. As shown in Table 8, the model uses a 10-point scale to rate the affect of an action on a participant. For example, the action “A is kidnapped by B” ranks as -4 for A, while “A promote B” ranks as +3 for B. The gestures for different points on the scale do not capture the specifics of the actions, but they do capture some of the emotions involved.

4.2.2 The Spatial Module (SM)

The Spatial Module takes care of the locomotive movement of the actors. In particular, this module allows us to create conditions which can be used to answer RQ2.2: What role does the use of space play in performances? And RQ3.1: How do spatial schemas support performances?

The robots in an embodied narrative do not need to do very much to tap into the image-schematic intuitions of an audience. When robots portray distinct characters, their movements can be interpreted figuratively, as the movements of their characters in the abstract spaces of emotion or social relationships. Even simple back-and-forth motion can convey nuance in an embodied performance. Each actor can move toward the other (as in marriage); or one can move forward as the other remains stationary (as in one-sided actions, such as trust and pursue); or both can move away at the same time (as in divorce) or one can move back as the other remains stationary (as in disgust and distrust).

Suppose robot A says “I’d like to get to know you,” and makes a flirtatious gesture toward robot B. Most gestures are culturally determined, so gestures like this are more reliable when they are rooted in

<table>
<thead>
<tr>
<th>Value</th>
<th>Movement</th>
<th>Example act</th>
<th>Value</th>
<th>Movement</th>
<th>Example act</th>
</tr>
</thead>
<tbody>
<tr>
<td>-5</td>
<td>Both arms raised</td>
<td>beaten, attacked</td>
<td>+5</td>
<td>Nodding</td>
<td>open up to, appointed by</td>
</tr>
<tr>
<td>-4</td>
<td>Both arms crossed</td>
<td>disappointed, threatened</td>
<td>+4</td>
<td>Nodding, Applauding</td>
<td>supported, favored</td>
</tr>
<tr>
<td>-3</td>
<td>Shoulders sank</td>
<td>underappreciated, spied on</td>
<td>+3</td>
<td>Nodding, Applauding</td>
<td>rescued, are loved by</td>
</tr>
<tr>
<td>-2</td>
<td>Shaking head</td>
<td>run away from, failed to impress</td>
<td>+2</td>
<td>Raising both arms</td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td>Shaking head</td>
<td>dismiss, lose faith in</td>
<td>+1</td>
<td>Raising both arms</td>
<td></td>
</tr>
</tbody>
</table>

Table 8: A 10-point valence scale is used to rate how an action affects a character emotionally. Apt gestures are associated with points on this scale.
image-schematic intuitions. If robot A moves closer to B as it gestures and says its lines, the meaning of all three signals – dialogue, gesture and spatial movement – becomes clearer. Table 9 presents three use-cases in this vein: one fosters emotional closeness; another increases emotional distance; and a third preserves the status quo.

<table>
<thead>
<tr>
<th>Speech Act</th>
<th>Spatial Metaphor</th>
<th>Spatial Move</th>
</tr>
</thead>
<tbody>
<tr>
<td>“I would like to get to know you”</td>
<td>emotional closeness</td>
<td>A → B</td>
</tr>
<tr>
<td>“I do not enjoy your company”</td>
<td>emotional distance</td>
<td>A ← B</td>
</tr>
<tr>
<td>“Did you read the news today?”</td>
<td>no emotional effect</td>
<td>A ............... B</td>
</tr>
</tbody>
</table>

Table 9: Three speech acts and the corresponding uses of space by the performer.

The dialogue and gesture models generate communicative events that punctuate the narrative but do not persist. Once a gesture or a speech-act has been articulated, it is done, and the story moves on. Spatial movements, however, do persist, insofar as they have a cumulative impact on the relative position and orientation of the embodied actors. At any point in the narrative, the position of the actors is a function of all the spatial movements that have occurred thus far. This spatial summary of the narrative is independent of any cultural and linguistic assumptions that are made by the DM and the GM (e.g. see [202]).

Each embodied actor maintains a record of its local spatial movements, while the blackboard maintains a global view, to ensure the robots do not crash into each other on stage. Employing a subset of conceptual scaffolding defined in [9], any story verb in Scéalability with a spatial dimension – whether literal or figurative – is coded with the corresponding on-stage movement.

Following Pope et al. [277], we define the personal space of a robot with respect to its step size. Figure 11 shows Hall’s definition of public, social, personal and intimate spaces [14]. The robot’s gestures are also adapted in this radial fashion, since we restrain our gestures accordingly when we occupy the intimate space of another. Figure 12 illustrates the imaginary rings that define the spatial proximity of a robot actor. Each ring is defined step-wise, with the robot storing its position relative to its starting point in the innermost ring. If two robots are just one ring-distance apart, they confine themselves to subtle gestures. When they are two ring-distances apart, they may use any regular gesture.

### 4.2.2.1 Schematic Gestures

When locomotive Spatial Movement is enabled in Scéalability, and iconic, pantomimic gestures are disabled, we still need the robot actor to perform some minimal gestures in order to provide a basic movement...
### Table 10: The 9 chosen image schemas and their corresponding gestures.

If a schema requires a direction along an axis, two gestures are provided, otherwise the same gesture is given in both columns. S-P-G is the Source-Path-Goal image schema (see Sec. 2.5).

<table>
<thead>
<tr>
<th>Image Schema</th>
<th>(Direction) Gesture</th>
<th>(Inverse Direction) Gesture</th>
</tr>
</thead>
<tbody>
<tr>
<td>UP-DOWN</td>
<td>(UP) Both arms up</td>
<td>(DOWN) Both arms/torso lower</td>
</tr>
<tr>
<td>FRONT-BACK</td>
<td>(FRONT) Both arms stretching to front</td>
<td>(BACK) both arms moving back</td>
</tr>
<tr>
<td>CENTRE-PERIPHERY</td>
<td>(CENTER) Hands unite from sides</td>
<td>(PERIP) Hands depart from centre</td>
</tr>
<tr>
<td>CONTACT</td>
<td>(SAME) Both hands touch in front</td>
<td>(SAME) Both hands touch in front</td>
</tr>
<tr>
<td>IN-OUT</td>
<td>(IN) arm-1 encloses, arm-2 reaches in</td>
<td>(OUT) arm-1 encloses, arm-2 reaches out</td>
</tr>
<tr>
<td>SURFACE</td>
<td>(SAME) Both hands circle horizontally</td>
<td>(SAME) Both hands circle horizontally</td>
</tr>
<tr>
<td>CYCLE</td>
<td>(SAME) arm-1 moves vertically 3 times</td>
<td>(SAME) arm-1 moves vertically 3 times</td>
</tr>
<tr>
<td>S-P-G</td>
<td>(SAME) arm-1 fixed front, arm-2 unites</td>
<td>(FAR) arm-1 fixed front, arm-2 departs</td>
</tr>
</tbody>
</table>

for the arms of the robot. For this, we choose the set of image schemas as explained in Sec. 2.5 (see Fig. 4). We only need a small set of image schemas in order to express a variety of different actions, hence we choose only those schemas that a Nao robot can easily perform. Consequently, we propose the selection of schemas outlined in Table 10.

Table 10 provides an overview of nine image schemas. Of these, five are sensitive to direction, meaning that a corresponding inverse gesture along the opposite direction also exists. For example, for the UP-DOWN schema both UP and DOWN have their own distinct gestures, though each can be seen as a directional inflection of the other. In all then, the nine schemas give rise to fourteen distinct gestures. The association between gesture and image schema builds on the evidence provided in other studies [218, 219] and its validity for our framework will be the subject of further investigation. To connect these image schemas to the storytelling system, we next consider the database of actions on which the storytelling system is built.

We must tie these schemas and their gestures to specific actions in the repertoire of *Scéalextric*, since its approach is explicitly action-centred. This mapping of gestures to actions was done in the absence of any mediating abstraction, and was thus opportunistically ad-hoc. However, we can now use image schemas as the motivating abstraction to tie gestures to story actions when iconic/pantomimic gestures are disabled. Given the causal underpinnings of *Scéalextric*, this will subsequently allow us to look for recurring causal patterns between the image schemas themselves.

Since these schematic gestures are active only for those conditions in which iconic, pantomimic gestures are disabled during empirical evaluation, we refer to [278] for a more detailed analysis of these schematic gestures.
4.2 Scéalability

Figure 28: Scene from a performance. The left robot (A) proposes to the right robot (B). In the background, an Amazon Alexa device narrates the action.

Figure 29: Schematic depiction of the spatial map that both robots use to signify their character relations. The robots will not cross the “+3” line in order to avoid bumping into one other. Moreover, their gestures become more subtle when the distance between robots is less than 3 units.

4.2.3 Evaluative Aspects

There are two dimensions of physical embodiment at the core of this implementation. One is the mapping of iconic gestures to story verbs by the Gesture Module. For example, Fig. 28 shows robot A performing the iconic “bending of the knee” associated with the action propose to. We will empirically validate that coherent gesture mappings are appreciated over incoherent ones in Chapter 5.

Another is the spatial movement between robots as guided by the Spatial Module, which maps character dynamics into physical space. Each robot knows its position relative to centre-stage, and can move 3 steps back or forth. Each coordinates with the other to avoid collisions (see Fig. 29). To prevent accidents and respect proxemics, subtle gestures (with no wild swings) are preferred when the robots are within 3 steps of each other. The empirical studies will show that coherent, schematic uses of space add as much as dramatic gestures to the appreciation of a story, and significantly more than random or incoherent uses.

The following sections will explain the four Scéalability configurations that have been implemented in order to address the research questions.
This implementation presents two modes of user-influenced story-generation. The first, a baseline, employs simple user interaction prior the generation of a story, while the other conducts a probing, therapist-like interview to elicit personal experiences from the listener that it can repackage – somewhat collaboratively – into novel stories.

4.3.1 Baseline Mode

The baseline mode of interactive storytelling employs minimum engagement with the user and explores the space of stories, whilst exploiting various gestural possibilities. Storytelling is initiated in this mode with a request to the user: “Please provide an action around which to build a story.” Any of Scéalextric’s 800+ action verbs may be offered in response by the user, as the Nao’s speech recognizer is primed with the corresponding words. Low-level engineering challenges include loading this vocabulary in a parallel thread and preventing robotic stutters during the loading phase.

A vocabulary of 800+ verbs diminishes the reliability with which the robot can correctly distinguish words, as e.g. the words “look” may be confused for “cook” or “pay” for “pray.” Fortunately the Nao’s acoustic module reports a confidence value for each word that it recognizes. Only when this confidence is above an upper threshold (0.65) does the robot accept the user’s response without question. When the confidence falls below a lower threshold (0.4) the robot remains in listening mode; only when the confidence falls between thresholds does the robot signal its uncertainty and seek explicit user confirmation with a “yes” or “no.”

Once the robot has obtained an action from the user, the story-generation around that pivot can proceed. Building on the representations of [152], the system selects its start and end points in the story forest to represent a meaningful character-development arc. Stories with the preferred number of actions (e.g., 6 to 10) are then generated by traversing the story forest between these end-points and retaining only those pathways/plots that contain the desired action. Of these matching pathways, the system selects one that provides maximal opportunity for gestural expression. Consider this example:

- Nao: This story is about Isolde the Loyalist and Tristan the Despot.
- Nao: Isolde swore loyalty to trustworthy Tristan.
- Nao: Isolde gave Tristan Tristan’s full support.
• Nao: But Tristan took full advantage of impressionable Isolde.
• Nao: Tristan intimidated others with threats of violence from Isolde.
• Nao: Thuggish Tristan threatened Isolde with violence.
• Nao: Isolde considered Tristan a disgusting monstrosity
• Nao: so Isolde’s feelings for trustful Tristan soured.
• Nao: As a result, Isolde sold out Tristan to Tristan’s enemies.

Each story begins by introducing a pair of characters which have been selected from a comprehensive database of familiar faces, real and fictional, called the NOC list (see [55]; access the list at github.com/prosecconetwork). The sequence of actions that have been selected is rendered with apt logical connectives (and, so, but, etc.) to convey the teller’s understanding of causality.

In this baseline mode, the user is given only one opportunity – a single action – to influence the direction of a story. An example video of this mode can be found at: bit.ly/2GGORhc. More opportunities for collaborative development and self-expression are offered by the interview-shaped interaction of Therapist mode, as we will see next.

4.3.2 Therapist Mode

This mode is named Therapist mode because it frames its interactions with the user around a series of probing questions about the user’s own life experiences. These questions mirror the possibilities of the robot’s story space because they are built from an interrogative rendering of the same plot actions. The guiding intuition is that everyone has a story in them; the role of a good therapist is to bring this tale to light by collaboratively reassembling the user’s own experiences.

The story’s actions should reflect the individual experiences while perhaps changing the sequence and filling gaps, thereby offering a new perspective on the user’s past behaviour. Potential story actions are rendered as yes-no questions (so, e.g., are_inspired_by is rendered as “Have you ever been inspired by someone?”), and an affirmative answer adds that action to the story’s plot. Consider this example:

1. Nao: Let us craft a story. I will start to ask you some questions. You only have to answer with yes or no.
2. Nao: Have you ever felt inspired by someone?
3. User: Yes.
4. Nao: But have you ever taken a hard look at someone who inspired you?
Figure 30: Example of a tree traversal for the story generation process. Grey nodes and edges indicate that a path has not been explored by the system. Red nodes and edges denotes those that have been answered with a “No” by the user. Green nodes have been validated by the user, giving the sequence Action 01, Action 02, Action 03 and Action 04 that start the story. An orange node is one that has been rejected by a direct question, but accepted by the user to be a hypothetical possibility.

5. User: Correct.
6. Nao: Have you ever been criticized by that person?
8. Nao: Can you imagine this is possible?
10. Nao: Have you ever felt censured by that person?

The initial story verb is chosen at random from the robot’s repertoire. If an affirmative response is given, the verb provides the next action in the story, and the potential follow-ons to that action suggest the next set of questions. If a negative response is given, and asserted again, the robot jumps to another branch in the story forest, and tries to move forward from there. We want the robot’s stories to build on the user’s experiences but to expand upon them too, to suggest what might have happened if events had taken a different turn. Thus, in line 6 the robot asks whether the user was ever criticized by a role model. Given a negative response, the robot presses on, asking instead if the user can conceive of this possibility. If the user now replies affirmatively, the story can incorporate this sequence of events (inspiration followed by criticism). To further allow for stories that go beyond the specific facts, users are also encouraged to reply with “maybe.” As the questions asked of the user will differ from
session to session, a different – yet highly personal – story will result each time.

Each story involves two characters, a protagonist A (the user, or “you”) and an antagonist B. Notice how the questions above relate each action back to the previous action by assuming the antagonist to be common to both. It is likely that the user will have multiple antagonists in mind when answering questions, and the antagonist presumed in answer 3 is not the antagonist presumed by answer 13. The questions are sufficiently general to allow this artistic freedom to operate, so that the robot can create stories that merge several people from the user’s life into a single thought-provoking antagonist. Once the session ends at the user’s request, the selected actions can now be woven into a two-character plot:

This is the story about you and a pioneer. This spectacular pioneer became a shining inspiration for you. You kept the pioneer under close observation. You mimicked the popular pioneer’s style and adopted it as your own. “You’ve let me down” said the Pioneer plaintively, so the domineering pioneer gave you a very public rap on the knuckles, and to conclude, the pioneer brought suit against you in open court. That’s the end of your story.

This story features all of the actions that the user has assented to, and may include additional actions as well to conclude the protagonist’s arc (in the sense of [152]) and bring the tale to a satisfying conclusion. Notably, the protagonist of the story is addressed as “you”, while the protagonist is referred to not by name but by character-type (here, the pioneer, as this is the kind of character that inspires others). The resulting story diverges from the user’s own experiences, but in doing so sheds new light on them.

A schematic view of the generation-by-interrogation process is provided in Figure 30. Here we can see that the first selected action (Action 01) is translated into a question form and posed to the user by the robot. The node is selected (green) if the user validates this experience, or assents to the hypothetical, and this selection cues the node’s children as possible follow-ons. However, if the user negates the question and its rephrasing as a hypothetical, the node and its action are blacklisted so that it cannot be asked again of the user. Instead, the robot selects another sibling and another branch to explore. In the example the node Action 03 has been validated as a meaningful hypothetical by the user.

This mode differs significantly from the baseline in the amount of interaction it demands from the user. However, this interaction is not deterministic, and the user’s answers merely suggest, rather than dictate, the robot’s path through the story forest. This mode demonstrates that every story can offer a probing interrogation of
one’s own experiences, and vice versa too. An example video of this mode can be found at: bit.ly/2ouZbY.

4.3.3 Evaluative Aspects

The Therapist Mode is the first implementation, which tests the various challenges involved in embodied storytelling. It probes the capabilities of the Nao robot and how well the story generator can be connected with movement and gestures. As such, its implementation provides some preliminary answers for \textbf{RQ1}: What layers of performance can be combined for an automated performance? This research also reveals that manually annotated gestures can support a pantomimic approach to embodied storytelling, but this does not suffice to investigate the role of space and movement in performances. Furthermore, \textbf{RQ2} (What can embodiment add to a performance?) can better be answered if we contrast an embodied agent with a non-embodied agent. Hence, the next mode adds a non-embodied agent to form the Double Act.

4.4 DOUBLE ACT

This implementation introduces a non-embodied agent to the storytelling performance, in which a Nao robot and an Amazon Echo are used to implement a storytelling double act (see Figure 31). In this mode their complementary strengths and weaknesses are exploited to make a virtue of failings that might be intolerable in one alone. The next section presents a storytelling skill for the Echo’s speech-driven Alexa front-end, before an embodied, Nao-based robot storyteller, for the same space of computer-generated stories, is described. As part of \textit{Scéalability}, the story space is again built using \textit{Scéalextric}. In this mode \textit{Scéalextric} imposes a top-down shape on its plots and supports the generation of narratives with more than two characters (see Sec. 2.1.2.2). These tales are performed by a double-act of Alexa and a Nao robot, in which Alexa narrates a tale as the Nao embodies its actions. Coordinating their interactions via \textit{Scéalability’s} blackboard architecture obviates the need for any overt communication. Nonetheless, we focus here on the ways in which their joint performance is built upon the interplay of the spoken and the physical.
4.4 Double Act

Figure 31: Double Act of Nao and Amazon Alexa on an Echo device.

4.4.1 Storytelling with Alexa

We can contrast the embodiment provided by the Nao robot with that of the less embodied Amazon Alexa. This interactive smart speaker takes speech input and provides its responses with a high quality synthesised voice. As a consumer-oriented product, it can answer questions, order products online, control music or lighting and more. Through Amazon’s developer platform, these functionalities can be extended with additional “skills”. The implementation of a storytelling-skill allows us to use Alexa as a companion for the Nao. This implementation is part of the double act presented by Veale et al. [1]. The software engineering behind Alexa’s storytelling skill\(^4\) has been implemented by Mildner [279].

The skill can be initiated with the voice prompt “Alexa, open the narrator.” The user can then choose the theme of the story with prompts like “Alexa, tell me a story about love” or “Alexa, tell me a Star Wars story.” Internally, the skill has access to hundreds of thousands of pre-generated Scéalextric stories and to the NOC list. Therefore, the chosen theme may correspond to any of the hundreds of themes from the NOC list. Mildner [279] discusses some of the technical obstacles

\(^4\) A description of the skill and its code is available at: https://github.com/ThomasMildner/Scealability/tree/master/01_Masterskill
that come with implementing an Alexa skill. First, the developer interface can seem counter-intuitive from a conventional programmer’s perspective. Second, communication with Alexa during the execution of the storytelling-skill is fragile, since reentering of a skill requires the specific “Alexa, open the narrator.” prompt. When Alexa is supposed to work with another embodied robot, this point of reentrancy requires a high degree of synchronisation. Third, every prompt that Alexa is supposed to receive (e.g. “continue”, “go on”, “then what?”) must occur within a short time window, otherwise Alexa will exit the storytelling skill.

For long stories that rely on two actors, Nao and Alexa, the source of error can be on both sides, as when one robot speaks over the other, a voice-prompt is unheard, or Alexa unexpectedly exits the skill. At the same time, the shared blackboard in Scéalability can flag an error in the performance space which can be remedied “behind the curtain” by both robotic actors. For example, when Alexa has missed a prompt and exits the skill, potentially ruining the performance, this provides an opportunity for the other robot to insert comic relief and restart the skill by offering a snappy remark that prompts Alexa to reenter the skill.

4.4.2 Adding Nao to the Double Act

Alexa has a voice but no body. The Nao has both a body and a voice, but the limitations of the latter often struggle to transcend the former. Although the Nao’s capacity for physical movement is a major selling point, its gestures can be so noisy as to dominate its vocalizations. Moreover, Nao’s processing of speech is limited in comparison to Alexa’s, and frequently forces its human users to repeat themselves on even short commands. The limitations of Nao are further addressed in Sec. 4.7. A pairing of Alexa and Nao makes sound technical sense for a language-based task like storytelling, since Nao’s utility as an embodied storyteller has already been demonstrated by Ham et al. [4], Gelin et al. [11], Pelachaud et al. [13] and our Therapist Storyteller [125, 128].

4.4.3 Domain Model and Meta Model

Veale et al. [1] make a conceptual distinction between the domain and the meta model in the double act. The domain model represents and models the world of the story. In contrast, the meta-model manages the actors’ behaviour when telling the story. While the domain model is responsible for character-to-character relations, the meta-model is responsible for actor-to-audience and actor-to-actor relations [1]. This meta-model description is likely unique to this specific system, but meta level descriptions within storytelling systems
are not new. For example, Pérez y Pérez [280] describes how a storyteller must track character-to-character relations. Indeed, character-to-character relations are especially important to any schematic movement that maps mental closeness to physical closeness. As far as the domain model is concerned, the audience does not exist, and is irrelevant to the world within the story. In contrast, the meta-model only works when it acknowledges the presence of the audience or the interactions of the actors relative to each other. The meta-model can allow an actor to change its role and break out of the domain model whenever the situation is appropriate. For example, the actor can break the fourth wall (i.e. by directly addressing and/or acknowledging the audience) to become a narrator or even an audience member themselves.

Another example of meta-models in storytelling is provided by Montfort’s Curveship [148] (discussed in Sec. 2.1.1). Curveship can modulate the dynamic of the storytelling by changing the perspective from which the story is told or by changing the style in which the story is narrated. Just as the communication between robots to align their meta and domain models is done via the blackboard, Montfort et al. [107] also use a blackboard when they integrate a storytelling system with a metaphor generator.

With respect to our Double Act mode, we must define how the domain and meta-models are applied. Some parts of the performance are better enacted by the embodied Nao, while the verbal narration is best performed by Alexa. As such, both agents need separate meta-models and domain models. This separation allows us to tailor each model to the capabilities and strengths of each performer. For example, it allows us to delegate movement and gestures to Nao and speech acts and narration to Alexa. Ultimately, both performers are part of the same act, hence they need to share a unifying meta-model to support their coordination. This coordination allows us to remedy the potential issues with Alexa’s reentering of skill commands.

It is worth noting that in addition to the Nao’s physical affordances for pantomime, it also offers some support for vocal mimicry. So while its built-in voice is of low quality, the robot permits programmers to upload arbitrary sound files and recordings, making the use of 3rd-party voice synthesis tools a viable option. This option is used whenever Nao must communicate directly with Alexa and have its voice prompts understood as commands, since Alexa does not react to the Nao’s normal speaking voice. It can also be used to associate a different speaking voice with different meta-model functions, from making comments about the current story to making fun of the audience. A key use of this ability is the coordination of meta-models. The Alexa narrator articulates each beat of the story before waiting for the Nao to respond in an embodied fashion. Since neither knows
how long the other will take, they use conversational markers to align their own private models.

4.4.4 Implementing the Double Act

Choreography is needed to align the actions of partners to ensure that they read from the same script while staying in sync from one beat to the next. For a given beat it is impractical for one to infer the timing of another, as a Nao cannot reliably infer how long it will take Alexa to speak the text of a beat, just as Alexa cannot know how long the Nao may take to enact it. If our duo is not to become hopelessly co-dependent, a third system needs to manage their coordination. This third system is Scéalability, the ideal architecture for synchronizing the cooperative strangers of a distributed system. We use Scéalability’s blackboard to store key elements of the domain- and meta-models of the performers, as well as their current positions in each.

The double act is initiated by a command to Alexa, such as “Alexa, tell me a story about Donald Trump.” So it is the responsibility of Alexa to retrieve an apt tale from her story space, as already pre-generated using the augmentations to Scéalextric described above. Each story is fully rendered as text when retrieved, and Alexa segments it into a sequence of individual story beats of one action apiece. It is this sequence that is placed on the blackboard for Nao to see. In the dance of Alexa and Nao, Alexa leads and Nao follows. Alexa starts the tale by articulating the text of the first beat, then waits for Nao to respond. The robot, seeing the cued beat on the blackboard, reacts appropriately, either with a pantomime action for the plot verb, or with a gesture that signifies its response to the story so far. But Alexa does not proceed with the story until she is given an explicit vocal command to do so, e.g., “continue”, “go on”, “then what” or “tell me more.” This can come from the audience, but Nao will provide it itself if none is forthcoming. When it replies to Alexa, the robot looks down at the Echo device to maintain the social contact of a double act. Both agents are engaged in a back-and-forth conversation, and it should show.

This baseline conversation uses only the domain models. But as more substance is added to the meta-models of each partner, sophisticated artifice is possible. So Nao can peek at the next story beat on the blackboard, and determine its causal relation to the last. It can then use this to choose its cue to Alexa to proceed with the tale. Suppose the next beat is “But Donald spurned Hillary’s advances”. Seeing the but, Nao can prompt Alexa to go on by asking “But then what?” In this way a single initiative task becomes a mixed initiative task, in which Nao draws the tale out of its companion, and seems to shape it as it is spun. As Nao uses pre-recorded sound cues for these interactions (recall that Alexa does not understand Nao’s built-in voice),
Figure 32: Blackboard logic for the system’s meta-models. Idea and conceptualisation of the blackboard logic has been developed jointly by Veale et al. [1]. Graphical rendering by Thomas Mildner, co-author of [1].
it can use sound effects here as well as high quality voice recordings, to give the interactions a greater social dimension.

An integrated depiction of the double-act’s meta-models is shown in Figure 32 (and explained in more detail in [1]). A key responsibility of a meta-model is to predict an audience’s response to an unfolding story and allow performers to take elaborative action as needed. Suppose Alexa articulates three successive story beats that begin with then, so, or and. A meta-model may see this as characteristic of a flat stretch in a story in which one action leads predictably, and boringly, to the next, and so spur the robot to reply with a structural reaction, such as a yawn.

If Nao peeks ahead to see that the current flat stretch is about to lead to a “but” it can announce “I see a but coming.” Alternatively, the robot might reply with laughter when a silly act is described, or, more insightfully, when a character gets what they deserve. The robot can also pass remarks on characters as they are introduced into the story, by querying the NOC list for relevant qualities. So it may, for instance, say that “Donald Trump is so arrogant” when that character is introduced for the first time. Each meta-model may also be capable of its own small acts of creativity. For instance, the meta-model can generate dynamic epithets for characters as they evolve in a tale, such as Hillary the Death-bringer, Bill the Seducer, or Donald the Lie-Teller. These epithets can be the robot’s spoken contribution to the plot delivery. So the meta-model allows performers to switch from narrator to actor to Greek chorus as the story context demands.

The joint meta-model of Fig. 32 supports the following reactions to a tale as it is told: gestural reactions (the Nao makes an appropriate gesture for a given action); character reactions (Nao or Alexa react in an apt fashion whenever a character is introduced); structural reactions (Nao reacts to the logical shape of the tale); emotional reactions (Nao reacts with emotion to a plot turn that is highly positive or negative); and evaluative reactions (Nao or Alexa react to their cumulative impression of a story so far, if this opinion is sufficiently positive or negative to be worthy of remark). Since our content model is Scéalextric, a symbolic CC system, all stories have predictable markers that allow our meta-models to be implemented as rule-based systems. The next section illustrates the reaction of the meta-models within an annotated transcript of our double act in action.

4.4.5 The Double Act in Action

As the only embodied agent in the duo, it is the responsibility of the robot to create the duo’s shared physical space. Nao must address itself to Alexa to present their interactions as a conversation, and not just a pairing of devices that speak past each other in a synchronized manner. To begin with, Nao asks Alexa to “open your notebook” so
they can create a story together. Alexa then asks Nao for a subject, which it provides (such as “Star Wars”) and Alexa reacts by noting her satisfaction with the story to come. This tale then unfolds, beat by beat, with Nao asking Alexa to continue between beats once it has enacted its own reply. But Alexa has responsibilities too, and must do in words what it cannot do with physical acts. Alexa must acknowledge the robot’s contributions to show that they do indeed share the same space. For when one agent acts as an audience to the other, they can collectively shape our feelings for the tale.

Alexa’s weaknesses have been well-documented in other work (e.g., Kapadia et al. [281]), and her recovery mode is not sufficiently transparent to avoid failed interactions with the Nao. So if Alexa does not receive her next prompt in a timely manner, she will reiterate several requests for input before eventually quitting the narrator skill. Unfortunately, this reiteration cannot be unpacked so as to tell the blackboard of a failed interaction, so Nao will remain unaware of the failure. However, the robot can detect a long pause in a conversation and prompt Alexa to continue. With each of its reprompts, the robot’s exasperation grows, until it finally tells Alexa to “reopen your notebook” and continue. At this point, the narration continues from the last beat that was processed. What follows is a sample transcript of the double act in action. All speech is presented in italics, while each gesture and meta-model reaction is annotated in square brackets:

Nao: Alexa, open your notebook.
Alexa: Please provide a topic for our story.
Nao: How about Harry Potter?
Alexa: I have written a great story about Harry Potter.
Nao: Let’s hear it.
Alexa: What if Harry fell in love with Hermione Granger?
Nao: Sounds promising.
Alexa: At first, Harry’s thoughts of Hermione were of love.
Nao: Proceed. [Love Gesture]
Alexa: I didn’t get that. [Failed Interaction]
Nao: Please go on. [Reprompt] [Waiting Gesture]
Alexa: Because Hermione openly favored Harry.
Nao: Tell me more.
Alexa: Hey, Bubblehead! We need another character.
[Introduction of new character]
Nao: Lizzy Bennet works for me. [New character Reaction]
Alexa: Yet Hermione governed the land with Lizzy Bennet.
Nao: Oh damn. [Structural Reaction]
Alexa: Should I continue? [Reprompt]
Nao: Continue. Is there more? [Waiting Gesture]
Alexa: But Harry spurned Lizzy Bennet’s advances.
Nao: Oh my goodness! [Structural Reaction]
Alexa: So Lizzy Bennet took a whip to Harry’s back.
Nao: Please go on.
Alexa: Well, Harry complained to Hermione.
Nao: But then what? [Structural prediction]
Alexa: But Lizzy Bennet spread slander about Harry.
Nao: That is interesting. [Spread Gesture]
Alexa: Yet Hermione's suspicions about Lizzy grew.
Nao: That escalated quickly! [Structural Reaction]
Alexa: Are you still there? [Reprompt after brief failure]
Nao: Tell me more. [Waiting Gesture]
Alexa: Then Hermione banished Lizzy from the kingdom.
Nao: I am entertained. [Evaluative Reaction]
Alexa: But Lizzy was not as dumb as Hermione imagined.
Nao: Oh my goodness! [Structural Reaction]
Alexa: So Lizzy unseated Hermione from her perch.
Nao: Then what? [Flat Structural Reaction]
Alexa: Then Lizzy Bennet had Hermione locked up.
Nao: So then what? [Locked-Up Gesture]
Alexa: So Harry's denunciation of Lizzy was scathing.
Nao: Please go on. [Humiliated Gesture]
Alexa: Well, Lizzy turned Harry into an indentured slave.
Nao: What a terrible person. [Emotional Reaction]
Alexa: But Harry escaped from under Lizzy Bennet's nose.
Nao: I am surprised! [Structural Reaction]
Alexa: Now expect to be poisoned with bad cooking!
Nao: Bridget Jones, a great add! [Character Reaction]
Alexa: Meanwhile Bridget guarded Hermione closely.
Nao: Cool, tell me more. [Guarding Gesture]
Alexa: So Harry's wiles proved too strong for Bridget.
Nao: And then what next? [Flat Structural Reaction]
Alexa: Then Harry put poison in Bridget Jones's cup.
Nao: That took a dark turn. [Emotional Reaction]
Alexa: After this, Harry gave Hermione her freedom.
Nao: I love to hear that. [Emotional Reaction]
Alexa: So Hermione fomented rebellion against Lizzy.
Nao: And then what? [Rebellious Gesture]
Alexa: [Interaction failure. Alexa exits story skill.]
Nao: Continue. [Reprompts after delay]
Nao: Are you there? [Reprompts Alexa again]
Nao: Hello, anyone home? [Emotional Reaction to Alexa]
Nao: Alexa, reopen your notebook. [Reopens skill]
Alexa: Then Hermione usurped all of Lizzy's power.
Nao: So what? [Grasping Gesture]
Alexa: So Hermione turned Lizzy into an indentured slave.
Nao: That took another dark turn. [Locked-Up Gesture]
Alexa: In the end, Harry walked Hermione down the aisle.
Nao: *That’s just lovely.* [Emotional Reaction]
Alexa: *The end.*

A single transcript can be revealing about specifics, but not about the generic tendencies of a system. To appreciate the latter, we simulated the double act $136,787$ times, choosing a different story to tell each time. Our goal was to estimate the relative occurrence of alternate meta-model reactions to the story in each case. In particular, we considered the following: the BUT structural reaction to a turn in the plot; the BORED evaluative reaction to a predictable stretch of plot; the STRONG emotional reaction to a highly-charged plot verb; the GOOD evaluative reaction to an exciting stretch; the NEW character reaction to the introduction of another named entity to a story; and the GESTURE reaction, which delivers a mimetic response to a given plot action. Overall, the BORED evaluative reaction accounts for $18.4\%$ of all reactions, the BUT structural reaction accounts for $16.6\%$, the STRONG emotional reaction accounts for $15.5\%$, the NEW character reaction accounts for $7.7\%$, and the GOOD evaluative reaction accounts for $4\%$. In all remaining cases, or $37.8\%$ of the time, the Nao responds structurally, with a prompt to “continue” or “go on” and a downward glance at the Echo unit by its side. The GESTURE reaction is independent of these other reaction types, since the robot can make a gesture and utter a spoken response in a single turn. For $49.6\%$ of story beats the robot performs a gesture that is visually mimetic of the current plot verb; for the other $50.4\%$ of beats, Nao makes a “holding” gesture - such as folding its arms, putting its hands on its hips, or shifting its weight from one leg to another - in the manner of human listeners who wish to emphasize their physical presence.

### 4.4.6 Evaluative Aspects

The main contribution of the *Double Act* is the introduction of a non-embodied agent (Alexa). This implementation allows to contrast the embodied agent (Nao) and the non-embodied agent against each other, while highlighting their differences and individual strengths. At the same time, this expands the single storyteller of the *Therapist* mode by another device. Hence, an empirical evaluation of the double act will show that the pairing of two or more devices enhances the user’s experience of storytelling. Chapter 5 shows that an embodied agent contributes more to a performance than a non-embodied (voice-only) agent. Specifically, the evaluation provides evidence that a mobile anthropomorphic agent like the Nao adds an extra dimension to the audience’s appreciation of the performance. Moreover, the enactment of a story as an embodied performance is rated higher than the story as a spoken text. As such, the *Double Act* and its evaluation in the next Chapter directly address research question **RQ2**: What can embodiment add to a performance?
This implementation fully instantiates *Scéalability* with two Nao robots and one Alexa, and so makes use of all of *Scéalability*’s modules. It introduces the use of interpersonal space through the spatial module (see Sec. 4.2.2) and an additional Nao robot. Now, both Nao robots embody the characters of the story, while Alexa simply narrates it.

A representative performance resulting from the implementation can be viewed at: [youtu.be/d4xNMFXTZLU](http://youtu.be/d4xNMFXTZLU). The script of this excerpt follows the format first presented in Section 2.1.2.3:

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

*(two robots)*

N: Say hello to Hillary Clinton.
A:[waving]
N: And let us welcome Donald Trump.
B:[waving]

*(on black-screen)*

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.

fiction

< A propose to B >
A: [gesture: move_closer] “It’s time we took our relationship to the next level. Will you marry me?”

*(on black-screen)*

N: But Donald felt a deep love for Melania. So Donald turned a cold eye to Hillary’s entreaties. Well, Hillary took Donald hostage ...

fiction

< A release B >
B: [gesture: back-away] “Just let me go.”
A: “I will release you.”
N: Thereafter Hillary would say that it was the other way around: that it was Hillary who dumped Donald!

4.5.1 Evaluative Aspects

This multi-agent, embodied interpersonal mode using spatial movement and gestures is the most relevant mode for the empirical evaluation. Its use of gestures, spatial movement and narration allows us to design experimental conditions which contrast spatial movement and pantomimic movements. These conditions allow us to test whether the coherent use of spatial movement and gestures is more appreciated than incoherent use. As such, this implementation will answer RQ$_3$, by providing conditions which show that audiences do appreciate the use of spatial movement in an embodied performance, and prefer coherent over incoherent uses of space (whether metaphorical or literal) to convey the impact of story events.

4.6 Interactive Storytelling

With recent advances in language generation [3, 151], writing stories is no longer left to the creativity of human writers, and can exploit promising results by machines [282, 283]. However, there is more to storytelling than text generation, and the enactment of a tale through movement and gesture facilitates a greater engagement with the plot [2]. Having an embodied agent, more specifically, a robotic one, can elevate a creative AI system from rule-bound generation to creativity by providing live-feedback and interaction [17].

Other implementations in this Chapter make use of minimal interactivity that relies on simple vocal prompts. With prompts like “Yes” and “No”, a user is able to shape the plot to be performed, by essentially engaging in a co-creative search in a dense graph of branching story lines (see Therapist Storyteller 4.3). In this more interactive implementation (also described in [21]), we aim to further engage the user in the telling, and make them feel greater responsibility for plot turns as the tale is told. Specifically, we shall look at two modalities, gesture and facial expression, which have previously been studied in the context of storytelling with robots [4, 11, 13], and interface those modalities with the interacting human.

In much the same way that Emperor Commodus dictates the fate of the gladiator Maximus in the movie Gladiator (see Fig. 33), we want
users to be able to decide the fate of protagonists during a robotic performance. To achieve this, we combine two neural network models: One to classify hand gestures, another to recognize facial emotions. The former identifies diverse hand shapes, but we focus here on the detection of *verso pollice*, the turned thumb, while the latter suggests the sentiment of the action by recognizing emotions on the user’s face. As an example, Fig. 33 depicts a cartoon of actor Joaquin Phoenix raising his right arm and showing a thumbs up gesture. An overlay of lines depicts the tracked joints and a boundary box. The latter tracks facial emotions and annotates the actor as mostly neutral but somewhat sad. This combination of non-verbal signals can prompt the system to alter its story-line in mid-telling, with the robots reacting accordingly. While robotic storytelling is not new [100, 116] and neither the use of robotic or human gestures for interaction [284, 285], the hallmark of the robot movements used in this work is their schematic foundation. Recall, an *insult* action increases emotional distance, prompting the robots to move further apart [275]. When spatial motions and pantomimic gestures are coherently used to mirror the plot, audiences show greater appreciation for a tale and its telling [2]. Gestures by audience members should be equally schematic, even if they exploit a vertical spatial metaphor (up=good, down=bad) rather than a lateral one (closer=positive, farther=negative).

This schematic underpinning is described in greater detail in the system description of the next section.
Figure 34: Two Nao robots act out a story using space and gesture. A display with a webcam provides live-feedback to users while tracking their emotions and gestures. An Echo device narrates the tale.

4.6.1 System Implementation

This implementation extends the Scéalability storytelling framework [2] with two neural networks: one to recognize hand shapes and another to recognize facial expressions in live images of a human user. These neural models are incorporated into the system to permit dynamic interaction with a user, who is presented with the setup shown in Fig. 34 & 35. Photographic recordings are depicted, both of which show the two robots standing to the left and right of a screen. The screen shows the user and their joints being tracked by a webcam. Figure 34 depicts the start of the performance with descriptions of the individual devices. Figure 35 shows the user on the screen during an interaction where a “thumbs up” and a smile have been detected by the system. At the start of a performance, the actors (two Nao robots) are introduced in character, by the smart speaker (Amazon Echo, with Alexa). The story-generating system generates a plot and furnishes it with stage directions, narrative and dialogue, and sets up several decision points at which the user may influence the outcome via gestures and facial expressions. Consequently, the robots use their own gestures, spatial movements and speech to act out their parts. For a demonstration of the system in action, please watch this online video: https://youtu.be/Xe8gaa0YJXQ.

4.6.2 Inserting Interaction Points

Two Nao robots embody the main characters of the story, while an Echo device narrates the plot. The novelty of the current setup over
the previous modes resides in the interaction of story events. Every story can be regarded as a thread of actions. When a decisive action is reached, one in which a character needs to make a decision, the robot will pause, turn towards the audience and ask for a sign as to what it should decide. While the robot speaks directly to the audience, a user can provide a thumbs up or a thumbs down (just like Emperor Commodus), with a corresponding facial expression. Users need not react at all, but if they do, their inputs decide the robot’s next action and the plot is regenerated from that point. The evaluation of user responses is explained in Section 4.6.3.3.

This functionality exploits a feature of the Scéalextric story-generator, which can generate branching plot-structures as well as linear structures (see Sec. 2.1.2). In the case of a branching plot, the plot can follow one of two directions after a Boolean decision point. It is at this decision point that Scéalextric inserts its interactions with the audience. As shown by [51], each story is mostly self-correcting. When user decisions force a detour, the plot still rejoins its original narrative arc before the conclusion is reached.

4.6.3 The Neural Network Models

This section briefly revisits the pair of neural classifiers that allow the system to recognise hand gestures and facial expressions in an interactive performance.
4.6.3.1 Hand Shape Recognition

The system tracks the user’s pose in real time using the OpenPose\(^5\) framework for Python [286], which provides 135 keypoints on single images for multiple joint positions (torso, head, legs, arms, hands and face). The keypoints of the hands [287] become the inputs to a neural network that is trained to classify different hand shapes\(^6\). The model classifies nine different hand shapes (with an accuracy of 96%), of which only the thumbs up and down gestures are used by the system. Although gestures are culturally dependent, the thumbs up gesture is commonly understood as “good” or “positive” across cultures [288]. Figure 33 shows how OpenPose provides a colored skeleton of the limbs, head and fingers, even when parts of the arm are occluded. The hand shape model classifies the finger and displays the label on screen (“Thumbs up” in the example).

Whenever a thumbs up or thumbs down is detected in the camera-input, a marker is recorded for the next decision point in the storytelling process. The absence of a hand signal is deemed a neutral response. The blackboard architecture of the Scéalability framework, which allows backstage coordination between the robot actors and the Echo device, also allows the cameras and classifiers to be integrated into the system as yet another information source for storytelling.

4.6.3.2 Facial Expression Recognition

An additional channel of sentiment that accompanies the hand shape is provided by facial expression, so that an angry thumbs up carries a different meaning than a surprised one. The Python package FER\(^7\) is used for Facial Expression Recognition [289]. This identifies faces with the MTCNN face detector [290] and uses a convolutional network to classify expressions [291] by emotion, whether Angry, Happy, Sad, Surprise, Fear, Disgust or Neutral. Arriaga et al. [291] report human-level performance for their network architecture.

Figure 33 shows the boundary box for the face of emperor Commodus, as drawn by the MTCNN face detector, while FER displays percentage scores for the most salient emotions. A confidence threshold of 80% is imposed to reliably identify the most salient expressions for audience feedback: Angry, Happy and Surprise. These three prove to be the most robust classes in our setup. This information is then combined with the classified hand shapes to derive a meaningful signal from the user to guide the story.

---

5 https://github.com/CMU-Perceptual-Computing-Lab/openpose
6 https://github.com/Fasko/Hand-Gesture-Recognition Credit is due to Michael Fasko Jr and Jacob Calfee for allowing free access to their hand shape recognition model
7 https://github.com/justinshenk/fer Credit is due to Justin Shenk for the implementation of FER
Table 11: Signal encoding examples of hand gestures and facial expressions for decisive questions during storytelling.

### 4.6.3.3 Signal Evaluation and Enactment

The detected hand signals (thumbs up, thumbs down, neutral/none) and facial expressions (happy, angry, surprise and neutral) can form twelve (3x4) combinations, such as “enthusiastic thumbs up” when a thumbs up and a happy smile is detected. The robot re-articulates the combination that it detects and alters the plot if a narrative detour is necessitated. For example, the robot in the linked video asks “Should I give my heart to this person or not?” When the user provides a thumbs up and smiles, this is interpreted as an “enthusiastic thumbs up”, and the narrative is adapted accordingly. Each gesture is interpreted in the context of the robot’s specific request, so a thumbs down in response to the question “Should I turn down this offer” is interpreted as a yes (turn it down), while a thumbs up to the suggestion of a marriage proposal is also understood as a yes (do propose). A knowledge-base connects signals and event actions to appropriate decisions. Examples of decisive actions and the signal valences (summed for hands and face) that motivate them are provided in Table 11 for responses to robot questions of the form “Shall I ...”. If the robot asks “Shall I lay a trap for ... ”, a frown is not necessarily negative, but can express Schadenfreude about the deed. Likewise, the thumbs down has a positive valence for the question “Shall I kill ... ” since it puts the user in the same position as emperor Commodus. Given this context-specificity, the signal-interpretation matrix requires fine-tuning and evaluation through additional studies and experiments.

### 4.6.4 Evaluative Aspects

This mode implements an interactive robotic storytelling system that integrates non-verbal cues from both the performers (robots) and the audience (human users). The focus is on the naturalness of these
cues, which are grounded in pantomime (iconic actions that are the gestural equivalent of idioms and clichés) and spatial metaphor (in particular, schematic movement that has an emotional interpretation). An evaluation of the relative merits of pantomime and metaphor for robot performance has been conducted for the Story Enactment mode and is presented in the next Chapter.

The novelty of this implementation lies in its heterogeneous combination of cognitive linguistic models of space, symbolic AI approaches to story generation, robotic models of performance, and neural network models of visual signal detection. This combination is expedient but relatively deep, since each element must ultimately connect at the plot level [54]. Gestures and movements and user signals must all integrate with an explicit sense of what is happening in the tale, which the system represents at both a surface text level (narration and dialogue) and a deep semantic level.

In an important sense, therefore, the system benefits from its heterogeneity. Although neural language models generate fluent textual narratives [3, 151], the plot-driven use of space and gesture necessitates access to the symbolic deep-structures of the narrative that are not apparent at the surface level. Hence, this mode addresses RQ4: How do the layers of performance interact with each other and the audience? Specifically, the inclusion of emotion within the interaction between audience and performing robots through facial emotion recognition adds a new layer to the mix and presents an answer to RQ4.1.

4.7 NAO ROBOT AS STORYTELLER

Despite the embodied performances presented in this chapter being applicable for any form of embodied robot with similar affordances and hardware, there are some specific technical challenges related to the Nao robot. Since the Nao is an off-the-shelf consumer-grade robot, it has been used for studies in psychology, sociology and linguistics (see Sec. 1.3.3). Hence, a consideration of the technical challenges posed by Nao and possible solutions is warranted.

4.7.1 Technical Solutions

4.7.1.1 Speech Recognition

The Nao’s speech recognition software can respond to pre-assigned trigger words. There is no practical limit on the size of the trigger vocabulary, but even a few thousand words require an onerous loading time and slow 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
4.7 Nao robot as storyteller

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 when the system expects the user to reply with a single 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 can be achieved by framing the interaction using yes-and-no questions (as demonstrated in the Therapist Storyteller). We empirically determine the threshold for trigger recognition to be $p(\text{targetWord}) > 0.6$.

4.7.1.2 Text-To-Speech

Nao’s software package (NAOqi) 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, each text string is processed so as to access each embellishment prior to its output. We also shorten the pause between sentences to create more fluency and momentum in the telling of a story. At the same time, storytelling modes that include Amazon Echo create a strong contrast between the speech quality of the devices. Since the Nao’s speech synthesis falls short of Alexa’s, Google’s WaveNet is used for Text-To-Speech synthesis in real-time [292]. Scéalability provides information about the gender of each character, and this is used to inflect the spoken output accordingly.

4.7.1.3 Creaky Joints

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. A subtitle feature in our framework has thus been added to the framework to pipe the output of the TTS module onto a screen. As shown in in Fig. 36, the audience is thus able to read the robot’s verbal output in large-print in real-time.

4.7.1.4 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 the core storytelling 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 eye color does not contribute much to the comprehension of its outputs and is consequently disregarded [293].

4.7.2 Why Choose the Nao?

Qualities that are desirable from an interpretative perspective may be undesirable from a performance perspective, and vice versa. For instance, a decision to link excited enactment (high arousal) with the energy and speed of the robot’s actions must consider issues of unwanted noise (from the robot’s gears) and balance (it may fall over if it reacts too dramatically). The latter also affects its use of space. As robots move closer together, to e.g. convey emotional closeness, their gestures must become more subtle, otherwise they accidentally strike one another in the execution of a sweeping motion. While the interpretation and performance frameworks might look different than they are had a different platform been chosen, these modular and extensible frameworks can grow to accommodate other choices in the future.

Because the Nao platform has been used in a variety of related research (e.g., [125], [294], [28], [4], [11], and [13]), this speaks well to the reproducability of the current approach. Despite its limitations,
<table>
<thead>
<tr>
<th>Mode</th>
<th>Setup</th>
<th>Video Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>Therapist Storyteller</td>
<td>![Icon]</td>
<td>bit.ly/2ouZbY</td>
</tr>
<tr>
<td>Double Act</td>
<td>![Icon]</td>
<td>bit.ly/2SNeeHQ</td>
</tr>
<tr>
<td>Story Enactment</td>
<td>![Icon]</td>
<td>bit.ly/2Ud0GYx</td>
</tr>
<tr>
<td>Interactive Storytelling</td>
<td>![Icon]</td>
<td>bit.ly/3q7UhKl</td>
</tr>
</tbody>
</table>

Table 12: Storytelling modes with their respective properties and links to recordings. The ![Icon]-icon stands for a Nao robot as an embodied actor. The ![Icon]-icon indicates that an Amazon Alexa is part of the performance. If the mode is interactive, the ![Icon]-icon signals that an audience is participating in this mode.

The Nao suits the needs of an embodied story performer, not least because it has 25 degrees of freedom and the ability to move its limbs independently of each other (see Fig. 9). The robot’s fixed facial expression is clearly a limitation, one that prevents it from conveying emotion with facial cues, yet this also helps it to avoid unwanted bias in the user studies. It also means that the robot’s manual gestures will not occlude any facial expressions at key points. Other limitations can be addressed in a satisfactory fashion. For instance, the Nao cannot turn on the spot, but turning can be implemented as a composition of spatial and rotational movements (see Fig. 20). So, although the movements of the performance framework are shaped in large part by the abilities of the robots, they are not wholly determined by their limitations.

4.8 Demonstration

To conclude this Chapter on the Practical Implementation, an overview of all four modes is provided and links to videos for each mode are provided. These videos either show descriptive videos of the mode or recordings of robotic performances of the mode. The overview and listings are presented in Table 12.

The Practical Implementation comprises four different configurations of the Scéalability storytelling system. Each condition addresses a different research question by contrasting different layers of performance against each other. The Therapist Storyteller is able to provide a basic implementation of an interactive storyteller. The exact response to RQ1 and RQ2 that this mode can provide will be discussed in Chapter 6. The same holds for the Interactive Storytelling mode, which includes the audience’s non-verbal behaviour and emotion during
a storytelling performance. The *Double Act* and the *Story Enactment* modes are both ideal implementations to empirically test the foundational hypothesis on spatial schema and embodiment. These tests sit at the core of the empirical evaluation in the next Chapter.
EMPIRICAL EVALUATION

Analysing the effectiveness of embodied gestures and locomotion in storytelling performances requires that the audience provides feedback on their perceptions. Moreover, a meaningful assessment of the layers of performance requires contrasting conditions and rigorous statistical methods. The evaluation is divided into two parts:

- **Experiments: Embodiment** investigates the contribution of an embodied actor in a storytelling performance. The *Double Act* will be presented to multiple cohorts of test subjects under different conditions. Those conditions contrast different degrees of embodiment. This empirical evaluation addresses RQ2 (What can embodiment add to a performance?). The relevant publication is [1]. The design and execution of this experiment, together with its evaluation, was done as a shared contribution together with Thomas Mildner [279].

- **Experiments: Space & Gesture** investigates how gestures affect the appreciation of a storytelling performance, when they are meaningfully connected with the actions of the plot. Specifically, different conditions of the *Story Enactment* mode are assessed and evaluated by participants. This empirical evaluation addresses RQ3 and RQ4. The relevant publications are [2, 258].

The empirical evaluation is split into two studies because each study focuses on a different hypothesis and uses different modes. However, the structure for both is the same: the hypothesis is introduced, followed by an explanation of the methodology, analysis and results. All experiments including human participants have received an official exemption from ethics approval by UCD’s ethics committee as explained in Appendix A.

5.1 EXPERIMENTS: EMBODIMENT

5.1.1 Hypothesis

RQ2 (What can embodiment add to a performance?) is addressed through a set of specific hypotheses, which are tested using the *Double Act* mode of Scéalability. Since the *Double Act* allows us to present one or more embodied (Nao) or non-embodied (Alexa) actors in a performance, we can pose more fine-grained hypotheses regarding this RQ. First, we assume that pairing more devices will enhance the storytelling experience (H1), because each device can play to its strengths
individually. As argued in Sec. 4.4, Alexa can provide narrative skills, whereas Nao provides enactment through embodiment. Second, this not only permits us to check whether an additional agent is beneficial to the storytelling, but also whether the contrast between embodied and non-embodied agents enhances the performance (H2). Third, the *Double Act* introduces a humorous component to the storytelling performance and, with respect to embodiment/non-embodiment, this allows us to investigate if different devices contribute differently to the perceived humour of a performance (H3). Fourth, we can broadly hypothesise that different devices contribute differently to the performance (H4). The resulting hypotheses are as follows:

**H1**: A pairing of two or more devices enhances the user’s experience of a storytelling performance.

**H2**: A contrast of embodied and non-embodied agents enhances a performance.

**H3**: Different devices contribute differently to the perceived humour of a performance.

**H4**: Different devices are appreciated differently in a performance.

### 5.1.2 Experimental Setup

A storytelling performance in the *Double Act* is typically performed by the two devices Nao and Alexa. Scéalability allows us to alternate the performance to either be executed as a single device (Alexa or Nao) working alone (*Solo-Act*), or as a *Partner-Act* of two or more interacting devices. Our focus is on the theatre of devices-as-agents. We can vary the number of narrative, enactment and commentary agents to explore emergent phenomena, such as the humour that often emerges from the interaction of devices with their own characteristics. Each of these acts has two or more configurations to be evaluated here. A Partner-Act can be realized as an Alexa with a Nao, or as two Alexa devices. The Solo-Act can be realized as an Alexa or a Nao working alone. These conditions allow us to explore the contributions of each device in a storytelling performance.

#### 5.1.2.1 The Partner-Act

In this condition, a pairing of two agents performs a story that has been generated using *Scéalextric*. Narrative agents advance the plot, while commentary agents react to plot developments. Alexa works mainly as a narrative agent, while Nao’s commentaries serve five broad purposes: it may express emotions raised by a story; it may react to structural qualities in the plot (e.g. if it seems too linear, or too erratic); it may pantomime the current event; it may use gestures
5.1 EXPERIMENTS: EMBODIMENT

<table>
<thead>
<tr>
<th>Condition</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Devices</td>
<td>Alexa</td>
<td>Nao</td>
<td>Alexa + Nao</td>
<td>Alexa + Alexa</td>
</tr>
<tr>
<td>Configuration</td>
<td>Solo-Act</td>
<td>Solo-Act</td>
<td>Partner-Act</td>
<td>Partner-Act</td>
</tr>
<tr>
<td>Properties</td>
<td>Non-Embodied</td>
<td>Embodied</td>
<td>Non-Embodied + Embodied</td>
<td>Non-Embodied</td>
</tr>
</tbody>
</table>

Table 13: A total of four studies tests the different story-telling aspects of *Scéalability*. These studies are concerned with the perception of multi-agent story-telling in different configurations.

to reflect the relationship between characters; or it may use voice and gesture to comment on the entertainment value of the narrative so far. These complementary functions can give Alexa and Nao the appearance of conflicting personalities, and the pairing has a comedic dimension that neither has on its own.

5.1.2.2 The Solo-Act

In this condition, a single agent on a single device presents the narrative, which is generated as per the Partner-Act condition. This condition represents the most typical consumer interaction with a smart device. It is considered here for the contrast it offers to the double act on two fronts: an Alexa working alone is a voice without a body, while a Nao acting alone emphasizes physical embodiment. When Alexa works alone, the story is read aloud in the manner of an audio book; when Nao works alone, it speaks and gestures as the plot dictates.

5.1.3 Methodology

In order to explore the contribution of communication and embodiment in cross-device or cross-agent interactions, four contrasting but comparable conditions (see Table 13) are investigated.

To remove the choice of story as a variable, a single computer-generated story is used in each of the four conditions. We pick a story at random, yet want to choose one that affords each device as many opportunities as possible to showcase its abilities. We first build a simulator of *Scéalextric* stories, and chose the one whose simulation most utilizes the devices’ abilities. So, regardless of how many Alexa or Nao devices are used, the following *Scéalextric* story is used as a test basis:

**What if Edna Krabappel fell in love with Homer Simpson? So at first, Edna fell deeper in love with Homer than she ever had with Matt Groening. But Homer was bored almost to death by**

---

1 A sample video of the Partner-Act can be viewed at: [https://bit.ly/2kSp1jN](https://bit.ly/2kSp1jN).
Edna. Yet Edna charmed Homer with a sly smile. So Homer started an illicit affair with Marge Simpson. Well, Edna viciously sliced Homer with her sharp-edged ruler. So Homer reported Edna to the police. Well, Edna begged for Homer’s forgiveness. So Homer lowered the boom on Edna. Then Marge Simpson went down on bended knee and proposed to Homer. So Homer introduced Marge Simpson to his social circle. But Marge Simpson invited Barney Gumble for a romantic meal. Yet Barney Gumble turned a cold eye to Marge Simpson’s entreaties. But Marge Simpson harassed Barney Gumble continuously. So Barney Gumble passed information to Homer. “We’re no longer an item” said Homer to Marge Simpson. “Forgive Me” begged Homer of Edna. Then Edna forgave all of Homer’s transgressions. After this, Edna went down on bended knee and proposed to Homer. Then Homer forged a bond with Edna. So in the end Edna married Homer in Springfield they honeymooned in Springfield. Thereafter Edna and Homer were utterly inseparable; wherever Homer went Edna was sure to follow. The End

Experiments based on this common narrative should allow us to consider, in the context of the Experiments: Embodiment, the hypotheses H1-H4. We set out to test these hypotheses by presenting all four conditions (Tab. 13) to a class of university students. The four conditions are set up as follows:

1. **Nao - Solo-Act**: The robot alone presents the story, using voice and gesture.

2. **Alexa - Solo-Act**: The Alexa device alone presents the story, using spoken narration.

3. **Nao & Alexa - Partner-Act**: The Alexa device uses speech to narrate the story, and the Nao reacts accordingly with speech and gestures of its own.

4. **Alexa & Alexa - Partner-Act**: One Alexa leads the narration and another responds accordingly.

The following transcript shows how the Partner-Act interprets part of the shared story:

— Nao opens both arms towards the audience and closes them in front of its body.
Alexa: “What if Edna Krabappel fell in love with Homer Simpson?”
Nao: “Ok.”
— Nao looks down at Alexa.
Alexa: “So at first, Edna fell deeper in love with Homer than she ever had with Matt Groening.”
Nao: “But then what?”
— Nao looks down at Alexa.

**Alexa:** “But Homer was bored almost to death by Edna.”
— Nao slumps forward with boredom.

**Nao:** “That is interesting.”

**Alexa:** “Yet Edna charmed Homer with a sly smile.”
— Nao raises its right hand, looks at it and opens and closes it.

**Nao:** “Wow, I did not expect that!”

— **ALEXA introduces a new character by using knowledge of its unique weapon**

**Alexa:** “Get ready to be stung with an angry beehive.”
— Nao retrieves its name from the blackboard

**Nao:** “Marge Simpson! Oh yes, right.”

... 

The cohort of Computer Science students who volunteered for the experiment were divided into four groups of 13, 14, 20, and 14 participants respectively, to evaluate the four conditions above in a *between-subjects* fashion. Each group was shown just one condition, and each participant filled out a questionnaire about just one performance (see Appendix B). Participants were asked to rate the quality of this performance, with a focus on entertainment value and humorousness. The questionnaire comprised statements to be rated on a Likert-Scale, such as “I liked the story”, “The performance as a whole was humorous” and “The performance as a whole was entertaining.” A forced choice scale with six points [295] was used to preclude neutral responses. Importantly, we distinguish between the story itself and the performance of the story. The former is the textual rendering of the narrative, which is shared across all four conditions, whereas the latter is the real-time performance of the story by one or more devices. We discarded 12 answer sets that were incomplete or invalid.

### 5.1.4 Analysis

To address the first hypothesis (H1), we compared the aggregate ratings across all four conditions for the items “I like the story/performance”, “The story/performance as a whole was humorous” and “The story/performance as a whole was entertaining” (The full questionnaire can be found in Appendix B). The Kruskal-Wallis H-test for independent samples produces a p-value = 0.00094 (test statistics = 10.945). There is thus a significant difference in how subjects appreciate the underlying story and its actual performance. Comparing measures mean $\mu_{\text{story}} = 4.1170$ and standard deviation $\sigma_{\text{story}} = 1.1334$ for the likeability ratings of the story with $\mu_{\text{perf}} = 4.5029$ and $\sigma_{\text{perf}} = 1.1717$ for the likeability ratings of the performance, subjects appear to prefer the performance over the story. See Fig. 37 for a visual comparison of these values.
5.1 experiments: embodiment

Figure 37: Differences in likeability ratings for “Story” and “Performance” in the four conditions of the first experiment. Ratings for appreciation of the story are shown in blue, and for the performance in red.

To test whether either of the devices contributes more to audience perceptions of humour or entertainment (H3 & H4), we compared the ratings for the two items “The robot added a humorous dimension to the performance” and “The speaking device added a humorous dimension to the performance” for the Partner-Act condition of Nao and Alexa. The Wilcoxon signed-rank test shows a significant difference (p-value = 0.002098, test statistics = 16.0), suggesting the robot is seen as the more humorous device. Moreover, if we compare ratings for “The robot greatly contributed to the performance” and “The speaking device greatly contributed to the performance”, the robot is shown to make the greater contribution overall (p-value = 0.018996, test statistics = 19.5). These findings are visualized in Fig. 38.

5.1.5 Results

This study shows the value of physical presence and embodiment when a humorous outcome is desired. Not only do audiences perceive the physical performance of a story as more entertaining than the story itself, it seems that the more embodied agent contributes most to this perception. The evaluation of the Partner-Act has shown that cross-device interaction can yield promising results with respect to humorous storytelling.

While H1 cannot be reliably answered, a robot actor adds significantly to the perceived humour of performances, thus providing a
5.2 experiments: space & gesture

5.2.1 Hypothesis

This empirical study utilizes the Story Enactment mode of Scéalability to create a network of artificial story-tellers, both embodied and non-embodied, so as to transform a purely textual narrative into a physical performance. By exposing the inner structure of a narrative at every level, it also provides explicit symbolic hooks on which interacting robots can hang their embodied performance. Hence, this study focuses on specific aspects of embodiment, namely gesture and space. In line with the previous Chapters, this study will provide evidence for the shared image schematic underpinnings of spatial movement and gestures.

The schemas that underpin gestures [16, 80] do what conceptual metaphors do for spoken language [8]: they generalize, connect, and
support a degree of intuitive metaphoricity. When spatial schemas – a specific kind of image schema – are used in an apt and timely fashion, they naturally enhance the actions of the story. The nature of spatially-situated storytelling raises research questions that we will explore in a series of empirical studies. For instance, in the previous section we have shown that the physical body of a robot contributes more to effective storytelling than a non-embodied voice-only agent. Robots add a degree of humour through the physicality of their presence [1], and story events are enhanced when their actions are spatially-coherent at the image-schematic level. Our studies show that spatial movements governed by a schematic logic enhance the audience’s appreciation of a narrative performance. Specifically, we focus here on one dimension of spatial movement involving two robot actors, to show how relative distance can convey changing interpersonal attitudes, and add as much to a performance as more showy, pantomimic gestures.

Spatial schemas are imagistic cognitive structures that apply in a great many contexts [8], and unlike pantomime, they generalize well across domains and cultures. To illustrate the difference, consider a story event in which character A proposes marriage to character B. Robot A might pantomime this action in an iconic fashion, by going down on one knee and extending its arms toward robot B. Although suggestive, this action is culturally determined and far from universal. In contrast, image-schematic reasoning offers a more subtle response: robot A should simply move closer to robot B. If B accepts the proposal, it should move closer to B; but if it declines, it should step back. We hypothesize that the latter strikes a chord with audiences even if it is less showy than the iconic alternative.

We hypothesize that robots which use space coherently, in line with our image-schematic intuitions, will be appreciated as better storytellers than those that are indifferent to these concerns. Although spatial movements are more subtle than showy, pantomimic gestures – for instance, a robot going down on one knee to propose to another, or flexing its biceps in a show of strength – we hypothesize that audiences will register them, perhaps unconsciously, and appreciate them in their ratings of a performance. We thus formulate the following additional hypotheses:

**H5:** All things being equal, audiences appreciate performances that make coherent, image-schematic uses of space more than those that do not.

**H6:** All things being equal, audiences appreciate the coherent use of space in a performance as much as the use of iconic/pantomimic gestures.
The experimental studies to follow include two pilot studies and one major study. The analysis and results of the findings will ground the subsequent discussion in the final Chapter of the thesis.

5.2.2 Pilot Studies

It is in the nature of Scéalextric stories to run long. Its plots often contain many twists and turns, and involve multiple characters: not just the pairing of A and B, which are embodied in the performance, but supporting characters such as A-friend, B-spouse and A-enemy-lawyer. The audience never sees the latter, but only hears of them in the narration, or in the dialogue of A and B. The dialogue module plays a central role in ensuring that the audience is not confused by the succession of characters, speech-acts and events. It does this in a number of ways (see Sec. 2.1.2.3). First, it inserts the name of the addressee into every speech-act that it generates. For instance, A may say to B, “Marry me, Catwoman,” to which B may reply, “I must decline your proposal, Joker.” In this way, the audience is constantly reminded of which robot is which. Another heuristic exploits the vocal range of the robot actors. When A and B are of opposite genders, the speech synthesizer is directed to reflect this. Thus, if robot A plays a male character, it will use a male voice; and if B plays a female character, it will use a female voice.

Even with these measures, the stories are still too long to reliably test for the coherence of gestures and spatial movement, especially in a crowd-sourcing evaluation via AMT (Amazon Mechanical Turk). The framework for evaluation has thus evolved by necessity. This section sketches two pilot studies that explore potential issues, relevant questions, test materials and presentation formats.

5.2.3 Questionnaire Items

Ritschel et al. [112] have investigated two different modes of storytelling with a robot using the AttrakDiff questionnaire [22] as their experimental instrument. One particular construct in this questionnaire, called ATT (for rating overall attractiveness), is most relevant here. This construct comprises 7 test items, in which subjects answer questions of the form “The performance of the robots is ...” with a value ranging from 1 to 7 on the following AttrakDiff dimensions:

1. unpleasant $\leftrightarrow$ pleasant
2. ugly $\leftrightarrow$ attractive
3. disagreeable $\leftrightarrow$ likeable
4. rejecting $\leftrightarrow$ inviting
5. bad $\leftrightarrow$ good
6. repelling $\leftrightarrow$ appealing
7. discouraging $\leftrightarrow$ motivating

The pilot studies rather naively presented full stories and the entire AttrakDiff questionnaire to AMT judges. Each was presented as a video recording of an embodied performance that is over 3 minutes long. The videos for each can be viewed for one robot at youtu.be/0ZQewRzo7gU and for two robots at youtu.be/d4xNMFXTZLU.

5.2.4 Preliminary Findings

Each pilot study was conducted on Amazon Mechanical Turk, with 100 judges for the first and 200 for the second, yielding the following preliminary findings.

Pilot I: One Robot In this repeated measures design, 50 subjects were presented with a robot performance that included a single robot either doing subtle, schematic movements or showy, pantomimic gestures. The second group of 50 subjects were presented with a one-robot performance that featured these gestures in either coherent or incoherent fashion (i.e. the movement is informed by the accompanying narration or not). All 100 judges were asked to rate the full AttrakDiff questionnaire\(^2\). With 7 items per construct, 28 ratings have been acquired by each judge. With the AMT platform, we automatically rejected any judge that failed to answer a few test questions, hence all 100 responses could be evaluated. Each subject was paid $0.40 USD for their judgements.

When judges rated videos of Coherent Spatial Movement and of Coherent Gesture\(^3\) usage in a one-robot performance, the ATT construct showed a significant difference ($p = 0.00055$) with means and standard deviations $\mu_{\text{Spat}} = 4.129$, $\sigma_{\text{Spat}} = 1.596$ and $\mu_{\text{Gest}} = 3.606$, $\sigma_{\text{Gest}} = 1.807$. For Incoherent versus Coherent Gestures, the one-way ANOVA yielded a significant difference ($p = 4.228 \times 10^{-15}$) for the AttrakDiff construct, with means and standard deviations $\mu_{\text{IncohGest}} = 3.984$, $\sigma_{\text{IncohGest}} = 1.425$ and $\mu_{\text{CohGest}} = 3.566$, $\sigma_{\text{CohGest}} = 1.300$. A comparison of three AttrakDiff constructs in both conditions is depicted in Fig. 39.

Coherent spatial movements thus seem to be appreciated more than coherent gesture usage in the embodied performances. Oddly, however, coherent gesture usage also seems to rate lower than incoherent gesture usage, and this casts a shadow over our ability to meaningfully compare gesture and space.

\(^2\) The questionnaire as presented in Appendix B.2 and contains the four constructs PQ (pragmatic quality), HQS (hedonic quality with respect to self improvement), HQI (hedonic quality with respect to user’s identity) and ATT (overall attractiveness)[22].

\(^3\) Labels: “Gest” for gesture, “Spat” for spatial, “Coh” for coherent and “Incoh” for incoherent.
Pilot II: Two Robots This second pilot was a between-subjects design (i.e. each subject only saw and rated one condition) with four conditions and 50 subjects per condition. Again, subjects were paid $0.40 USD and we excluded judges that failed test questions, which totalled in 200 valid responses. In the second pilot, judges rated videos with two robots either using spatial movement or gestures in coherent or incoherent fashion. Informed by the results of the first pilot, we drop the first construct (PQ) of the AttrakDiff questionnaire and kept the remaining 3 constructs (hedonic quality) with 7 questions each.

Using videos with two robots, we conducted 2-way ANOVAs for Spatial versus Gesture and for Coherent versus Incoherent, which yielded significant differences on the AttrakDiff scale for the accumulated constructs. A post-hoc t-test shows a significant difference between Spatial Movement and Gesture ($p = 0.002$, Cohen’s D = 0.094) with means and standard deviations $\mu_{\text{Gest}} = 4.591$, $\sigma_{\text{Gest}} = 1.636$ and $\mu_{\text{Spat}} = 4.430$, $\sigma_{\text{Spat}} = 1.785$. There is also a significant difference ($p = 0.002$) in favor of coherent action (Cohen’s D = 0.094). Coherent spatial movement is significantly more appreciated on the AttrakDiff scale than incoherent Movement (Cohen’s D = 0.272). A comparison of three AttrakDiff constructs in both conditions is depicted in Fig. 39.

Overall, the pilots show that audiences have a preference for coherent over incoherent spatial movement. This is the simplest dimension to appreciate, and is grounded in image-schematic intuitions in ways that pantomimic gestures are not. It appears that the length of the videos (3 minutes) worked in favour of spatial coherence, since space – when used coherently – serves to summarize the cumulative state of the story so far. Conversely, the length of the videos seems to have worked against the appreciation of coherent gesture usage. In the main study, shorter videos with more compressed effects will be used, to reflect these preliminary findings. Moreover, the ATT dimension of the AttrakDiff questionnaire seems to be the most important dimension, hence only the 7 items of the ATT construct are kept. To this core of 7 items, the main experiment adds 7 more that specifically relate to the embodied performance of narratives by robots:

8. The robots appear human-like
9. The robots show their intentions
10. I could act like one of the robots
11. The robot mirrored how I would react
12. I sided with one of the robots
13. I am curious as to how the story continues
14. The robots’ movements are appropriate to events in the story
Figure 39: Three AttrakDiff\cite{22} constructs have been rated in two pilot studies. The first pilot presented a one-robot performance with spatial movements and gestures. The second pilot presented two robots with spatial movements and gestures. All 21 questionnaire items can be found in the Appendix: HQS & HQI in App. B.2 and ATT in App. B.3.
The revised questionnaire used for the main study is listed in Appendix B.3.

5.2.5 Revised Presentation Format

Crowdsourcing is typically used for simple annotation tasks or for straightforward evaluation. It is an efficient means of eliciting large quantities of annotated data that are comparable in quality to university participant samples [296]. However, with respect to average response time and the quality of elicited judgments, the pilot studies indicate that a 3 minute video is too long to hold the attention of participants. The video of each performance is thus edited down to just 1 minute, to focus on the most representative sequences for test purposes.

Each video now illustrates just two story actions from the larger narrative. As shown in Fig. 40, each action is introduced with a text panel that summarizes the story so far, giving participants a context for what is to follow. This text is also vocalized using speech synthesis, as though delivered by the narrator. The video begins with an introduction from the robot actors, who announce, with a bow, the characters they will play in the subsequent clips.

The first panel in Fig. 40 shows this context: “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...”. What follows is a story event in which robot A, playing Hillary, proposes to robot B, playing Donald. The action then jumps forward in the narrative, as the next panel explains: “But Donald felt a deep love for Melania. So Donald turned a cold eye to Hillary’s entreaties. Well, Hillary took Donald hostage ...”. The last clip then shows A and B enact Hillary’s abduction of Donald. The video concludes with a final text panel that reveals the ending of the story.

5.2.6 Main Study

Each of Scéalability’s modules can be used in a coherent or incoherent fashion. In coherent mode, a module generates communicative acts that are suited to the underlying story action. But these acts can also be executed incoherently, insofar as they ignore the underlying event, or deliberately pursue the wrong course of action. More specifically, the gesture model can choose its gestures randomly, from its full repertoire of articulations, while the spatial model can do the opposite of what an action calls for. It can move forward when it should step back, or vice versa. If gestures and relative motion actually enhance a performance, we can expect audiences to prefer the coherent mode over the incoherent alternative.
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...

But Donald felt a deep love for Melania. So Donald turned a cold eye to Hillary’s entreaties. Well, Hillary took Donald hostage ...
### Conditions

<table>
<thead>
<tr>
<th></th>
<th>Spatial Movement</th>
<th>Gestures</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Coherent</strong></td>
<td>SM - Coherent</td>
<td>GM - Coherent</td>
</tr>
<tr>
<td><strong>Incoherent</strong></td>
<td>SM - Incoherent</td>
<td>GM - Incoherent</td>
</tr>
</tbody>
</table>

### Example

<table>
<thead>
<tr>
<th></th>
<th>A likes B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coherent</td>
<td>A moves closer to B</td>
</tr>
<tr>
<td>Incoherent</td>
<td>A backs away from B</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A hates B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coherent</td>
<td>A backs away from B</td>
</tr>
<tr>
<td>Incoherent</td>
<td>A moves closer to B</td>
</tr>
</tbody>
</table>

Table 14: Above: four experimental conditions for evaluating the contribution of Scéalability’s SM and GM models to a story performance. Below: two examples, “A likes B” and “A hates B”.

The main study assesses embodied performances with two robot actors and a single Alexa narrator. Each performance employs the dialogue model to augment the narrative with speech acts, but the gesture and space modules are not used together. Rather, for comparison purposes, we evaluate performances that make coherent/incoherent use of gesture separately from those that make coherent/incoherent use of space. The four test conditions are summarized in Table 14.

### 5.2.7 Methods

The main study was again conducted via Amazon Mechanical Turk. However, any participant who previously participated in one of the pilots is now excluded, and access is limited to raters at the level of Master Worker only. This is a registered participant that has been granted a Master qualification by the AMT platform. These participants are said to demonstrate consistently high quality work for a wide range of tasks. Each rater was paid $0.40 per trial. We assigned the task to 40 participants in each of the four trials/conditions (N = 160).

**Instructions:** Each participant is asked to watch a video and answer a questionnaire about the experience. Each video is presented in the format described in Section 5.2.5 and depicted in Figure 40. When the video ends, a new page brings the participant to the questionnaire. In addition to the 14 questions described in Section 5.2.3 (see Appendix B.3), another 3 items are added to test workers on their engagement with the task. One such meta-question asks “How many robots are on screen in the videos?” Any rater who incorrectly answers any of these
5.2 Experiments: Space & Gesture

engagement tests is automatically excluded. The order of the 14 + 3 test questions is randomized for each rater.

Conditions: The four trial groups are presented with the same Scéalextric story. The script of this excerpt is presented in Section 4.5 and a storyboard for this one-minute story is shown in Fig. 40. The four conditions differ only in how the robots embody the same story events on screen4. The four conditions are: Coherent Spatial Movement, Incoherent Spatial Movement, Coherent Gesture and Incoherent Gesture (see Table 14). To address the central research questions, we investigate the effect of mode (coherent versus incoherent) and embodiment type (spatial movement versus gesture) using a two-way ANOVA. If significant, we apply a two-sided t-test to identify the effect sizes for each factor independently (mode and embodiment type). We also apply a Bonferroni correction for the 2-factor analysis.

5.2.8 Analysis

The data was collected over the course of a week. After filtering invalid responses to the 3 tests of engagement, data was collected from 4 independent groups comprising 32 (Coherent Spatial Movement), 29 (Incoherent Spatial Movement), 29 (Coherent Gesture) and 28 (Incoherent Gesture) participants, for a combined total of 118 valid responses. An overview of the average rating for each item in all four conditions is shown in Fig. 41, while the accumulated results for each condition are shown in Fig. 42. A two-way ANOVA shows a significant p-value for the factor of coherence (mean squares = 48.138, F-values = 16.147 and p-value = 0.000061).

The post-hoc t-test shows significant differences between coherent and incoherent conditions with Cohen’s D = 0.197, which indicates a small to medium effect size in favor of the coherent conditions. The mean rating across coherent conditions is $\mu_{\text{Coherent}} = 3.820$, and across incoherent conditions it is $\mu_{\text{Incoherent}} = 3.480$. There is no significant difference between the Spatial and Gesture conditions.

---

4 To see the video materials for each condition, please visit tiny.cc/s21fiz
5.2 Experiments: Space & Gesture

Figure 42: Accumulated mean ratings for each item in all four test conditions. Note, y-axis displays cropped values of the actual 1-7 rating scale. The whiskers indicate the standard error of the mean.

<table>
<thead>
<tr>
<th>Condition L</th>
<th>Condition R</th>
<th>Mean L</th>
<th>Std L</th>
<th>Mean R</th>
<th>Std R</th>
<th>p-value (uncorr)</th>
<th>p-value (Bonf. corr)</th>
<th>Cohen’s D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial Overall</td>
<td>Gesture Overall</td>
<td>3.600</td>
<td>1.243</td>
<td>3.716</td>
<td>1.725</td>
<td>&lt;0.001</td>
<td>0.174</td>
<td>0.069</td>
</tr>
<tr>
<td>Coherence Overall</td>
<td>Incoherence Overall</td>
<td>3.820</td>
<td>3.394</td>
<td>3.486</td>
<td>1.794</td>
<td>6.4×10⁻⁴</td>
<td>0.197</td>
<td>0.115</td>
</tr>
<tr>
<td>Coherent Spatial</td>
<td>Incoherent Spatial</td>
<td>3.728</td>
<td>1.794</td>
<td>4.458</td>
<td>3.457</td>
<td>0.023</td>
<td>0.047</td>
<td>0.155</td>
</tr>
<tr>
<td>Coherent Gesture</td>
<td>Incoherent Gesture</td>
<td>3.724</td>
<td>1.565</td>
<td>3.303</td>
<td>1.734</td>
<td>0.003</td>
<td>0.003</td>
<td>0.244</td>
</tr>
</tbody>
</table>

Table 15: Overview of the t-test statistics for all four conditions. The four factors concern spatial movement (coherent vs. incoherent), gesture (coherent vs. incoherent), embodiment type (space vs. gesture), and coherence (coherent vs. incoherent). Cohen’s D is the effect-size.

(p-value = 0.174). The mean rating across Spatial movement conditions is $\mu_{\text{Spatial}} = 3.600$, and across the Gesture conditions it is $\mu_{\text{Gesture}} = 3.716$. A significant difference is found between coherent and incoherent uses of Spatial movement (Bonferroni corrected p-value = 0.047) and between the coherent and incoherent uses of Gesture (Bonferroni corrected p-value = 0.001). Each shows a positive Cohen’s D effect-size in favor of coherent conditions. See Table 15 for an overview. Additionally, we can observe the differences on the level of the individual items for Coherent vs. Incoherent conditions in Fig. 43 and for Spatial vs. Gesture conditions in Fig. 44.

5.2.9 Results

The results offer solid evidence in support of hypotheses 5 & 6:

- H5: Do audiences appreciate performances that make coherent, image-schematic uses of space more than those that do not?

The empirical findings suggest that audiences do appreciate the use of spatial movement in an embodied performance, and prefer coherent over incoherent uses of space (whether metaphorical or literal) to convey the impact of story events. Audiences also appreciate the use of gesture in a performance, and once again appear to appreciate action-appropriate (coherent) gestures over action-insensitive (incoherent) gestures.
5.2 Experiments: Space & Gesture

Figure 43: Average ratings of all questionnaire items (Sec. 5.2.3) for Incoherent (cyan) and Coherent (purple) conditions.

Figure 44: Average ratings of all questionnaire items (Sec. 5.2.3) for Spatial (blue) and Gesture (green) conditions.
• **H6:** Do audiences appreciate the coherent use of space in a performance as much as the use of iconic/pantomimic gestures?

No significant difference can be found between the audience’s appreciation of gestures and of spatial movement during an embodied performance. Since audiences do appreciate the coherent use of both of these embodiment types, and appear to view them as equally beneficial to the narrative. This is a useful result for a variety of reasons. First, spatial movement is less dramatic and showy than many pantomimic gestures. Second, it is easier to implement in a general fashion. We have shown that just two kinds of motion – forward (reduce distance) and backward (increase distance) – can contribute as much as a large repertoire of complex gestural scripts.

5.2.10 **Complementary Study**

Lastly, we argue that the combination of coherent iconic/pantomimic gestures and coherent use of space presents the most appreciated version of the performance. Our final, complementary study provides evidence that this is in fact the case. Consequently, these results address **RQ4:** How do these layers interact with each other and the audience? We introduce this last empirical study with the following hypothesis:

**H7:** All things being equal, audiences appreciate a combination of coherent gestures and coherent use of space more than coherent gestures alone or coherent spatial movement alone.

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.

5.2.10.1 **Methods**

This study evaluates three performance modes:

1. **Pantomimic Gesture:** the tale is performed with narration, dialogue and gesture, but no schematic spatial movements back and forth.

2. **Spatial Movement:** the tale is performed with narration, dialogue and schematic spatial movements back and forth, but no iconic 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 back and forth as the plot demands, and they also make pantomimic gestures for each story verb in the plot.

The one-minute video\(^5\) of the *Combined Movement* condition is presented to 40 raters on the AMT crowd-sourcing platform. These results will be compared to the two conditions of the previous experiment (*Spatial Movement* and *Pantomimic Gesture*). Therefore, each rater is shown just one of the three performances and then asked to evaluate it using the \(14 + 3\) item questionnaire (App. B.3). In return, each AMT rater is paid 0.40\$ per questionnaire.

### 5.2.10.2 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 \(\mu_{\text{Spatial}} = 3.728\) and standard deviation \(\sigma_{\text{Spatial}} = 1.792\) for *Spatial Movement* and a mean value \(\mu_{\text{Combined}} = 4.131\) and standard deviation \(\sigma_{\text{Combined}} = 1.762\) 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 the analysis can be found in Table 16. Statistical tests have been conducted on the accumulated test construct (of all \(14\) items) and the results are visualized in Fig. 45. The whiskers indicate the standard error of the mean (\(\frac{\sigma}{\sqrt{N}}\)).

### 5.2.10.3 Results

Our findings suggest that a mix of embodiment strategies – what has been 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

---

\(^5\) See all of the recordings here: [https://tinyurl.com/wpes3jl](https://tinyurl.com/wpes3jl)
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 45 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 $\mu_{\text{Spatial}} = 3.728$, $\mu_{\text{Pantomime}} = 3.921$ and $\mu_{\text{Combined}} = 4.131$ show 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.

<table>
<thead>
<tr>
<th>Condition L</th>
<th>Condition R</th>
<th>L: Mean/Std</th>
<th>R: Mean/Std</th>
<th>p-value</th>
<th>Cohen D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Space</td>
<td>Gesture</td>
<td>3.728/1.792</td>
<td>3.921/1.691</td>
<td>0.316*</td>
<td>-0.111</td>
</tr>
<tr>
<td>Space</td>
<td>Combined</td>
<td>3.728/1.792</td>
<td>4.131/1.762</td>
<td>0.002*</td>
<td>-0.227</td>
</tr>
<tr>
<td>Gesture</td>
<td>Combined</td>
<td>3.921/1.691</td>
<td>4.131/1.762</td>
<td>0.206*</td>
<td>-0.121</td>
</tr>
</tbody>
</table>

Table 16: Post-hoc test for all three conditions. L and R denote conditions named in first (L) and second (R) column. *Bonferroni corrected p-value.
Chapter 5 has presented two large studies including the empirical evaluation of seven hypothesis through multiple experiments and different embodied performances in contrasting conditions. The first experiments focused on embodiment. They primarily addressed RQ2 and used the Double Act mode of Scéalability to test H1-H4. The second experiments focused on Space & Gesture. They primarily address RQ3 and RQ4 and used the Story Enactment mode of Scéalability to test H5-H7. The results of the studies can be summarized in the following statements about robotic storytelling performances:

- **Embodiment**
  - The embodied performance is more entertaining than the story itself.
  - The more embodied actor is perceived as more contributory.
  - An embodied actor can add significantly to the perceived humor of a performance.

- **Space & Gesture**
  - Audiences appreciate the coherent use of spatial movement and gestures.
  - Audiences appreciate the use of spatial movement as much as gestures.
  - Audiences appreciate the combination of spatial movement and gestures more than any single embodiment type.

The way in which these results respond to the research questions is the content of the next and last chapter, Chapter 6. This will relate the findings of this chapter to the theories presented in Chapter 1, then situate the claims of the work in the context of related works (Chapter 2) and finally conclude with a critical discussion of the work’s limitations and its future possibilities.
DISCUSSION AND CONCLUSION

6.1 DISCUSSION

The story of this thesis has been told over the past five chapters. With the Conceptual Framework as its theoretical contribution, the Practical Implementation as evidence for its feasibility, and the Empirical Evaluation as its supportive evidence, this last chapter serves as the epilogue to the story. It will revisit its most prominent characters, Nao & Alexa, and it will look again at the research questions which provided the story arc. The epilogue will also look at potential plot-holes (Limitations) and how the story of this work might be continued (Future Work). This final chapter starts by revisiting the main acts of the story so as to critically discuss their individual contributions.

6.1.1 Connecting Start and End

The opening chapter of this thesis presents different layers of performances (Fig. 1) and examines the role of computational approaches in embodied storytelling. First, Section 1.1 connected ideas in computational story generation with cognitive foundations of human thought. This set the tone for the rest of the thesis. Working our way through the layers of performance, we started with semiotics and ended with gestures. Signs present meaning through their iconicity, whereas movement can alter and reframe this meaning. This process of construal relies on the same image schematic underpinnings, which are grounded in spatial reasoning and which can be utilized throughout the different layers of performance. We show how image schemas are especially useful for embodied performances that rely on spatial movements and gestures. Chapters 4 and 5 are reflective of this argumentation. Working with Scéalxtric on a symbolic level of meaning, we presented multiple extensions to the baseline story generator via its integration into Scéalability, which additionally introduces dialogue, embodiment and gestures, which were built around the schematic patterns identified in language. Moreover, the empirical results presented novel evidence for audience appreciation of those movements. Apart from the empirical results, the thesis also makes a theoretical contribution, which will be discussed next.
6.1.2  A Conceptual Framework

The conceptual framework in Chapter 3 supports both fine-grained and gross-level insights into the unfolding storytelling performance. For instance, we have seen that aggregate assessments of valence – for a given role of a specific action at a particular point in the plot – allow for aggregate judgments about characters and their changing feelings. These gross judgments, which reveal positive or negative shifts in a character’s overall feelings, can suggest equally reductive actions for robot performers to execute on stage, such as moving closer to, or further away from. In this way, gross interpretations support powerful spatial metaphors that are equally summative and equally persistent. We have largely focused on the semantics of gross spatial actions, but the literature provides a formal basis for the more fine-grained forms of expression that can be pursued in future work, such as those from the domain of dance [266, 297].

But fine-grained insights are also supported by the framework, which is to say, insights based on movements along a single emotional dimension. Spatial movement back and forth, of the kind evaluated in Chapter 5, is reductive and general. But metaphors that allow a performer to construe one action as another in a given context, such as by construing an insult as an attack, or an act of praise as an act of worship, are more specific. They work at the conceptual level of plot action, and do more than suggest an embodied response. Rather, they increase the range of choices available to a performer because they operate at a deeper and more specific level of interpretation. As such, they provide the Conceptual Structures presented in Sec. 1.2.3.

We reach for a metaphor when we want to extend our expressive options, and so too can robot performers. But metaphor is just one choice that leads to others. Irony is another. A performer can, for example, choose to react ironically to a script directive. Suppose character A is expected to show fealty to character B, and the story so far firmly establishes this expectation in the minds of the audience (and in the view of the interpretation framework). Irony is always a matter of critiquing a failed expectation, by acting as though it has not failed while clearly showing that it has. Suppose now that the plot calls for A to rebel against, or stand up to, or to break with B. When the interpretation framework compares the emotions established by previous actions with those stirred by this new action, it recognizes a rift that should, if it is large enough, influence how the performers react. The robot portraying character A might thus act out an action more in line with the expected emotions, such as bowing down to B, while speaking the dialogue associated with the current action, such as “I’ve had enough of you!” . The bifurcation of irony, of expectation versus reality, easily maps onto the parallel modalities of speech and physical action, so that a performer can follow both branches at once.
We have examined irony in Sec. 3.2.4, showing that robotic performances of complex semiotic structures, such as stories, open many avenues for an interpretative performer, at both the conceptual and the expressive levels. These choices, which include construal mechanisms such as metaphor and irony and more besides, open more choices in turn, if a performer has the wit to perceive and exploit them. As such, it is fair to say that we have barely scratched the surface of what an interpretative approach to embodied performance can yet bring to domains such as storytelling. As we go deeper, we may need to use a richer model of the gestures and motions that realize the embodiment, such as by drawing on insights and representations from the world of dance, where more nuanced actions can be observed.

6.1.3 Implemented Performances

Full embodiment is more than a matter of mere physical presence, as discussed in Sec. 1.3.1. Even Alexa is minimally embodied via the Amazon Echo device, in a way that can be acknowledged by robot actors. In the one-robot-one-Alexa performances of Sec. 4.4, the Nao robot interacts with Alexa as a physical, if immovable object. It looks at Alexa, and addresses its remarks to her Echo instantiation.

But full embodiment for a stage performer requires animacy, and motion about the stage. The Nao robot provides this with its capacity for spatial locomotion and animated gesture. It is tempting to think of gestures as motions of the robot’s upper body, and of locomotion as motions of the robot’s lower limbs. But this would miss the point about the essential difference between both kinds of embodied movement. Most human gestures – and most of the robot’s – are culturally determined, and open to misconstrual. The most pantomimic are the embodied equivalent of clichés and idioms. In contrast, when motion on stage is performed in accordance with deeply-rooted image-schematic intuitions, a more universal meaning can be communicated.

In Section 5.2, we have contrasted spatial movement with pantomimic gesture, and evaluated an embodied system that makes use of both. We have shown that audiences appreciate the combination of coherent gestures and image-schematic spatial movements more than any single embodiment type on its own. These results provide the necessary depth to respond to the Research Questions (Sec. 1.4).
6.1.4 Addressing the Research Questions

6.1.4.1 RQ1: What layers of performance can be combined for an automated performance?

On a theoretical level we have shown how symbols, movement, agents, gestures and emotions can be integrated into one conceptual framework (Chapter 3). The literature review shows that most robotic performances combine gestures, movement and interaction on a superficial level. Conversely, our implementations combine the different layers (e.g. gesture, movement, agents, emotion) with a unifying scaffolding. We propose that the spatial schemas, which we have identified in each individual layer, can provide tangible connections.

To a large extent, Chapter 1 presents an answer to RQ1, especially with the layers of performance presented in Figure 1. In support, Chapter 4 presents several successfully integrated performances and the empirical evidence (Ch. 5) suggests that those performances can be appreciated by different pools of audiences. In particular, the analysis of H1 in Sec. 5.1.4 shows that observers appear to prefer the embodied performance over the textual story. We conclude with a list of the layers that have been successfully combined in this work:

- **Symbols**: The story generator Scéalextric provides the symbolic foundations on which the performances are built. Every action, character and beat of the generated stories has a symbolic representation that can be augmented with additional semantics. For example, “are_insulted_by” is one of over 800 representations in the Scéalextric database. It provides links to idiomatic renderings for the narrator (e.g. “Rick called Jerry a jerk”, see Sec. 2.1.2), renderings for the dialogue (e.g. “You are such a jerk!”, see Sec. 2.1.2.3), instructions for gestures (e.g. flipping the finger, see Sec. 4.2.1), instructions for spatial movement (e.g. moving away, see Sec. 4.2.2) and interpretive choice (e.g. is the insult perceived as an attack and should be displayed as such? See Sec. 3.2). The symbols at our disposal are transformed by means of the performance, i.e. the database entry “are_insulted_by” becomes an iconic gesture or a spoken insult. More importantly, the symbolic nature of the knowledge base provides access to the fabula and the discourse (Sec. 2.1). Access to this rich database motivated our choice of Scéalextric (Sec. 2.7) from among the available story generators, and it is through these symbolic hooks that we can investigate the schematic underpinnings of gestures and movements.

- **Embodiment**: With advances in robotics and availability of cheap computing power, the transition of performances from a simulated, virtual domain to a physical domain becomes more and
more relevant. We have reviewed the importance of embodiment in studies of performance [168, 171, 194, 266] (Sec. 2.2.1) and studies of creative systems [117, 118, 255, 256] (Sec. 2.6). In line with these studies, our own experimental results show that an embodied performance can be perceived as more entertaining than the story itself. Moreover, the embodied actor is perceived to contribute more than its non-embodied partner. Yet, simply introducing a body without movement into a performance does not automatically increase audience appreciation. Hence, we looked at the use of gestures and space:

- **Spatial Movement**: We have shown how the coherent use of spatial movements can be beneficial for the appreciation of a performance (Sec. 5.3).

- **Gestures**: We have shown how the coherent use of gestures can be beneficial for the appreciation of a performance (Sec. 5.3). Moreover, we have shown how the combination of coherent spatial movement and gestures is appreciated over any single embodiment type.

- **Words and Speech**: The stories in our performances are presented with narration and dialogue (Sec. 2.1.2.3). Even though we do not present an explicit empirical evaluation of this layer, we have outlined evidence as to how deeply the accompanying words relate to gestures [79, 199, 201] (Sec. 2.4) and the schematics of movement [8, 67, 241] (Sec. 2.5.3).

**Limitations** This answer to RQ1 focuses on the most important layers which were defined in Fig. 1. Arguably, an embodied performance can feature all sort of additional layers, e.g. stage props, soundtrack, stage design, etc. We have argued that these layers start at the most primitive units of meaning (signs) and, in a bottom-up fashion, we derived the spatial primitives (Sec. 2.5) that elevate meaning into more embodied forms (movement, gestures). This development stops at the level of interaction, which introduces an entirely new dimension to a performance. Contrasting every layer with their respective baselines produces the different conditions that were presented in our empirical evaluation (Chapter 5). Moreover, we argue that additional layers, such as stage props, soundtrack, stage design, etc., may be grounded in a similar iconicity that can be found in our gestures and may be redundant. For instance, in Sec. 1.3.2 we explained how a cowboy can be identified through posture and gestures as well as through a typical costume. Consequently, we acknowledge that our embodied performances could draw from further, additional layers, yet we argue that the chosen layers represent a foundational baseline for the study of embodiment in automated performances.
The role of verbal interaction is less relevant for the presented investigation of gestures and spatial movement, yet dialogue and conversations between robots and humans can contribute greatly to a performance. Even non-embodied, text-based agents can reliably simulate a personality such that a user perceives personality traits as intended [208].

6.1.4.2 RQ2: What can embodiment add to a performance?

First, as a philosophical perspective we introduced the notion of embodiment as a constituent of cognition in Sec. 2.5.4. Embodied cognition argues that cognitive phenomena, such as linguistic creativity, have to be observed not only by looking at the brain but also by looking at the body and the environment it is situated in. The literature establishes a strong connection between language and the body [66, 72, 200, 201]. We exploit this connection by drawing from spatial schemas, which can be identified in the text we want to perform on stage [218, 219]. Once the most useful schemas have been identified, embodiment adds them to the performance, grounding language and providing additional semantics.

Second, embodiment adds authenticity and relatability to the performance. Related works show that even simple shapes or moving objects elicit anthropomorphism and invite the audience to attribute emotions and intent [5, 6]. As outlined in Section 2.6.3, our systems present a form of organismoid embodiment [117]. This helps the audience to better identify with the storyteller and thus, the audience is better positioned to empathize with the story that the teller wants to convey. This argument is supported by our experimental results, which show that the embodied performance is generally more entertaining than the story itself. Specifically, our insights from the study of coherent movements is supportive here. The second set of items in Sec. 5.2 (Appendix B.3) include ratings regarding relatability, e.g. “I could act like one of the robots”, “The robots show their intentions” and “The robots appear human-like.” As such, this construct was rated higher for performances with coherent use of bodily movement (gestures and spatial movement), than those with incoherent usage. This shows that if embodiment is used in a way that it “appears human-like” or in a coherent context, it can improve the perceived appreciation of the performance. Or, as Guckelsberger et al. [117] wrote with reference to our systems: “physical embodiment can overcome symbolic representations and ground meaning in sensorimotor interaction” [117].

RQ2.1: What role do gestures play in performances?

The literature on gestures in performances with robots (Sec. 2.2.1) shows that they have often been scripted to suit the needs of one particular aspect of the performance, e.g. discourse markers [171, 172],
emotional expression [157, 158] or facilitating attention or participation [159, 160, 161]. The studies that focus on gestures in a storytelling performance [4, 13], as well as our studies, show that gestures have a positive effect on the appreciation of the performance, because they provide speech acts with additional contextual semantics [79]. As embodied signs, gestures introduce many properties into the performance, which we have outlined in Chapter 3. Iconic gestures provide an obvious meaning and their enactment carries a theatrical property (see Sec. 3.1.1.1). Deictic gestures are referential and as such they can draw the attention towards or away from someone or something. Metaphoric gestures provoke the audience’s abstract thinking or sublimely communicate additional meaning. Cohesives can support the performance if they are used recurrently in order to reinforce overarching aspects of the plot. Beats can be used in event-related gestures in order to emphasise or draw attention to a specific moment in the story.

We have drawn from all of these gestures to enrich our embodied performances (Chapter 4). The choice of gestures is directly informed by our conceptual framework. Moreover, our empirical results confirm these choices, because their coherent use during the evaluated performances is more appreciated than their incoherent use (Sec. 5.2.8).

In addition, we have outlined the connection between image schemas and gestures in Sec. 2.5.4 with reference to work by Cienki [92] and Mittelberg [16]. Practically, gestures can serve another role in performances, namely that they share the same schematic underpinning that can be found in language and thought. This is a role that we shall discuss further in our responses to RQ4.1.

RQ2.2: What role does the use of space play in performances?

The use of space in a performance brings a structural component. Through movement by actors on a stage, an audience receives a dynamic sense of relationships and emotional distance. A set of immovable actors is unable to influence the audience’s perception of space. This influence however, can be used to provide implicit information about how the story develops. For example, we have shown that the space between two actors can be schematized (Fig. 29) such that their distance can be codified within meaningful sub-spaces, i.e. close distance requires more subtle movements since the actors have entered each other’s personal space. This use of space is inherently connected to spatial schemas (Sec. 2.5), since it allows actors to exploit spatial schematics as a conceptual metaphor. Here, we have shown how the emotional closeness of the characters can be mirrored by the physical
closeness of the actors. Our experimental results confirm the effectiveness of such a coherent use of space (Sec. 5.2).

Apart from this structural component, spatial movements serve the aim of bringing liveliness and action onto the stage. When we strip away the complex humanoid form of our robots, the narration, the gestures and only leave moving shapes, we are left with what Heider and Simmel [6] called “apparent behaviour”. The intuition is that movement in itself can elicit an anthropomorphic attribution of intent and liveliness. Our conceptual framework (Chapter 3) categorizes these basic movements in space and provides an overview of its implications for different kinds of performances.

Summarizing the role of gestures and spatial movement, the former provide a variety of properties due to their broad semiotic nature (e.g. iconic gestures, metaphoric gestures, cohesive gestures, etc.). In contrast, the interpretation of spatial movements relies on primitive schemas which require much less context to appreciate. Which role provides the stronger impact on the performance? Our empirical results suggest that the combination of both, gestures and spatial movement, receives more appreciation than any single embodiment type. This might indicate that both roles have equal benefits for the performance.

Limitations  The roles of embodiment for performances are manifold. Moreover, there are many different types of performances and ways in which embodiment can play a role in them. In Section 3.1.3, we addressed the application of the performance framework in dance, storytelling, jokes and conversation. Nonetheless, neither all types of performances nor all types of gestures fall under the descriptions we have provided. For example, we have deliberately excluded recurrent gestures (see Sec. 2.4.1) despite their pragmatic function, because we wanted to exploit a consistent set of types to derive our conceptual framework. While some consensus on recurrent gestures exists [93, 299], we decided to choose those types which we could reliably identify in the related works (Sec. 2.4).

6.1.4.3  RQ3: What connects the layers of performance?

The foundations of our performances are symbolic and their enactment is embodied. As outlined from the first to the last chapter, we have used image schemas to connect symbols and embodiment. For each individual layer, we can identify its image schematic underpinning. Two main properties of image schemas make them suitable and stable connections throughout the different modalities.

First, we reviewed the position by Mandler [233] on image schemas in Sec. 2.5.2. Mandler argues that image schemas are derived from spatial primitives and that through schematic integrations, they can
be used as conceptual metaphors. Their inherently spatial nature thus makes them integral to many movements in a performance. At the same time, these spatial properties can be observed in symbols (e.g. the symbol \(\leftarrow\) \[241\]) and also in gestures \[16, 92\].

Second, image schemas provide combinatorial properties which can connect simple schemas (e.g. UP/DOWN, LEFT/RIGHT) for the description of complex processes (see Sec. 2.5.3). Work by Hedblom \[69\] demonstrates how the right formalism enables researchers to use image schemas for different kinds of tasks \[70, 239, 240\]. This makes image schemas useful for performances with more complex interactions. In our Interactive Storytelling mode, we have expanded the use of image schemas from NEAR/FAR (signalling the emotional distance between characters) to the use of UP/DOWN (using thumbs up and thumbs down) as an interactive tool with which audiences can influence the course of the story.

**RQ3.1: How do spatial schemas support performances?**

The empirical results of the studies presented in Sec. 5.2.6 show that the coherent use of spatial movements is appreciated over its incoherent use. In this study we have kept story, dialogue, gestures and narration the same, and only changed the way that the robots move towards and away from each other with respect to the mental closeness of the characters they portray. This suggests that the audience is meaningfully integrating the spatial movement in accordance with the other layers of performance (e.g. narration, dialogue and gesture).

**Limitations** The degree of support that spatial schemas provide is difficult to measure in absolute terms. Obviously, inverting the schemas and making the robots approach when they ought to retreat (and vice versa) does not have the same effect on the appreciation of the story as inverting the meaning of words and sentences. Therefore, it is non-trivial to assign an absolute measure to the effect of spatial schemas. Further studies with more fine-grained movements could be implemented to derive a more precise measure of their impact on the performance.

**RQ4: How do these layers interact with each other and the audience?**

This research question has been partially addressed in the previous discussion. The aspect that has been overlooked so far, is the one on the connection between a performance and the audience. The thesis has presented the Therapist Storyteller (Sec. 4.3) and the Interactive Storytelling (Sec. 4.6) as two interactive modes. In the former, the robot interacts with the audience through verbal means while, in the latter, the audience can interact with the robot through non-verbal means.
Human-Robot interaction is a broad domain that offers various strategies on how to successfully create an interaction between robot and audience. Nonetheless, we have discussed (in Sec. 2.3.3) how only little research has been conducted that utilises image schemas to motivate the interaction [13, 195, 218].

RQ4.1: How can image schemas help to connect performers with the audience?

As shown, all the layers of a performance are imbued with spatial schematics in more or less transparent ways. Since image schemas are derived from bodily interactions with the environment, they establish the ideal link for interaction between a human and a robot. One of the core traits of image schemas is their near-universal transparency. Consequently, our system’s schematic basis can exploit this universality to facilitate a more intuitive interaction. Moreover, moving from verbal to non-verbal communication removes language barriers and it connects the semantics of the performance more closely to bodily experiences.

The Interactive Storytelling (Sec. 4.6) mode provides an instantiation of non-verbal interaction. It connects the audience with the story through two modalities. First, the thumbs up and the thumbs down gestures tap directly into the spatial schematics of the UP and DOWN schemas to signal positive and negative responses. Here, we can observe the plus-minus parameter of image schemas (Sec. 2.5.1) at work. The (thumbs) UP is a positive response and the (thumbs) DOWN is a negative response that the user can provide to alter the course of the story. Hence, image schemas connect performers with the audience in a natural and intuitive way. The system does not need to rely on a specific language or complicated instructions, it can simply ask for a non-verbal response.

Furthermore, the Interactive Storytelling establishes an embodied connection between performer and audience. Interaction that is purely verbal allows the audience to interact off-stage. They can shout their responses towards the stage whilst staying anonymous among a crowd of spectators. However, non-verbal interaction does the opposite. The user has to present themselves to the system and has to be directly involved. In our setup, the user becomes part of the physical performance by providing their own hand movements. Their physical actions can alter the physical and fictional parts of the performance. We hypothesise that this direct, physical response heightens their involvement and that the act of deciding pollice verso (“with a turned thumb”) creates a feeling of power.

Limitations The interactive storytelling at the core of the response to RQ4 has been implemented and demonstrated (Sec. 4.8), yet it
has not been possible to conduct any in-person research during the COVID-19 pandemic, while alternative evaluation strategies are not feasible due to the system’s complex setup. Hence, this last response offers hypotheses and some suggestions for future work.

6.2 Conclusion

We conclude by taking one more look at our *Westworld* allusion from Chapter 1, which was presented as the pinnacle of robotic performances. In turn, we have proposed potential layers of performance, and then implemented and evaluated these with real robots. Surely, the resulting performances are far away from the realism and immersion that *Westworld* exhibits. This work has assessed all necessary ingredients that are required to create a robotic storytelling performance. We argue that, as long as humanoid robots do not exhibit a human-like level of appearance and motion, our suggested conceptual framework can be useful and creates a basis for embodied performances that are in need of a well-defined taxonomy of movement. Moreover, if robots eventually close the gap and become more indistinguishable from humans, we have demonstrated that most of the performative movement can be grounded in spatial schemas. These image schemas exhibit near-universal properties that emerge from interactions with the environment. As such, they also offer possible interactions that do not rely on language or cultural context. Moreover, we have contrasted the use of gestures with the use of these spatial movements in our evaluated performances. We identified that spatial movement and gestures work best when combined, but also that gestures and spatial movement, if used coherently, appear to be equally appreciated.

6.2.1 Contributions

This thesis has drawn from a variety of different research fields. Its contributions can consequently inform researchers from different domains.

6.2.1.1 Contribution: Human-Robot Interaction

There are two main contributions to the field of HRI provided by this research. First, we present the integration of a story generator (*Scéal-Stric*) with a robotic performance in a unified framework (*Scéalability*), which enables researchers to assess different variants of embodiment and interaction. We contribute not only the conceptual framework, but multiple empirical evaluations that show how embodiment and movements can increase audience appreciation. Moreover, the conceptual framework provides a comprehensive overview of possible
movement types from which other HRI research can draw. As outlined in Sec. 3.1.3, the taxonomy is relevant for storytelling and can be applied to other HRI subfields (e.g. dance, jokes, dialogue, etc.).

Second, we also contribute a detailed assessment of spatial movement and gestures. The implementation of gestures is, despite computational advances in gesture production and modelling (Sec. 2.4.2), often a laborious task that is prone to ambiguity. Moreover, the role of gestures can distort a performance if their cultural and social context is incompatible with the audience’s understanding of those gestures (e.g. Núñez and Sweetser [216] highlight cultural differences in gestural meaning). The contribution is rooted in a comparison of both, gestures and spatial movements, which shows that the perceived appreciation of a performance benefits from a coherent mix of the two together. We further demonstrate that implementing a large set of complex gestures for a robotic performance does not automatically contribute more to its appreciation than a few well-chosen spatial movements.

**Recommendation** We recommend that HRI researchers who are considering movement in their performances might start on a more primitive, schematic level of movement. Such primitive, spatial movements are much easier to implement and can convey meaning independently of cultural context. Additionally, if further movement is desirable, our taxonomy clearly assesses the possible implications and effects of any added movements. Since gestures can increase audience appreciation of a performance, we provide a taxonomy and various examples in the accompanying repository (see Sec. 6.2.1.4).

6.2.1.2 Contribution: Gestures and Image Schemas

The Gesture Module within our system is by no means a complete model of human gestural capabilities, since those capabilities are limited by the physical model of the robot. Nonetheless, this thesis contributes a practical robotic model of gesture use in the context of storytelling. Moreover, in addition to a mapping from hundreds of actions to gestures that can be used by other storytelling systems, it also generalizes the mapping through image schemas enabling researchers to include new actions or gestures into the system if they follow the conceptual framework. Its practical implementation and assessment builds on proposals by Spitale and Matarić [218] and by Ravenet et al. [219]. Moreover, it goes further than these works by extending gestural movements with spatial locomotion.

Additionally, this work contributes to ongoing research in the field of cognitive linguistics. It provides a physical model of several spatial metaphors that rely on image schemas. Since the model is built with an extensible knowledge base, it can easily be expanded with additional spatial metaphors. For example, we have demonstrated how
the NEAR/FAR relation can be assessed using our framework. Other research that focuses on conceptual blends, schematic integration or more complex combination of schemas can also be assessed with the framework.

LIMITATIONS There are many aspects of gestures which have not been investigated here. For instance, recurrent gestures are an important focus of gesture studies that we did not include as a category in our analysis. Moreover, we did not specifically quantify how different types of gestures have contributed to the performances. To this end, we have included annotations of the gestures (Sec. 6.2.1.4) by taxonomic classification, to make their future analysis more convenient for other researchers.

6.2.1.3 Contribution: Computational Creativity

Scéalability is a modular system that allows other researchers to augment their creative story generators with embodiment. We have noted that Scéalextric provides the semantic access points needed to render a story with spatial movements, gestures and dialogue, but other systems, such as PropperWryter [134] or MEXICA [109], might be adjusted in order to work with Scéalability. In the same way that Montfort et al. [107] created Slant by fusing MEXICA with Curveship [148], we envision that future work can fuse Scéalability’s modules with other story generators.

Various publications related to this work [1, 128, 258] have already contributed to the field of computational creativity by providing instantiations of embodied, creative systems. Here, Guckelsberger et al. [117] noted that our work is an example of “leveraging embodiment to affect the perception of their robot’s creativity” [117]. As such, this research contributes to the growing body of work that suggests that embodiment is a valuable and needed dimension in CC research.

RECOMMENDATION In accordance with Guckelsberger et al. [117] we recommend that the dimension of embodiment in computational creativity research needs to receive more attention. Our system shows how many other CC systems might benefit from an expansion into the physical domain. We agree with Guckelsberger et al. [117] that the definitions and terminology around embodiment in CC research need to become more coherent in order to understand the benefits and shortcomings of various embodiment types. To this end, we have provided a detailed explanation of what this work understands as “embodied” (see Sec. 1.3.1) and classified our individual implementations with respect to the terminology provided by Guckelsberger et al. [117] (see Sec. 2.6.3).
6.2.1.4 Research Repository

Lastly, we want to present the accompanying research repository hosted at:

osf.io/e5bn2/?view_only=2e30ee7e715342d59c371b5d30c014e0

The repository contains three data sets. Some entries are directly connected with the knowledge base of the Scéalextric story generator\textsuperscript{1}, while others correspond to aspects of the NAOqi operating system and are only useful if a SoftBank Nao robot is available. Nonetheless, most gestures come with descriptions, such that researchers can compare or recreate Nao-specific gestures.

**Nao\_Behavior\_Actions.tsv**: This is the central data sheet that connects the database of Nao’s built-in gestures with the database of Scéalextric’s story actions. In its first column, the data sheet refers to paths in the NAOqi operating system. The second column specifies the time (in seconds) that the robot needs to execute the gesture. The next three columns list the story actions that best fit the gesture (as described in Sec. 4.2.1.1).

**Gesture\_Emotions.xlsx**: In this data sheet various gestures are annotated with respect to the four valence scales presented in Sec. 3.2.3 in our Interpretation Framework. Moreover, each gesture is described with a short and a detailed explanation, which makes them accessible for other researchers and enables them to reproduce the gestures even if they do not have access to a Nao robot. Additionally, this data sheet also contains the gesture type labels.

**verbs\_schemas.xlsx**: This data sheet contains the suggested spatial dimension for each of Scéalextric’s story actions as described in Sec. 2.1.2.2.

6.2.2 Future Work

The story that this thesis has told is coming to an end, but the material that it has uncovered can make for a sequel. The implemented frameworks provide starting points for further research. Scéalability allowed us to assess specific gestures in a narrative context. Here, gesture researchers can take this work further by investigating other gestures, e.g. recurrent gestures in a narrative context. Due to Scéalability’s knowledge-base structure, the set of gestures can be changed and meaningful connections between narrative and movement can be established.

\textsuperscript{1} available here: https://github.com/prosecconetwork/Scealextric
Nonetheless, any new gesture must be scripted for the robot, and its repertoire of existing gestures must be carefully mapped to story verbs so that they can be used coherently. We have shown that the impact of spatial movement on a story-telling performance is comparable to that of more iconic gestures. The latter kind of animacy, as exemplified by our practical implementations, is rather ad-hoc in nature and laborious to implement. In contrast, schematization affords generalization. A few simple, cognitively-rooted principles [8] allow a few simple motion procedures to generalize over a great many story verbs, and conveniently allow for future extensions to the vocabulary. We have thus far implemented just one aspect of the conceptual scaffolding defined in [9], but anticipate further gains as we integrate other image schemas – such as vertical orientation and containment – into the Scéalability framework.

With regards to the system itself, we need to address the perception of machine-generated stories and give them a more natural feel. Plot and performance should elevate, rather than undercut, each other. Since both are inseparably connected in Scéalability, improvements to one will have effects on the other, enhancing the overall perception of the act.
All experiments including human participants have received an ethics approval exemption by UCD’s Ethics Committee. In order of the experiments:

- The crowd-sourcing study presented in Section 5.1 received exemption from ethics approval under the protocol number UCD HREC-LS, Ref. No.: LS-E-19-97-Wicke-Veale Exemption. The study is exempt as it used an anonymous survey that did not elicit identifiable data, or target vulnerable groups. All participants took part voluntarily, agreeing with the terms and condition of the crowd-sourcing platform. All procedures were in accordance with the ethical standards of the institutional and/or national research committee (UCD HREC-LS, Ref. No.: LS-E-19-97-Wicke-Veale) and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

- The crowd-sourcing study on Amazon Mechanical Turk presented in Section 5.2 received exemption from ethics approval under the protocol number UCD HREC-LS, Ref. No.: LS-E-19-125-Wicke-Veale. The study is exempt as it used an anonymous survey that did not elicit identifiable data, or target vulnerable groups. All participants took part voluntarily, agreeing with the terms and condition of the platform Amazon Mechanical Turk. All procedures were in accordance with the ethical standards of the institutional and/or national research committee (UCD HREC-LS, Ref. No.: LS-E-19-125-Wicke-Veale) and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.
The experiments on different types of embodiment in a performance required four different versions of the same questionnaire. Depending on whether a Nao or an Alexa was present, some items of the questionnaire have been changed, e.g. “The robot greatly contributed to the performance” was replaced by “The speaking device greatly contributed to the performance”. In condition 3, where both embodiment types were present, both items were included. For each item, the participant was asked to rate the statement on a Likert-Scale with five options from 1 (Strongly disagree) to 5 (Strongly agree). The following questionnaire shows all possible items (the items were presented in random order):

- I like the story.
- I like the performance.
- I had no problem understanding the performance.
- The story as a whole was humorous.
- The performance as a whole was humorous.
- The story as a whole was entertaining.
- The performance as a whole was entertaining.
- The robot added a humorous dimension to the performance.
- The speaking device added a humorous dimension to the performance.
- The performance hindered the understanding of the story.
- The performance enhanced the understanding of the story.
- The robot greatly contributed to the performance.
- The speaking device greatly contributed to the performance.
B.2 AttrakDiff - Items

In our two pilot studies and the main study, we have used the standardized AttrakDiff questionnaire by Hassenzahl et al. [22]. This questionnaire is widely used in user experience design and development. It can provide a measure for a product or system with respect to multiple dimensions. The ATT construct, presented in Appendix B.3, measures the overall attractiveness and is part of the hedonic quality construct. Two additional sub-constructs are part of the hedonic quality construct. First, the HQS construct measures the perceived ability of a product or system to meet the user’s desire for self-improvement. Second, the HQI measures the perceived ability of a product or system to communicate an identity to the user. Apart from these three hedonic qualities, the AttrakDiff questionnaire also contains the PQ construct that provides a measure of pragmatic quality of the product or system. All items of PQ, HQS and HQI are listed below:

PQ
- technical $\leftrightarrow$ human
- complicated $\leftrightarrow$ simple
- impractical $\leftrightarrow$ practical
- cumbersome $\leftrightarrow$ straightforward
- unpredictable $\leftrightarrow$ predictable
- confusing $\leftrightarrow$ clearly structured
- unruly $\leftrightarrow$ manageable

HQS
- conventional $\leftrightarrow$ inventive
- unimaginative $\leftrightarrow$ creative
- cautious $\leftrightarrow$ bold
- conservative $\leftrightarrow$ innovative
- dull $\leftrightarrow$ captivating
- undemanding $\leftrightarrow$ challenging
- ordinary $\leftrightarrow$ novel

HQI
- conventional $\leftrightarrow$ inventive
- unimaginative $\leftrightarrow$ creative
- cautious $\leftrightarrow$ bold
- conservative $\leftrightarrow$ innovative
- dull $\leftrightarrow$ captivating
- undemanding $\leftrightarrow$ challenging
- ordinary $\leftrightarrow$ novel
The first seven items are the ATT (for rating overall attractiveness) construct of the AttrakDiff questionnaire by Hassenzahl et al. [22]. Subjects answer questions of the form “The performance of the robots is ...” with a value ranging from 1 to 7 on the following AttrakDiff dimensions:

**ATT**
- unpleasant $\leftrightarrow$ pleasant
- ugly $\leftrightarrow$ attractive
- disagreeable $\leftrightarrow$ likeable
- rejecting $\leftrightarrow$ inviting
- bad $\leftrightarrow$ good
- repelling $\leftrightarrow$ appealing
- discouraging $\leftrightarrow$ motivating

Seven additional items have been added that specifically relate the embodied performance of robot narratives.

- The robots appear human-like
- The robots show their intentions
- I could act like one of the robots
- The robot mirrored how I would react
- I sided with one of the robots
- I am curious as to how the story continues
- The robots’ movements are appropriate to events in the story
EXAMPLE: GENERATED STORY

A=Miss Jane Marple(F); B=Reginald Jeeves(M); B-friend=Bertie Wooster(M); A-enemy=Hercule Poirot(M); N=Narrator

N: What if Miss Jane Marple fell in love with Reginald Jeeves? Miss Jane Marple and Reginald Jeeves had not met since they were young children.

R0 - fall in love with:
A: You are as beautiful as a picture
B: Why thank you
N: So at first, Jane fell deeper in love with Reginald than she ever had with Agatha Christie.

R1 - A are blackmailed by B-friend:
N: But Bertie Wooster used blackmail to manipulate Jane
A: Your friend needs to be silenced, Reginald.

R2 - borrow from:
A: Could you spot me some cash?
B: I think I can do that.
N: So Jane borrowed a lot of money from Reginald.

R3 - distrust:
N: But Reginald had every reason to distrust Jane
B: I don’t quite trust you yet
A: We must learn to trust each other.

R4 - are judged by:
A: I don’t have a good opinion of you
B: How dare you judge me.
N: So Reginald judged Jane and found her wanting.

R5 - accuse:
B: You are a villain through and through
A: Do you have any evidence to back that up?
N: Well, Reginald accused Jane of heinous wrongdoing.

R6 - offend:
A: Frankly I’m insulted
B: It wasn’t my intention to offend
N: Then Reginald offended Jane’s sensibilities.

R7 - refuse to pay:
A: I’m not paying you another penny
B: Don’t make me call in the heavy mob
N: So Jane left Reginald hanging for what was owed it.

R8 - are sued by:
A: My lawyers will hear of this
B: Aren’t you taking this too far?
N: Well, Reginald sued Jane for damages inflicted when solving crimes.

R9 - settle with:
A: Let’s put our differences behind us
B: I’m sure we can agree a settlement
N: So Jane buried the hatchet with Reginald.

R10 - resent:
N: But Jane resented Reginald for being manipulative, controlling and pious
A: I deserve much better than this
B: What are you grumbling about now?

R11 - dump:
A: It’s not me it’s you
B: So you’re moving on without me?
N: So Jane did not let Reginald down gently.

R12 - A are kidnapped by A-enemy:
N: Then Hercule Poirot kidnapped Jane and kept her hostage
A: Don’t just stand there. Call the police, Reginald.

R13 - B track down A-enemy:
N: So Reginald followed the crumbs and tracked Hercule down
A: I knew you could do it, Reginald.

R14 - A-enemy underestimate B:
N: But Hercule crucially underestimated Reginald
A: I think you may have the upper hand, Reginald.

R15 - A-enemy are lured by B:
N: So Reginald lured Hercule out of hiding
B: Soon I will spring my trap, Jane.
B murder A-enemy

N: Then Reginald murdered Hercule in his bed.
A: I would call that justifiable homicide, Reginald.

B are saved by

A: Save me. Save me
B: I got to you in the nick of time
N: After this, Reginald saved Jane's life.

B are thanked by

A: Thank you for everything
B: You're welcome
N: So Jane thanked Reginald effusively.

marry

A: I think we should get married
B: Yes let's get married
N: Then Jane married Reginald in Kent and they honeymooned in Mayfair.

argue with

N: But Jane argued with Reginald over the best way of rescuing aristocrats
A: You are so wrong
B: You wouldn't know the truth if it bit you.

are bored by

A: Is that a glazed look in your eye?
B: Talking with you is about as thrilling as wallpaper
N: So Jane bored Reginald to death.

love favor with

A: Why the cold shoulder?
B: You are in my bad books
N: Well, Jane was in the dog-house as far as Reginald was concerned.

cleat on

B: I won't be home for dinner tonight
A: I trust you implicitly
N: So Reginald fooled around and cheated on Jane.

stand up to

A: You are about as frightening as a wet kitten
B: Do not talk like that to me
N: Well, Jane called an end to Reginald's abuse.

are abused by

N: But Reginald abused and degraded Jane
A: I don't deserve this abuse
B: Stop over-reacting you snowflake.

are reported by

A: I'm putting you on report
B: Report me all you want
N: So Jane reported Reginald to the police.

are threatened by

A: Was that a threat?
B: I don't make threats just promises
N: Well, Reginald threatened Jane with talk of petrifying her with his earnest stare.

hide from

A: Catch me if you can
B: You can't hide forever
N: So Jane hid from Reginald in the darkest corner of Kent.

are found by

B: I bought you. I own you
A: Slavery is no longer legal
N: But in the end Reginald turned Jane into an indentured slave. Thereafter Jane would dance to Reginald's tune and jump at Reginald's orders.
BIBLIOGRAPHY


[227] Lena Ekberg. The mental manipulation of the vertical axis: How to go from ‘up’ to ‘out’, or from ‘above’ to ‘behind’. In Lexical


