The 5th Computational Creativity Symposium

In conjunction with the 2018 Convention of the Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB 2018)

4th April 2018

Understanding MEXICA: an analysis of an Engagement-Reflection system

Juan Alvarado[1] and **Geraint A. Wiggins**[2]

Abstract. We employ the Wiggins' Creativity Framework to explore the Engagement-Reflection (E-R) model of creativity the way it was implemented in MEXICA The purpose of this exploration is to have a deeper understanding of the E-R model by abstracting its constituent parts.

1 INTRODUCTION

MEXICA [4] is a system that generates short stories based on a set of constraints and a generative cycle of *Engagement* and *Reflection*. According to [4] MEXICA has shown good results, but it is still a system and a model that can be improved. For example, to include a frame of reference for the modification of its rules of operation and evaluation.

Creative computer systems have been generated for many purposes but are still being developed and can be useful tools to simulate and understand the creative processes in machines The E-R model has proven to be a useful tool for the implementation of creative systems and to understand the process of creative behaviours.

MEXICA, the computer model of creativity E-R and Sharples' [7] E-R proposal, on which the computational model is based, are complex. They consist of a set of many elements with complicated relationships. The computational model of creativity E-R [4] has left out several elements in the original proposal of Sharples but even so, the elements that have been considered are many and complicated to manipulate.

Using Wiggins' [8] framework, we intend to achieve a greater understanding of the components of MEXICA and thus be able to include improvements to the existing model and later also include features that have been left out.

2 BACKGROUND

2.1 Conceptual spaces

Boden [1] points out that there are *Conceptual Spaces (CS)* where creative ideas exist. She suggests that CS have origin in the culture of the creator and are any disciplined way of thinking that is familiar to (and valued by) a certain social group. For any CS there are rules or constraints which form it and there, new ideas (concepts) may be found. Boden [1] defines CS as a structured style of thought and she points out that conceptual spaces are normally learned from the culture. They include ways of writing prose or poetry; styles of sculpture, painting or music; theories in chemistry or biology; etc.

According to Boden [1] concepts in a CS can be found by *Exploration* and *Transformation*. She explains that by exploring the CS someone may be able to see possible concepts that had not been discovered yet. By transforming the CS its form changes because the rules or constraints have been changed and different concepts may be available to be found.

2.2 Mike Sharples' theory of Engagement and Reflection

Sharples [7] proposes an account of writing as creative design. Here, a writer generates new material by imposing appropriate (internal and external) constraints. External constraints may include an essay topic, previously written material, or a set of publisher's guidelines. Internal constraints (knowledge of the writer) may be interrelated concepts, genres, and knowledge of language [7].

Constraints are a combination of external resources and the writer's knowledge and experience. Sharples explains that a writing episode starts not with a single goal, but with a set of external and internal constraints. According to Sharples, by imposing constraints on a generative system it is possible to form what Boden [1] describes as a conceptual space.

Sharples explains that creativity in writing occurs through a mutually supportive cycle of *Engagement* and *Reflection*, both guided by constraints. An engaged writer devotes full attention to creating a chain of associated ideas and turning them into text. Reflection interacts with engaged writing through the activities of reviewing, contemplation, and planning. Reviewing involves reading the written material and interpreting it which leads to contemplation of it. This involves knowledge exploration and transformation of conceptual spaces and then planning takes the result as a source for creation of plans and intentions to guide a further period of engaged writing [7]. Figure 1 shows the E-R cycle.

Boden [1] argues that one important feature of creativity is the novelty of the artefact and Sharples agrees with this but he argues it is not enough to be novel, writing must also be appropriate to the task and to the audience, otherwise it degenerates into a ramble of nonsense [7].

In the next section, an example of the use of the Engagement and Reflection Model, applying the same principles of internal and external constraints, no specific objectives and a cycle of Engagement and Reflection is shown. This example is based on Sharples' ideas, explained above, but adapted to a computer model and implemented in a computer system called MEXICA.

[1] Queen Mary University of London, UK, email: j.alvaradolopez at qmul.ac.uk

[2] Vrije Universiteit Brussel, Belgium & Queen Mary University of London, UK, email: geraint.wiggins at vub.be

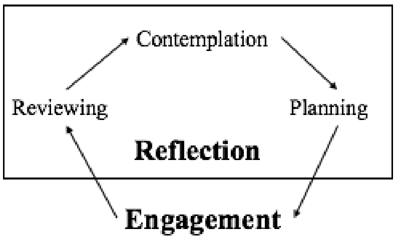


Figure 1. During Engagement the writer devotes full attention to create a chain of associated ideas and turning them into text. The box represents the Reflection stage. It interacts with Engagement through the activities of reviewing and contemplation of the material generated so far and then planning for the material to be generated back in Engagement.

2.3 MEXICA

MEXICA is an implementation of the computer model of creativity E-R proposed by Pérez y Pérez [4].

The main goal of MEXICA is to produce novel and appropriate short stories as a result of an Engagement-Reflection cycle without the use of predefined story-structures which was built with many modifiable parameters to experiment with the process of creating a new story plot [4]. In this model, there are no specific goals but there is also the notion of a set of internal and external constraints, also expressed by Sharples [7], that will guide the development of a story.

The system is divided into two phases.

1. An analysis of input files to produce knowledge structures,
2. An Engagement-Reflection cycle from which a new story will be obtained.

MEXICA needs two inputs provided by the user: a set of Primitive Actions (PA) and a set of Previous Stories (PS). The first gives the system the knowledge of everything that is possible to happen in a story, and the second are examples of stories, built with PA, that the system will use to build new ones. The PS set contains sets of actions in short stories which are supposed to be well organised and to be logical[3].

2.3.1 Constraints

MEXICA has the following categories of constraint:

Context Constraints are structures that represent the state of the current story. Events that occur in a story modify its context. Any action in a story has associated a set of consequences (postconditions) that modifies it. Actions added to the story in progress must respect the actual context. An appropriate action to continue the story must be selected by the system [4].

Knowledge Constraints are constituted by the experience, knowledge and beliefs of the writer. In MEXICA they are divided into three classes:

Abstract Representation encodes part of the knowledge necessary to retrieve an appropriate next action during the development of the story. For each action in each story in the Previous Stories, MEXICA obtains the story context, re-represents such context in more abstract terms and stores it in long-term memory as a new structure and links to that structure the following action performed in the story. MEXICA thus establishes a relationship between the structures in memory that represent the contexts and the next (logical) actions to be performed. The abstract representation establishes the universe of all possible events that MEXICA can retrieve from memory during Engagement [4].

Tensional Representation encodes part of the knowledge necessary to produce sequences of events in the current story that combine processes of degradation-improvement (conflict, complication and resolution). The tension produced in the reader is one of the central elements of fiction. In MEXICA a story is supposed to be interesting when it includes these degradation-improvement processes. [4].

Concrete Representation can be viewed as a copy of the previous stories file but in memory [4].

Guidelines constrain the material to satisfy the requirements of novelty and interest. During Reflection, MEXICA evaluates whether the material produced during Engagement meets the requirements of novelty and interest. As a result of this evaluation, MEXICA produces a set of guidelines, whose purpose is to influence the production of the material. A group of filters removes some of the actions, retrieved from memory during Engagement, that do not meet the guidelines [4].

General constraints include rhetorical and content constraints not included in the previous classifications. They are formed by a set of requirements that must be satisfied by all events retrieved from memory and are necessary for MEXICA to operate correctly [4]. They have two main objectives:

- To guarantee the flowing of the story. In MEXICA a story flows when a process of degradation or improvement occurs. The General Constraints remove those actions that do not contribute to the flow of the story. To ensure that a story flows the postconditions of any action retrieved from memory during Engagement must modify the context of the story [4].
- To prevent that the current story includes events that do not fulfil certain beliefs or basic knowledge on the subject of writing and the world in general or they are illogical[4] [4].

In MEXICA a story is a sequence of events or actions which are coherent and interesting. An action is an event in a story in which characters can participate. An action has pre-conditions and post-conditions, useful to give coherence to a story and to know the consequences of the execution of an action respectively.

When an action is executed, consequences arise and they generate a story context, this is how post-conditions are used. Story contexts are useful in MEXICA because from them it is possible to explore what can happen next in a story, they linked an action with the next. These linked actions are called logical actions. They are logical in the sense that an action is expected given a particular context. For example, given that a character A and a character B are friends (this is the context), a logical action in a story, that follows that context, could be: character A cures character B. Notice that the second action is expected given that they are friends. The fact that character B has not been wounded or is not ill is not part of the logic here. The reason to cure character B will be solved with the pre-conditions.

[3] The quality, interestingness and coherence of Previous Stories are system inputs too.

[4] In MEXICA, for example, it is not logical that the princess falls in love with the villain who harms her

Having an action linked to the next is not enough. In MEXICA it is also needed to link an action with the previous one in order to guarantee coherence, this is how pre-conditions are taken into account. In MEXICA a coherent sequence is that where all preconditions of all actions are satisfied. Taking the previous example, It could be said that, in order to have the second action a justification is needed, It means that character B can not be cured with no reason. So, for example, the action: character B cut his hand, can be inserted between the first and second action and so the precondition is satisfied and the story is coherent. Here we have an important concept in MEXICA: *coherence*. Coherence is a property of stories and they can only be coherent or non-coherent at a time.

Let us say we have the sets $Coherent\_Stories$ and $Non\_Coherent\_Stories$, then:

$$Coherent\_Stories \cap Non\_Coherent\_Stories = \varnothing$$

2.3.2 Engagement in MEXICA

During Engagement a sequence of actions linked by story contexts is produced. In order to do this, the PS are processed to get structures grouped by story contexts and related to next possible (logical) actions according to those contexts. Then, context associative structures are calculated from the story in progress MEXICA looks for a story context (like the one calculated) in memory and retrieves possible next actions linked to such a context to continue the story [4].

Contexts are calculated considering post-conditions; emotional links and tensions are part of them. The development of emotional links and tensions occurs due to an action in the set of PA. All actions have a set of post-conditions which are triggered when the action is executed, in that way a character A can develop an emotional link towards a character B due to the post-conditions of an action [4]. The same applies to the development of tensions.

The context of a character can suggest a possible action that follows a particular context, for example, if character A has a positive emotional link towards a character B and character B was wounded (this is the context), probably, the next action (based on the context) could be that A helps or cures B [4].

Once a set of possible actions has been retrieved from memory[5], Engagement selects one of the actions to continue the story appending it to the story in progress [4].

During Engagement MEXICA does not verify if the story actions satisfy pre-conditions. At this stage, as explained in [4], Engagement might produce a sequence of actions with unsatisfied pre-conditions (potentially non-coherent stories). But it might be the case that the sequence of actions is actually coherent. So, Engagement can produce coherent and non-coherent stories.

Let $\mathcal{C}_E$ be the conceptual space produced by the Engagement stage, then:

$$\mathcal{C}_E = Coherent\_Stories \cup Non\_Coherent\_Stories$$

2.3.3 Reflection in MEXICA

In contrast with Engagement, Reflection verifies pre-conditions for each action in the story in progress in order to produce a coherent story. Each Primitive Action (PA) has associated a set of preconditions. When unfulfilled pre-conditions are detected in the story in progress, MEXICA explores the space defined by all the PAs and fetches an action whose post-conditions satisfy such unfulfilled preconditions. Then, it inserts that action just before the event with the unsatisfied preconditions. Inserted actions can also have unsatisfied preconditions, which would cause the process to be repeated. Thus, whole episodes can be inserted to satisfy the preconditions of a single action. [4] The next is an example of a story with two actions. there is an unfulfilled precondition for one of the actions

Action 1: character A and character B are friends, thus they have a positive emotional link.
Action 2: character A cures character B.

It seems logical that A cures B because they are friends but there is a missing precondition, which is: B has to be wounded or ill first. An action should be inserted between Action 1 and Action 2 to fulfill such precondition. The inserted action might have unfulfilled preconditions. During Reflection, only coherent stories can be produced. Let $\mathcal{C}_R$ be the conceptual space produced by Reflection, then:

$$\mathcal{C}_R = Coherent\_Stories$$

The conceptual space of Reflection is a subset of that of Engagement as the first can contain only coherent stories and the second can contain coherent and non-coherent stories.

$$\mathcal{C}_R \subset \mathcal{C}_E$$

Reflection has access to elements in the Engagement Conceptual Space, even to those elements which are non-coherent. This is because one of the more important functions during Reflection is to guarantee coherence in the story, so, there is a mapping between elements in $\mathcal{C}_E$ and $\mathcal{C}_R$.

MEXICA also implements heuristics to test if the story in progress is interesting. When the set of PS is analysed the Tensional Representation has information about the way processes of degradation-improvement occur. MEXICA assumes that the stories in the set of PS supplied are interesting and so its Tensional Representation is a good example to follow [4]. MEXICA considers all the examples of Tensional Representation when it is evaluating whether the story in progress is interesting or not. Based on that information, and assuming that an interesting story includes degradation-improvement processes, when it is discovered that the story in progress does not increment tension, and it is supposed to do so, Guidelines are established in a way that the next execution of Engagement will favour retrieving actions able to produce tension to continue the story. So, the search strategy for Engagement can change with each cycle given that the guidelines are modified in Reflection.

Boden [1] suggests that novelty is one important characteristic of creative acts. Novelty is also considered in MEXICA and during Reflection, there are rules to assess novelty.

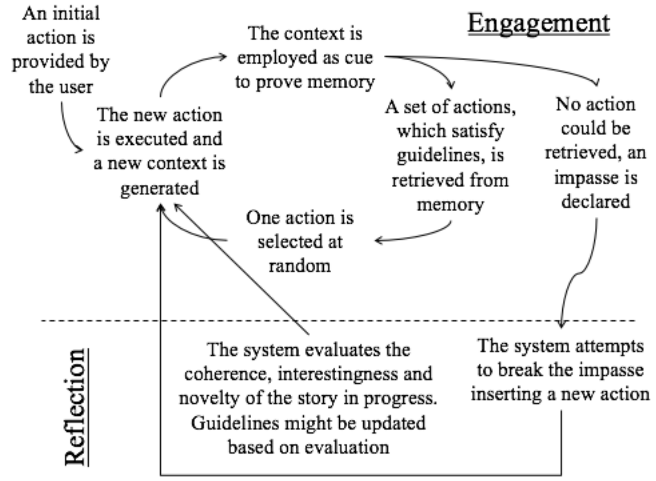
MEXICA verifies if the material produced during the Engaged state resembles too much any of the tales in the set of PS. The system has a parameter, called Novelty-Percentage, that determines the maximum percentage of similarity allowed between two tales; if it is exceeded, the guidelines are established to get a more original sequence of events during Engagement. This percentage has a default value of 50% and is modifiable by the user [4].

MEXICA is a computer system to experiment with the process of creating a new story plot. In order to achieve this, it has a number of modifiable parameters. There are many parameters and possible combinations thereof. Here, some of the parameters and purposes are mentioned. For example, the user can specify the number

[5] The set of possible actions retrieved from memory could be empty and in this case, an impasse is declared and Reflection is needed.

of times the complete cycle E-R will be executed. It also can be specified the number of actions that the Engaged state will append to the story before switching to Reflection. In order to assess novelty The Novelty-Percentage, which has a 50% by default, can be modified by the user. It is the maximum value of similarity allowed, so if the user sets a 90%, it means that the system is almost allowed to copy the Previous Stories. When the system is using the context as a cue to probe memory there are some parameters that can be modified to set the minimum and maximum percentages of context similarity to retrieve actions from memory.

Figure 2, similarly described by Pérez y Pérez [5] describes in a diagram the way MEXICA works.



Engagement
An initial action is provided by the user
The context is employed as cue to prove memory
The new action is executed and a new context is generated
A set of actions, which satisfy guidelines, is retrieved from memory
No action could be retrieved, an impasse is declared
One action is selected at random
Reflection
The system evaluates the coherence, interestingness and novelty of the story in progress. Guidelines might be updated based on evaluation
The system attempts to break the impasse inserting a new action

Figure 2. MEXICA has an initial action, it is executed and the context is calculated. The context is employed to retrieve a set of related actions from memory and one is appended to the story in progress. The context is calculated and the process is repeated a number of times defined by the user. In Reflection the system verifies the story in progress and the guidelines might be updated. Impasses are also broken in Reflection when no action was retrieved from memory (during Engagement) to continue the story.

2.4 Creative Systems Framework

Wiggins [8] formalizes the ideas on creativity expressed by Boden [1]. He argues that at first sight Boden's proposal lacks elements to use it in a consistent way, so he formalised the concepts in Boden's theory so they can be better applied.

Wiggins [8] builds a body of concepts starting from basic elements in computational creativity. He explains that artefacts are produced by a system (a creator), in a certain context, like P-creative acts explained by [1] which are related to the creator's mind and a culture that is familiar to a certain social group. Wiggins [8] indicates that novelty and value as features of artefacts produced by a system in its context and many authors coincide with this (e.g. [1, 4, 6, 2]).

Wiggins [8] defines different conceptual elements which are important in the analysis of a creative system.

Universe ($\mathcal{U}$) is a multidimensional space, whose dimensions are capable of representing anything and all possible distinct concepts correspond to distinct points in $\mathcal{U}$ [8]. Conceptual Spaces $\mathcal{C}$ defined by cultural agreements and for specific domains, in which concepts may exist, can be located inside the Universe $\mathcal{U}$.

Language ($\mathcal{L}$) is a common language from which framework's rules will be obtained.

Rules ($\mathcal{R}$) is a subset of $\mathcal{L}$ and are the rules which constrain a Conceptual Space $\mathcal{C}$; they define the nature of the created artefacts. In particular, in the societal context, they represent the agreed nature of what a concept is [8].

Traversing strategy ($\mathcal{T}$) is a subset of $\mathcal{L}$ and is the set of rules which allow us to traverse the Conceptual Space ($\mathcal{C}$). $\mathcal{T}$ defines the way a particular agent produces an artefact in practical terms [8].

Evaluation ($\mathcal{E}$) is a subset of $\mathcal{L}$ and is the set of rules for evaluation of concepts according to whatever criteria we may consider appropriate, they define the value of artefacts [8].

The Creative Systems Framework proposal [8] has some axiomatic points which are independent of the domain or type of the system.

Axiom 1 All possible concepts, including the empty concept, are represented in $\mathcal{U}$.

$$\top \in \mathcal{U}$$

.

Axiom 2 All concepts c_i represented in $\mathcal{U}$ are different

$$\forall c_1, c_2 \in \mathcal{U}, c_1 \neq c_2$$

Axiom 3 All conceptual spaces are strict subsets of $\mathcal{U}$.

$$\mathcal{C}_i \subseteq \mathcal{U}$$

Axiom 4 All conceptual spaces $\mathcal{C}$ include the empty concept $\top$

$$\top \in \mathcal{C}_i$$

$\mathcal{R}$ represents the rules which define the nature of the created artefacts. So, $\mathcal{R}$ constraints the Conceptual Space ($\mathcal{C}$) suggested by Boden [1]. Wiggins [8] explains that by using an interpretation function $[\![.]\!]$ it is possible to choose members of $\mathcal{U}$ which belongs to $\mathcal{C}$, assuming a well formed set $\mathcal{R}$.

$$\mathcal{C} = [\![\mathcal{R}]\!](\mathcal{U})$$

Similarly, for the search strategy $\mathcal{T}$, Wiggins [8] explains that another interpretation function is needed $\langle\langle ., ., . \rangle\rangle$ which, given three well-formed $\mathcal{R}$, $\mathcal{T}$ and $\mathcal{E}$ sets computes a function which maps two totally ordered subset of $\mathcal{U}$; c_{in}, and c_{out}. This function operates on members of $\mathcal{U}$ and not just on members of $\mathcal{C}$ because it is necessary to describe and simulate behaviours which are not completely well-behaved [8].

$$c_{out} = \langle\langle \mathcal{R}, \mathcal{T}, \mathcal{E} \rangle\rangle (c_{in})$$

Having different sets; $\mathcal{R}$ for the nature of the artefact, and $\mathcal{T}$ for the search strategy gives the possibility, explained by Wiggins [8], to have transformational creativity by transforming $\mathcal{R}$ into $\mathcal{R}'$ or $\mathcal{T}$ into $\mathcal{T}'$ or both. This is an important feature because, for example, changing $\mathcal{R}$ is a way to change the constraints of the conceptual space, and it might be called transformational creativity in Boden [1] terms and is equivalent to a paradigm shift. Changing $\mathcal{T}$ only affects the agent using that $\mathcal{T}$ [8] but the agreed nature of an artefact remains the same.

Wiggins [8] points out that in $\mathcal{C}$ there exist $\mathcal{C}_!$ and $\mathcal{C}_?$, concepts discovered and concepts not discovered yet respectively. Given $\mathcal{R}$ and $\mathcal{T}$ sets, some concepts in $\mathcal{C}_?$ may not be accessible, and even changing

$\mathcal{R}$ (transformational creativity in Boden's terms), they might remain non accessible. By changing the search strategy $\mathcal{T}$ the elusive concepts in $\mathcal{C}_?$ might be accessible. This means that by transforming the search strategy one may find by exploration concepts $\mathcal{C}_?$ in $\mathcal{C}$. Boden [1] suggests that transformational creativity is more significant that the explorational one. Wiggins [8] explains that this formulation shows that Boden's suggestion might not be true.

Wiggins [8] explains that Boden's idea of transformational creativity is to change the rules that define his conceptual space. Wiggins [8] defines two sets of rules, $\mathcal{R}$ and $\mathcal{T}$. Then the transformational creativity consists of changing either of them or both. The two sets are expressed in the language $\mathcal{L}$, which means that the result of the transformation(s) must also be in $\mathcal{L}$.

A syntax checker that selects $\mathcal{L}$ elements which are well formed is necessary. Therefore the transformations of $\mathcal{T}$ or $\mathcal{R}$ will be well formed in terms of any interpreter. Transformation means building new $\mathcal{L}$ subsets of the old ones [8].

Wiggins [8] explains that if we allow ourselves access to a meta-language, $\mathcal{L}_\mathcal{L}$, for $\mathcal{L}$, which can describe the construction of new members of $\mathcal{L}$ from old ones, we can pair it with an appropriate interpreter, to allow us to search the space of possibilities. The mentioned syntax checking task is structural meta-level (with respect to $\mathcal{L}$). $\mathcal{L}_\mathcal{L}$ can be used to describe this task too. Then, we can evaluate the quality of transformational creativity, with some Ω function [8].

Then it could be possible to specify interpreters, $[\![.]\!]$ and $\langle\langle ., ., .\rangle\rangle$, which will interpret a rule set $\mathcal{T}_\mathcal{L}$ applied to an agenda of potential sequences in $\mathcal{L}$, such an interpreter could work for both $\mathcal{L}$ and $\mathcal{L}_\mathcal{L}$ [8].

Then, the evaluation function Ω, could be express as a set of sequences $\mathcal{E}_\mathcal{L}$ in $\mathcal{L}_\mathcal{L}$ and use $[\![.]\!]$ to execute it [8]. The transformational creativity system can now be expressed as an exploratory creative system working at the meta-level of representation [8].

Wiggins [8] suggests that, for true transformational creativity to take place the creator needs to be in some sense aware of the rules he/she/it is applying. This self-awareness, suggested by [8], is what makes a creator able to formalise his/her/its own $\mathcal{R}$ and $\mathcal{T}$ in terms of the meta-language $\mathcal{L}_\mathcal{L}$. So without that self-awareness, a creator cannot exhibit transformational creativity [8].

Wiggins [8] points out that Boden's supposition that creative agents are well-behaved, in the sense that they either stick within their conceptual space, or alter it politely and deliberately by transformation may not be adequate. There are some situations in which agents may have a different behaviour which can be useful to analyse the system, they may also give information to switch to transformational creativity. They are grouped in [8] into the terms *Uninspiration* and *Aberration*.

Uninspiration occurs in three different forms:

Hopeless uninspiration: there are not valued concepts in the universe.

Conceptual uninspiration: there are not valued concepts in the conceptual space.

Generative unispiration: the search strategy of the creative agent does not allow it to find valued concepts

The first and second require redefining the universe and the constraints of the conceptual space respectively. The third indicates that the agent is not able, by the actual search strategy, to find valued concepts. A solution to this could be to modify the search strategy of the agent.

Aberration is a situation where a creative agent is traversing its conceptual space. The strategy $\mathcal{T}$ enables it to create another concept which does not conform to the constraints required for membership of the existing conceptual space.

Wiggins [8] terms this aberration, since it is a deviation from the norm as expressed by $\mathcal{R}$. The choice of this rather negative terminology is deliberate, reflecting the hostility with which changes to accepted styles are often met in the artistic world [8].

Aberrant concepts are very interesting because they are not part of $\mathcal{C}$ but the system might be able (by $\mathcal{T}$) to find concepts outside the constraints of the conceptual space defined by $\mathcal{R}$. The evaluation $\mathcal{E}$, of this concepts, has to be analysed carefully because, as expressed in [8] and it was also noted in [3], $\mathcal{E}$ should be capable of scoring the results of $\mathcal{T}$ even when they fall outside the set defined by $\mathcal{R}$.

3 AN EXPLORATION OF MEXICA

3.1 Introduction

Pérez y Pérez [4] describes that MEXICA operates not with a single goal or predefined story structures but a set of internal and external constraints. MEXICA inputs include a set of PA, which are all the possible actions that can be performed and a set of PS, which are examples of stories, both supplied by the user. They are related because the set of PS can only be formed with actions in the set of PA. According to Ritchie [6], the PS could be seen as the inspiring set.

MEXICA is divided into two main stages.

1. An analysis of input files to produce knowledge structures,
2. An Engagement-Reflection cycle from which a new story will be obtained.

In this analysis, the generation of knowledge structures is not relevant because they can be inferred from other elements. So, the generation of knowledge structures is a computation to make the operation of the system easier and more efficient. What is important is how the Engagement-Reflection cycle, with the input elements, is able to produce a story regardless the computations it uses. Now we apply the Creative Systems Framework [8] to MEXICA [4]

Universe ($\mathcal{U}$) is the set of short stories about the Mexicas.

Language ($\mathcal{L}$) is a common language from which rules will be obtained.

Rules ($\mathcal{R}$) is formed with the sets of constraints in MEXICA. Section 2.3 described that Engagement and Reflection do not have the same conceptual space because the concepts they can find are not always of the same type. MEXICA has a set of constraints but they are used to produce different results at each stage. So, there will be $\mathcal{R}_E$ and $\mathcal{R}_R$ sets of rules to produce $\mathcal{C}_E$ and $\mathcal{C}_R$, conceptual spaces for Engagement and Reflection respectively.

$$\mathcal{R}_E \rightarrow \mathcal{C}_E$$

$$\mathcal{R}_R \rightarrow \mathcal{C}_R$$

Traversing strategy ($\mathcal{T}$) represents the strategy by which an agent produces an output in practical terms, so they are the rules which define the way an agent will traverse $\mathcal{C}$. MEXICA produces a story through the strategy of an Engagement and Reflection cycle. The outcomes of each stage can be different because they do not perform the same operations to continue a story in progress. So, there are two sub-strategies, $\mathcal{T}_E$ and $\mathcal{T}_R$:

1. $\mathcal{T}_E$ to traverse the space $\mathcal{C}_E$ when the system is working in the Engagement state and when actions are being appended using story contexts and no pre-conditions of any action are verified.
2. $\mathcal{T}_R$ to traverse the space $\mathcal{C}_R$ when, in order to produce a coherent story, pre-conditions are verified. There is also a set of rules in the strategy $\mathcal{T}_R$ implemented to break impasses. An impasse happens when Engagement is not able to retrieve actions from memory to continue the story in progress.

Evaluation ($\mathcal{E}$) MEXICA does not have evaluation rules during Engagement but it has rules to evaluate novelty and interest implemented in Reflection. In the same way that there are two sets of rules $\mathcal{R}_E$ and $\mathcal{R}_R$ that define the conceptual spaces for the Engagement and Reflection stages, two sets can also be considered for the evaluation of concepts; $\mathcal{E}_E$ for Engagement and $\mathcal{E}_R$ for Reflection. Depending on the result of the evaluation, guidelines might be updated and the strategy $\mathcal{T}_E$ might change (this may be seen as strategy-transformation, $\mathcal{T}$-transformational creativity).

3.2 Concepts and rules

In a conceptual space $\mathcal{C}$, it is possible to find *concepts*. In MEXICA there are $\mathcal{C}_E$ and $\mathcal{C}_R$ conceptual spaces but they have different definitions of what a concept is.

- For $\mathcal{C}_E$ a concept is a sequence of actions related by story contexts.
- For $\mathcal{C}_R$ a concept is a coherent sequence of actions.

In MEXICA an *action* is an event that happens in a story which has characters, pre-conditions and post-conditions. MEXICA inputs include a set of Primitive Actions (PA) and a set of Previous Stories (PS) and there are grammars for generating such PA and PS sets (G_{PA} and G_{PS} respectively) defined in [4] and they generate the languages L_{PA} and L_{PS}.

$$L_{PA} = L(G_{PA}) \tag{1}$$

$$L_{PS} = L(G_{PS}) \tag{2}$$

In MEXICA [4] there are 4 types of constraints to develop a new story.

Context constraints are structures that represent the state of the current story.

Knowledge constraints are constituted by the experience, knowledge and beliefs of the writer.

Guidelines constraint the material to satisfy requirements of novelty and interest

General constraints group rhetorical and content constraints not included in any other group, useful so the system can work properly and basic beliefs about the world can be satisfied.

L_{Ctx_C}	→	Language of Context Constraints
L_{Knwl_C}	→	Language of Knowledge Constraints
$L_{Guidelines\_C}$	→	Language of Guidelines
L_{Gen_C}	→	Language of General Constraints

Table 1. Languages of constraints in MEXICA

Categories of constraints have particular definitions but it can be said that there is a common language to define them. Having the languages listed in Table 1 to define each category of constraints, the language of all constraints L_C could be represented by expression 3.

$$L_C = L_{Ctx\_C} \cup L_{Knwl\_C} \cup L_{Guidelines} \cup L_{Gen\_C} \tag{3}$$

Wiggins [8] explains that $\mathcal{R}$ and $\mathcal{T}$ sets are needed to have the rules for the conceptual space and the strategy by which it will be traversed. In order to build those sets, we need a common language to define them. There are languages which define PA, PS and Constraints, so, using expressions (1), (2) and (3) a common general language $\mathcal{L}$ can be:

$$\mathcal{L} = L_{PA} \cup L_{PS} \cup L_C \tag{4}$$

The set of rules $\mathcal{R}$, which defines $\mathcal{C}$, represent the agreed nature of what a concept is. $\mathcal{R}$ is a subset of $\mathcal{L}$ and can be described using (4). For this analysis, MEXICA has two sets of rules; $\mathcal{R}_E$ and $\mathcal{R}_R$, for $\mathcal{C}_E$ and $\mathcal{C}_R$ conceptual spaces. The expressions (5) and (6) can be produced.

$$\mathcal{R}_E \subset \mathcal{L} \tag{5}$$

$$\mathcal{R}_R \subset \mathcal{L} \tag{6}$$

By using an interpretation function $\llbracket . \rrbracket$, members of $\mathcal{U}$ which belongs to $\mathcal{C}_E$ and $\mathcal{C}_R$ conceptual spaces are chosen.

$$\mathcal{C}_E = \llbracket \mathcal{R}_E \rrbracket (\mathcal{U})$$

$$\mathcal{C}_R = \llbracket \mathcal{R}_R \rrbracket (\mathcal{U})$$

During Engagement, there is no evaluation of the story in progress and therefore it could be said that the set of evaluation rules $\mathcal{E}_E$, for concepts in $\mathcal{C}_E$, is empty. On the other hand, during Reflection the novelty and interest of the story is evaluated. For concepts in $\mathcal{C}_R$, the set of evaluation rules $\mathcal{E}_R$ is a subset of $\mathcal{L}$. Expressions (7) and (8) can be produced.

$$\mathcal{E}_E = \varnothing \tag{7}$$

$$\mathcal{E}_R \subset \mathcal{L} \tag{8}$$

There are also two strategies, $\mathcal{T}_E$ and $\mathcal{T}_R$ (Engagement and Reflection strategies respectively), useful to traverse $\mathcal{C}_E$ and $\mathcal{C}_R$ conceptual spaces. $\mathcal{T}$ is a subset of $\mathcal{L}$ and can be described using expression (4). Expressions (9) and (10) can be produced.

$$\mathcal{T}_E \subset \mathcal{L} \tag{9}$$

$$\mathcal{T}_R \subset \mathcal{L} \tag{10}$$

During Engagement new concepts (stories) are available by computing the context of the story in progress and then retrieving logical actions (related to the computed story context), filtered using guidelines. MEXICA appends one of the retrieved actions to continue the story in progress. This process is repeated the number of times the user defines.

Engagement does not verify preconditions when new actions are appended, this can produce non-coherent stories. However, in final outputs of MEXICA, coherence is a requirement which is fulfilled when the search strategy $\mathcal{T}_R$ is employed. This happens because $\mathcal{T}_R$

strategy contains a subset of MEXICA constraints where coherence rules can be found.

Reflection verifies preconditions of all actions in the story and if it finds them unsatisfied, actions which fulfil preconditions are inserted into the story until all preconditions are satisfied. During reflection also novelty and interest are evaluated and if an impasse[6] was declared, Reflection has procedures to break it.

Wiggins [8] explains that an interpretation function $\langle\langle ., ., . \rangle\rangle$ is needed, which given three well-formed $\mathcal{R}$, $\mathcal{T}$ and $\mathcal{E}$ sets maps two totally ordered subset of $\mathcal{U}$; c_{in}, c_{out}. The interpretation function is one, but there are two different sets of rules constraining the conceptual space $\mathcal{R}_E$ and $\mathcal{R}_R$, two sets $\mathcal{T}_E$ and $\mathcal{T}_R$ for the Engagement and Reflection search strategies and two sets $\mathcal{E}_E$ and $\mathcal{E}_R$ for evaluation of concepts. So, given a c_{in} input subset of $\mathcal{U}$, it is possible to obtain outputs (subsets of $\mathcal{U}$).

$$c_{out\_Engagement} = \langle\langle \mathcal{R}_E, \mathcal{T}_E, \mathcal{E}_E \rangle\rangle(c_{in})$$

$$c_{out\_Reflection} = \langle\langle \mathcal{R}_R, \mathcal{T}_R, \mathcal{E}_R \rangle\rangle(c_{in})$$

These functions can operate on members of $\mathcal{U}$ and not just on members of $\mathcal{C}_E$ or $\mathcal{C}_R$. They can describe and simulate behaviours which are not completely well-behaved as suggested by Wiggins [8].

3.3 Aberration in MEXICA

Wiggins [8] proposes the term aberration for the situation when an agent is able to create by $\mathcal{T}$ another concept which does not conform the constraints ($\mathcal{R}$) required for membership of the conceptual space. Pérez y Pérez [4] explains that a story is a sequence of actions but it is also important that the sequences are logical and coherent. A logic and coherent sequence of actions is that where the preconditions of all actions in the sequence are satisfied [4].

In Engagement there is no guarantee to produce a coherent story. When Engagement is appending actions to the story in progress using the story contexts, preconditions are not verified and that can produce potentially non-coherent stories. When Engagement receives a coherent story from Reflection, it appends a new action to the story and that operation can modify the story and, again, potentially produce a non-coherent story. When a non-coherent story is generated, that story does not conform the constraint of $\mathcal{R}_R$ and is, therefore, an aberrant concept for Reflection.

MEXICA has different ways to operate in which Reflection may not participate. If Reflection does not participate in the operation of the system then preconditions are not verified and the coherence requirement is not fulfilled, then potentially non-coherent concepts might be generated. If Reflection participates, when non-coherent concepts are given to Reflection, they are processed to produced coherent ones, they conform the constraint of $\mathcal{R}_R$ and, therefore, they belong to $\mathcal{C}_R$.

What is important to notice here is that as part of the MEXICA process the system is exploring options out of the scope of the main objective of MEXICA (out of the scope of $\mathcal{C}_R$ too, therefore aberrant concepts) which is to produce coherent stories.

3.4 Uninspiration in MEXICA

When Engagement is not able to find actions to append to the story in progress an impasse is declared. The search strategy $\mathcal{T}_E$ is not being able to create a new concept. This can be seen as uninspiration in Wiggins [8] terms. When the uninspiration is due to the generative process it can be fixed by changing the strategy $\mathcal{T}$. MEXICA can break an impasse by switching to Reflection. To break an impasse the context of the story in progress must change; if the previous context was not useful to find new actions, the new one might work. Once Reflection has appended a new action to the story in progress, the system switches to Engagement but now it can be considered that strategy $\mathcal{T}_E$ has changed.

4 CONCLUSIONS

After having analysed MEXICA with the Creative Systems Framework is interesting to note that MEXICA generates two different types of conceptual spaces and this is due to a marked difference in the result obtained in each stage.

During Engagement, MEXICA generates stories, adding new actions to the story in progress using the context the story generates. During this process, pre-conditions of the actions are not taken into account, which may give rise to actions that, although logical in the sense that they can be expected to happen, do not fully justify their appearance and provoke a story that lacks general coherence. This behaviour gives rise to the set of non-coherent stories.

During Reflection, MEXICA reviews the generated material and verifies that the preconditions of each action in the story are satisfied. When the preconditions have been met, the system has generated a story whose sequence of actions are all justified and the story is considered coherent. This gives rise to the set of coherent stories.

Due to the difference in the results obtained from each stage we have two conceptual spaces but is important to notice that it does not mean that they necessarily have distinct concepts. For this analysis it has been shown how the $\mathcal{C}_E$ has coherent and non-coherent concepts, (all kind of concepts for MEXICA) and $\mathcal{C}_R$ is a subset of $\mathcal{C}_E$, containing only the coherent ones.

Once Reflection finishes operating passes the turn to Engagement giving it a coherent story. The context of the story is calculated and is used to continue the story. Engagement applies only one operation; it adds a new action to the story and repeats this as many times as the user has defined. It could be the case that with the first action added, this coherent story changes to be a non-coherent story. The story moves out from $\mathcal{C}_R$ and becomes an aberrant concept for Reflection. This is interesting because we may analyse MEXICA creative behaviour based on unexpected aberrant concepts produced by exploration of $\mathcal{C}_E$. A further step can be the exploration out the boundaries of $\mathcal{C}_E$ which may be possible considering changes in the $\mathcal{T}_E$strategy that has been shown.

It should be noted that the mapping between the conceptual spaces $\mathcal{C}_E$ and $\mathcal{C}_R$ is not fixed because for a non-coherent story there could be more than one way to achieve coherence, also, there is more than one way by which a coherent story can become a non-coherent story (especially if the set of previous examples and primitive actions is large enough).

In MEXICA when evaluation based on $\mathcal{E}_R$ is applied (during Reflection), the intention is not to improve Reflection search strategy but to elaborate a plan for the next execution of the Engaged state. Based on this evaluation the Guidelines might be modified. The next execution of Engagement might have a modified version of the strategy $\mathcal{T}_E$. This will change the way the conceptual space $\mathcal{C}_E$ is traversed and this can be considered $\mathcal{T}$-transformational creativity in Wiggins [8] terms. The transformation occurs in the strategy and not in the conceptual space, so the agreed nature of the artefact ($\mathcal{R}_R$ and

[6] This happens when Engagement is unable to retrieve actions from memory to append to the story in progress.

$\mathcal{R}_E$) does not change but the way in which concepts are located in the conceptual space ($\mathcal{T}_E$) does. This will be part of a future work.

ACKNOWLEDGEMENTS

We would like to thank the referees for their comments which helped improve this paper and the National Council of Science and Technology (CONACyT) in México for its sponsorship

REFERENCES

[1] Margaret A. Boden, *The Creative Mind: Myths and Mechanisms*, Abacus, 1990.

[2] Simon Colton, 'Creativity Versus the Perception of Creativity in Computational Systems', in *Proceedings of the AAAI Spring Symposium on Creative Systems*, volume 8, pp. 14–20, (2008).

[3] Carlos León and Pablo Gervás, 'The role of evaluation-driven rejection in the successful exploration of a conceptual space of stories', *Minds and Machines*, (2010).

[4] Rafael Pérez y Pérez, *MEXICA : A Computer Model of Creativity in Writing*, Ph.D. dissertation, The University of Sussex, 1999.

[5] Rafael Pérez y Pérez, 'Employing emotions to drive plot generation in a computer-based storyteller', *Cognitive Systems Research*, **8**(2), 89–109, (2007).

[6] Graeme Ritchie, 'Some empirical criteria for attributing creativity to a computer program', *Minds and Machines*, **17**(1), 67–99, (2007).

[7] Mike Sharples, 'An account of writing as creative design', *The science of writing: Theories, methods, individual differences, and applications.*, (January), 127–148, (1996).

[8] Geraint A. Wiggins, 'A preliminary framework for description, analysis and comparison of creative systems', *Knowledge-Based Systems*, **19**(7), 449–458, (2006).

Afanasyev: A collaborative architectural model for automatic story generation

Eugenio Concepción and **Pablo Gervás** and **Gonzalo Méndez**[1]

Abstract. The present article focuses on detailing the characteristics of Afanasyev, an architectural framework for the construction of story generation systems through replaceable services. The basic idea behind this approach is the development of a collaborative environment for generating stories. This entails the inclusion of a common representation model to allow the interoperation between different story generation systems as a base for a collaborative environment to run an enhanced process of literary creation. In addition to this objective, this model aims at the development of a story representation formalism for creating a common knowledge base that can be fed in the future with the outcomes of new storytelling systems, without the need to adapt it to every system-specific representation model.

1 INTRODUCTION

Automatic story generation is a long-standing research field in the area of Computational Creativity (CC), which pursues the development of creative behaviour in machines [44]. A story generator algorithm (SGA) refers to a computational procedure resulting in an artefact that can be considered a story [20]. In other words, a story generation system is a computational system designed to tell stories. So, the terms story generation system and storytelling system can be considered equivalent.

From an architectural point of view, many automatic story generation systems have been traditionally designed as monolithic systems. This feature entails that a single application concentrates all the required functionality and assets. While this was a feasible solution for the earlier systems, mainly designed for research purposes and a limited-complexity functionality, nowadays it seems quite difficult to host the ideally expectable storytelling capabilities with such a model. So, as the story generation systems are becoming more complex, they are being designed in a much more modular way.

This paper introduces Afanasyev, a collaborative architectural model for automatic story generation which relates to a service-oriented architecture (SOA) [12, 38], and the microservices model [37, 45, 7]. It has been named after Alexander Nikolayevich Afanasyev, a Russian folklorist who compiled and published hundreds of Russian folktales [2].

The SOA paradigm provides a convenient framework for organizing complex software systems. In addition, the main contribution of the microservices architectural pattern to the service-based landscape is the development of highly distributed and decoupled applications. The application of this approach to the context of automatic story generation, along with the concepts taken from the API economy model [19], would allow the storytelling systems to create new functionalities and value.

This document is structured in four main blocks: a general review of the existing storytelling systems, with a special emphasis on collaborative story generation; a summarized statement of the problem; a detailed description of the proposed solution; and a final part focused on discussing some specific aspects of the solution and the conclusions.

2 BACKGROUND

The first story generation systems date back to the 1970s. The Automatic Novel Writer [27] is considered the first storytelling system. It generated murder stories in a weekend party setting. Its capabilities were quite limited, so the generated stories had an identical structure and the only variation came from the characters roles.

TALE-SPIN [33] was another of the earlier story generators. It was a planning solver system that wrote up a story narrating the steps performed by the characters for achieving their goals. TALE-SPIN generated stories about the inhabitants of a forest taking a collection of characters with their corresponding objectives as inputs. TALE-SPIN found a solution for those characters goals, and wrote up a story narrating the steps performed for achieving those goals.

Author [11] was the first story generator to include the author's goals as a part of the story generation process. Dehn considered that stories were mainly the result of a plot conceived in author's mind. In such a way, Author intended to emulate the mind of a writer. Conceptually it was a planner but, unlike TALE-SPIN, it used the planning to fulfill authorial goals instead of character goals.

Universe [29] was designed for generating the scripts of a TV soap opera episodes in which a large cast of characters played out multiple, simultaneous, overlapping stories that could continue indefinitely, without a closed end. Universe gave a special importance to the creation of characters, in contrast with Dehn's approach. It used complex data structures for modelling characters, using as input both predefined stereotypes and user-provided characterization.

Mexica [39] was developed as a computer model whose purpose was studying the creative process. It generated short stories about the early inhabitants of Mexico. Mexica was a pioneer in that it took into account emotional links and tensions between the characters as a means for driving and evaluating ongoing stories.

Fabulist [40] is a complete architecture for automatic story generation and presentation. Fabulist combines an author-centric approach together with a representation of characters intentionality, and an open-world planning for maximizing the quality of the stories.

Curveship [35] was a system for interactive fiction in which the user controls the main character of a story by introducing simple de-

[1] Universidad Complutense de Madrid, Spain, email: econcepc@ucm.es, pgervas@sip.ucm.es, gmendez@fdi.ucm.es

scriptions of what it should do, and the system generates descriptions of the outcomes of the character's actions. Curveship's storytelling approach differs from other story generation systems in the sense that it tells the story from different perspectives, without modifying the plot. For example, it makes use of a wide variety of techniques such as flashback, flash-forwards, interleaving of events from two different time periods, telling events back to front.

Regardless of whether the construction of the story plots relied on grammars [27], planning [33, 11, 29], or case-based reasoning [43, 22], a good part of the mentioned storytelling systems fitted the monolithic model. In addition to this approach, simulation-based systems [40, 35] were built mainly as distributed architectures. None of the aforementioned generators combined capabilities from other systems, nor considered the collaboration with others.

Slant [36] can be considered a remarkable example of storytelling systems working collaboratively for producing an enhanced outcome. It is an architecture for creative story generation that integrates several components from different systems: Mexica [39], Curveship [35] and Griot [25]. The latter is a collection of Computational Creativity related systems. The core of Griot is Alloy, a component which makes what its authors name "blending" [23]. Conceptual blending is an idea that comes from cognitive linguistics. It is a model of creative thinking in which two concepts can be integrated to form a new one. Namely, the thrust of this approach is the integration of different concepts in order to produce some creative results —for example, metaphors.

In a wider context, still within the computational creativity area, it is noteworthy the architecture proposed by Veale [44] for creative Web services. In an effort to accomplish both the academic and the industry needs, he proposes a solution for enhancing computational creativity systems by introducing an architectural model which categorizes the services according to their function in the application structure.

After the prior analysis of a representative subset of the existing storytelling systems, it seems quite clear that every system has been designed according to certain operational expectations that they are able to accomplish, but they can hardly produce stories beyond their predefined target model. Hence, it is quite uncommon to find a single story generation system producing stories that combine different narrative rhythms or that deal with diverse motifs in the thematic aspect.

3 STATEMENT OF THE PROBLEM

What makes a story captivating? The basic elements of a story have been largely analysed by classic Narratology [4, 3, 32]. The plot is an essential element in a story, but so are the characters depiction, the narrative discourse, the rhythm, the emotional arc and many others. All these elements produce an effect in the people watching a play or a film, reading a novel or listening to a narrator. The wise arrangement of all these components, adapting the length of each scene to the most convenient one, varying the speech and description passages, choosing the right timing for the key events and remaining faithful to the theme, help to create movement, tension and emotional value in the development of the story.

Despite the efforts made in the field of automatic story generation, the stories written by humans are considerably more complex than those generated by computational systems. Consider as an example any classic novel: they contain a main plot, several subplots, every chapter can be focused on a different theme, there are changes in the rhythm of the narration, there are passages that focus on a particular character and ignore the rest, and many other features that help to keep the readers attention in the narration. The existing storytelling systems are capable of creating a single-themed story, with a single narrative structure and a specific rhythm.

Coupled with the intrinsic limitations of the generation model, the monolithic architecture of many existing systems introduces an additional limiting factor.

Considering the collaboration between different storytelling systems as a simple way of generating more natural stories, it seems appropriate that a solution could involve using different systems, generating different types of content according to their capabilities. Due to the fact that a monolithic design hinders the collaboration with other systems, this paper considers the use of several systems working collaboratively for achieving the generation of richer and more complex stories by providing a service-based framework for automatic storytelling. This approach would allow to combine different services from different story generation models –or systems, so the outcome would be closer to the diversity of narrative resources that characterize the stories created by humans.

4 PROPOSED SOLUTION

Many of the existing systems have been designed as monoliths, which make the collaboration between them a really complex challenge. This happens because almost every system duplicates a considerable part of the common storytelling functions. If every storytelling system broke its architecture into finer-grain components, such as microservices, these components could be used separately and evolve independently.

The basic idea of the proposed solution can be seen as one of those toddler toys in which they have to classify different pieces by matching the shapes and drop every block through the sorter. In this case, the model supports the use of different types of automatic storytelling services, as long as they can implement every required interface.

Afanasyev is basically a collection of microservices orchestrated by a high-level service. The overall ecosystem can be considered a small storytelling API Economy [19]. Each service exposes their capabilities as REST-based API [14] and it understands and generates JSON messages. Due to the fact that the inner logic of any microservice can come from a different storytelling system, its interface must be adapted to this new purpose. This is the reason why Afanasyev includes the definition of the common REST interfaces provided by the services and leaves to every particular system the details of the implementation. This approach introduces several benefits. First of all, the whole architecture is highly decoupled. This means that every service is implemented and deployed separately, and it can evolve independently from the others. Another benefit of this model is that it can be extended in the future, by adding new microservices to the ecosystem without affecting the others. And finally, a very important feature, the ease of integrating a new system. To add a new storytelling system to the ecosystem, simply entails to implement at least one of the microservices interface, and registering it in order to be considered by the Story Director during the generation process.

From a certain point of view, the operation of Afanasyev may evoke the idea behind *Hopscotch*, a novel by Cortázar [10], whose chapters can be read in different order, giving rise to a good number of differing valid interpretations of the resulting plot. In this case, the architecture provides the structure and function, which must be covered by the different microservices that implement each API. This allows to use parts coming from different generating systems in a combined way, or to reconstruct a complete generator according to the architecture provided by the framework. An early approach to this model was proposed as part of a wider API-based collaborative environment [5].

The development of Afanasyev entails two main tasks: the definition of a shared knowledge representation model and the design of a microservice-based architectural environment. Both are addressed in the following sections.

4.1 Common knowledge representation model

In order to allow the combined operation, the microservices of the framework require a common representation model for stories. The knowledge required to generate stories depends heavily on a number of factors. One of these key factors is the system architecture. The components that participate in the generation process condition the structure of the knowledge. For example, in the case of storytelling systems built over planners, it is necessary to keep knowledge concerning states, preconditions, actions, effects of the actions, etc. Grammar-based story generators require a complete representation of the applicable rules for creating their stories. Simulation-based storytelling requires a detailed typification of the characters and their relationships. On the other hand, there is a common element for every storytelling system that can be interchanged: the story, which is the end product of the generation process.

The proposed representation model [6] focuses on the knowledge that is directly related to the story, instead of that related to the generation process, which would be hard to export between different systems. This model is strongly influenced by the components of narrative identified in the classic Narratology [4, 3, 32]. These concepts and structure are enhanced by various storytelling-related computational concerns.

The resulting representation model is summarized in Figure 1.

The model has been designed as a hierarchical structure, in which the root concept is the **story**. Most of the leaves of this tree-like structure are assertions representing a piece of knowledge. These assertions are expressed by means of sentences in a Controlled Natural Language (CNL) [41]. The use of a CNL for representing knowledge in storytelling systems has been proposed by the authors in earlier papers [8, 9]. The main advantage of using a CNL is that the concepts referred in the assertions can be expressed by domain experts in the knowledge base and then they can be translated to the variety of formal representations used by the various services. This feature allows the definition of rules in a system-agnostic language, useful not only for expressing the different concepts involved in the story, but also for exchanging these knowledge resources across the different storytelling services.

A story represents what both intuitively and narratologically can be considered a story, that is, a narration of the actions performed by the characters and the events happening in a setting. A story is composed by two main elements: the plot and the space.

The **plot** is represented as a sequence of scenes. A **scene** is conceptually related to the division of a play, that represents a single episode inside the plot. It is clearly conditioned by the time division, which means that it is a sequence of events that happen during a time frame. From a spatial point of view, it is also constrained to take place in a single spatial frame —considering the spatial frame definition mentioned before. So, the scene is composed by a sequence of **events**, that can be actions or happenings. An **action** is an act performed by one or more characters in the story, generating consequences. The resulting consequences of every action are expressed as a modification in the global state of the space —considering it as the whole setting and the existents. A **happening** is an event that happens in the plot, as an accident or as a consequence of a prior action or happening. A happening can be natural —it rains— or artificial —a car accident. Regardless of the type of event, both are characterized by their impact in the story world. This is represented as a pair of states: the previous state and the later state. Each state is represented by a set of assertions, expressed in a CNL.

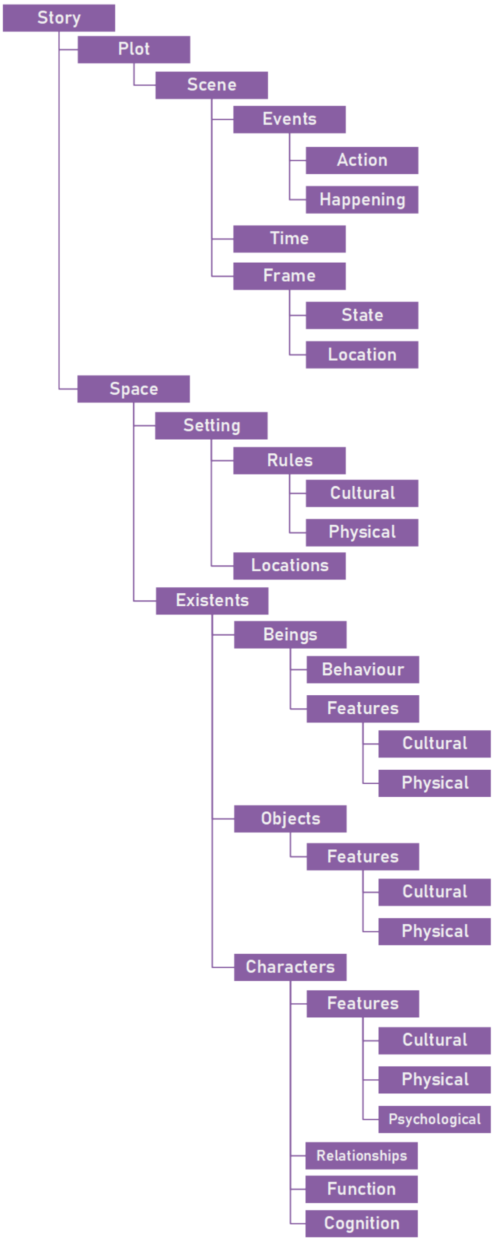


Figure 1. Stories common representation model.

The **space** encompasses the whole universe in which the plot is taking place and also all the places, beings and objects of which existence the characters are aware of, regardless of these elements are real or fictitious. The representation model considers that the space is composed by the setting and the existents. The **existents** are the whole set of actors that take a part in the story. They can be characters, living beings —an animal—, and an object in the setting. The two last types are mainly defined by their physical features and their cultural significance in the story. The **characters** are the most relevant, and also the most complex to represent, elements in the story. The proposed model considers not only their physical, psychological and social features, but also their cognitive-related characteristics. The cognition of the characters is represented in a very detailed manner due to its importance for ensuring story consistency and characters liability. The aspects considered have been chosen after analysing those used by the existing storytelling systems [42, 31, 11, 30, 34, 39] and theoretical studies about Narrative [3, 32]. So, the representation of cognition includes the following facets:

- Goals: The goals are the results or achievements toward which the character effort is directed. The model considers two types of goals: conscious and unconscious. In the first case, the character is aware of them, in the second, they drive the character's actions, but he/she is not aware of them.
- Intentions: The intentions refer to the general plan that every character has, and the drive for his/her actions.
- Knowledge: Despite the characters act and interact in the same space, every single character could have different levels of knowledge concerning it. That means that the characters are not considered to be omniscient. This knowledge can evolve over the time, so characters can be acquiring or discarding knowledge as the story develops.
- Memories: Unlike the general knowledge, the memories refer to some past situations that have relevance in the story. For example, a memory can be referred to a past scene in which the character took part.
- Beliefs: The beliefs are a very subjective part of every character's cognition. They refer to facts about the world which the character considers as axioms, regardless of they are true. They can be part of the character's cultural or religious code, or simply originate in a particular misconception of the world.
- Dreams: The dreams represent the unconscious aspirations of the character. He/she may not be aware of them, but they can operate at a subconscious level and inspire his/her intentions.
- Fantasies: The fantasies are product of characters' imagination. They are beliefs or notions based on no solid foundation, a fact which the character is perfectly aware of. They represent aspirations that the character considers unreachable, but he/she enjoys thinking about them.
- Emotions: The emotions are related to the feelings of the character. They are usually influenced by the relationships that the character establishes with the others, and the evolution of them during the story.

Another relevant element of character's representation is the **function**. The idea is to provide a way of representing the main two approaches concerning the role of the characters in the plot. There are models that consider the plot as the result of characters interactions in a simulated story world, but there is another line of thought which considers that characters are subordinate to the narrative action. There are storytelling systems [21] that describe characters in terms of a structure based on their roles in the plot. Hence, the function tag refers to this approach and provides a way for linking the functional role of the character to the underlying structure of the story.

The **setting** is a combination of a set of physical —or virtual— locations in which the action of the story takes place, and the set of cultural and physical rules that govern the story world. The **locations** can be considered the scenario in which every scene that composes the plot takes place. So, as shown in the model, every scene links to its corresponding location.

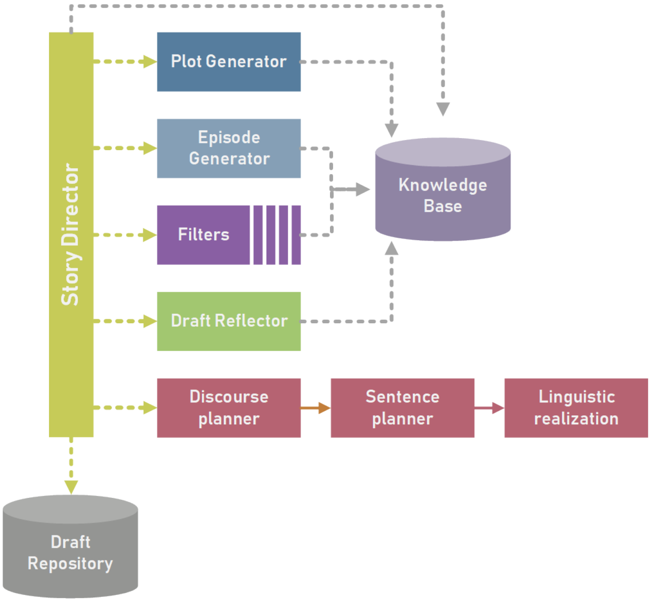


Figure 2. Architecture of Afanasyev.

4.2 Architecture of Afanasyev

The architecture of Afanasyev is based on a set of key microservices that provide the essential capabilities for story generation. Every microservice publishes an interface according to the REST model [14]. The joint operation of the microservices ecosystem is managed by the Story Director, which acts as an orchestrator of the services activity. It will request the APIs of the different services according to the steps of the generation process. This process will proceed iteratively, generating drafts that will be refined in each pass, until the established criteria for story completeness are met.

The main microservices in Afanasyev, depicted in Figure 2, are the following:

- Story Director
- Plot Generator
- Episode Generator
- Filter Manager
- Draft Reflector
- Discourse generation services (Discourse Planner, Sentence Planner and Linguistic Realization)

The key component of this framework is the **Story Director**, the inner architecture of which is depicted in Figure 3. It is strongly influenced by the Domain-Driven Design (DDD) principles [13].

The distinction between Application services and Domain services is precisely due to DDD. An application service has a clearly distinguishing role: it constitutes the environment for executing the domain logic, orchestrating the calls to the other components of the architecture: domain services, gateways and repositories. Domain services are only focused on performing domain logic which does not involve managing entities (Repositories) or calling external compo-

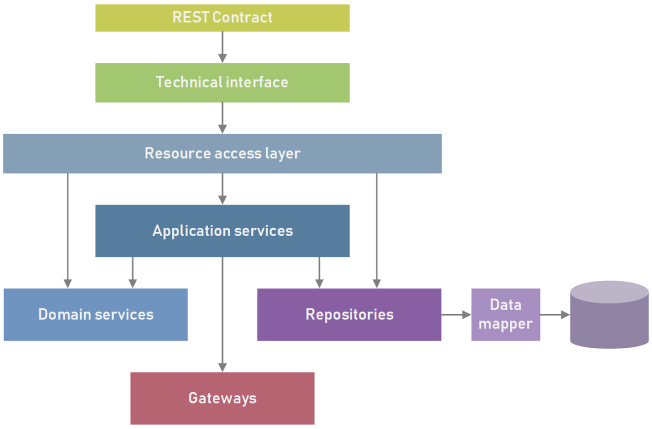


Figure 3. Story director architecture.

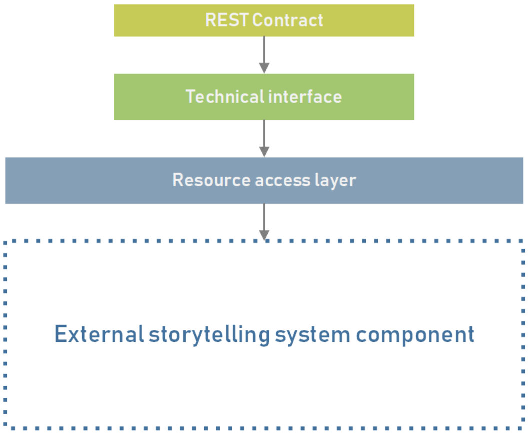


Figure 4. Marker microservices architecture.

nents (Gateways). So, they can rather be seen as components that provide procedural functionalities.

The Story Director has a clearly defined REST interface. The technical interface layer provides the logic necessary for implementing the communication-related requirements, allowing the isolation of the remaining components from them. The resource access layer provides a uniform interface for accessing the stories managed by the Story Director.

The repositories have been designed according to the Repository pattern [15], which provides a convenient abstraction for managing persisted objects. The inner database of the Story Director is an auxiliary store for persisting the life cycle of the ongoing drafts.

Persistence in Afanasyev is mainly composed by two stores: the Draft Repository and the Knowledge Base. The Draft Repository is a database that stores the ongoing drafts. The current implementation of this component is based on a NoSQL database [24] (MongoDB [1]). The knowledge base has the task of preserving all the knowledge related to concepts, relationships between concepts, rules, etc. It is a knowledge base generated from the contributions of the involved story generation systems. This model of knowledge syndication allows to increase the shared set of concepts each time a new system joins the ecosystem. Hence, every contributor performs an initial load expressing its rules by means of a controlled natural language expression. Namely, the current version counts on Attempto Controlled English (ACE) for this representation [18, 17, 28]. The use of a CNL for representing the knowledge allows the model to abstract from the programmatic representation used by each particular system, and to provide a greater robustness and consistency to the system architecture.

The **Plot Generator** main task is generating the complete plot structure. This includes the generation of the sequence of scenes that constitute the plot, the preconditions and postconditions that constrain every scene, and the articulation of the story in a high level.

The **Episode Generator** is in charge of developing the details of what happens in every scene of the plot. It must consider the preconditions and the postconditions defined for the scene by the Plot Generator, in order to create a scene detail that is consistent with them.

The **Filter Manager** is a service devoted to filter the population of generated drafts in order to select only the most promising stories, in terms of narrative tension or suspense. It is a very convenient tool for avoiding an explosion of irrelevant draft variants during the episode generation.

The **Draft Reflector** inspects the drafts for deciding if they are finished stories or if they must be improved in another iteration. For example, it checks if all the scenes of the plot have been detailed.

From a technical point of view, the **Plot Generator**, the **Episode Generator**, the **Filter Manager**, the **Draft Reflector** and the text generation services are basically marker microservices, with a predefined REST interface and a set of common architectural components. They are expected to be implemented by the particular story generation systems that collaborate in the generation process.

The internal architecture of these microservices, as Figure 4 shows, share partially the design of the Story Director. The components directly related to the intercommunication has been structured in the same way. They have a common layer for REST contract, with their corresponding technical interface, and the mandatory CNL mapping components. In their case, the resource access layer acts as an anticorruption layer [13] that isolates the inner logic of the service from the common framework infrastructure.

4.3 System operation

Afanasyev operates iteratively. Firstly, it generates a draft that will be completed by the various existing services in the architecture. The Story Director acts as the central component, orchestrating the requests to the different microservices. Table 1 summarizes the REST operations related to each microservice. The first step is always performed by the Plot Generator, which generates the basic structure of the plot. This provides a first basis for the story, with the sequence of scenes that make up the plot. Each scene is characterized by a previous state (precondition) and a later state (postcondition) of the world in which the action takes place. Every state is a collection of statements relating to the characters, living beings, and objects that exist in the story. In addition, each scene is associated with a specific setting. This setting is a reference to the list of existing settings defined in the story space.

Once the first draft is generated, the Story Director will persist it in the Draft Repository and then it will request the Episode Generator to generate the detail of what happens in each scene. For this, the Episode Generator receives as a parameter the draft, and the identifier of the scene that it must develop. Again, in this process the previous

Table 1. Afanasyev microservices operations summary.

Service	Method	Input	Output
Story Director	POST	Characters list, Pre/post spec	Story
Plot Generator	POST	Characters list, Pre/post spec	Draft
Episode Generator	PUT	Episode UUID, Draft	Draft
Filter Manager	POST	Episode UUID, Draft	Episode curves
Draft Reflector	POST	Draft	Draft Evaluation
Discourse planner	POST	Story	Text (NLG)

and final states of the scene are extremely important, since they will provide information to the Episode Generator about what can and can not happen in the scene. That is, the Episode Generator will only generate solutions for the scene that are coherent with the previous and final states, discarding the rest. The output will be a collection of possible continuations of the story, namely, a collection of drafts. Once again, every generated draft will be saved by the Story Director in the Draft Repository.

In the next step, the Story Director will request the Filter Manager to apply a sequence of filters on the generated drafts, and discard those considered as not promising. The number of filters is variable and they will always be applied in order, being the first the most important. Some of these filters can focus on aspects such as narrative tension or suspense. They allow us to make the stories more interesting by selecting those drafts that best fit the proposed parameters. The Story Director will remove the discarded drafts from the Draft Repository.

The final step in each iteration is provided by the Draft Reflector, which analyzes each of the drafts in progress and decides if the story has been completed, and therefore, stopping being a draft to become a finished story. The last step for the finished story is to generate the text in Natural Language. This task is performed by the discourse generation services, that work sequentially: Discourse Planner - Sentence Planner - Linguistic Realizer.

The whole operation of Afanasyev is summarized by Figure 5.

The main advantage of this operation model is that the components of the architecture are basically slots that can be fitted by different services that follow different strategies. For example, the criteria for story completeness depend totally on the implementation of the Draft Reflector. Furthermore, the architecture admits the coexistence of various draft reflecting services that can be called by the Story Director according to higher order criteria. This feature provides a wider variety of behaviours during system operation.

5 DISCUSSION

Unlike previous approaches to collaborative story generation [36], Afanasyev is not geared towards the ad hoc integration of specific pre-existing systems, but rather to provide a general service-oriented framework that allows the construction of different storytelling systems by assembling components from various systems (or from only one, in the simplest case).

From an architectural point of view, Slant consists of a blackboard architecture [26] and a shared XML based story representation, which allows different storytelling systems or components to contribute to the story generation. This approach entails that every contributing system can access a shared working draft and enrich it. As part of the generation process, Slant provides mechanisms for selecting the most convenient contents in every iteration and deciding when to finish a ongoing story.

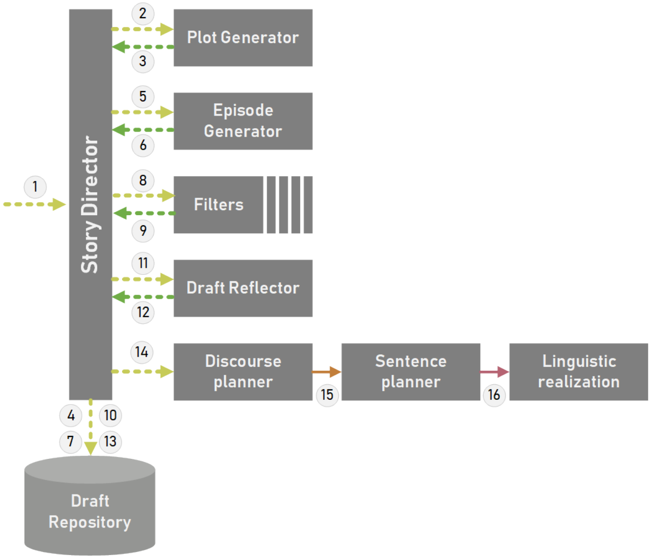


Figure 5. Operation of Afanasyev.

In contrast, in the service-based approach of Afanasyev, only the Story Director manages directly the ongoing drafts. The rest of the services can be invoked only according their interface and their operation is always orchestrated by the Story Director. This modularization, derived from the use of a microservices architecture, is not the only interesting feature. First of all, every service can be instantiated several times, and even exhibit different behaviour according to its configuration. For example, there can be several instances of the Plot Generator service, each with a different inner implementation, and the Story Director can request them to generate a draft in order to have a wider variety of plots. The same applies to the Episode Generator and the Draft Reflector services. In an API ecosystem, different versions of the same service can live together and be consumed independently. So, it would be possible to have an Episode Generator instance implemented from certain storytelling system, and another Episode Generator instance implemented from a different storytelling system.

Another interesting feature is that the architecture can be easily extended. The operation of every microservice in Afanasyev is completely independent from the others. If we wish to introduce a new microservice in the architecture, the only component that would require to be adapted would be the Story Director —in order to include this new service in the generation process that the Story Director manages.

Also, the Filter Manager service has been designed as an extensible sequence of filters that are applied in order to modify the draft received as a parameter. These filters are related to the degree of interest of the draft (for example, narrative tension and suspense). Adding a new filter simply requires to register the service that implements it into the Filter Manager.

Due to the coexistence of rules from various systems, it is assumed that there is no guarantee of consistency in the knowledge base. Achieving a full strict consistency would entail the validation of every new rule against the set of rules previously stored, and deciding which rule must be preserved in case of conflict. Another option would be the segmentation of the rules according to their origin as namespaces that would be locally consistent.

In the current version of Afanasyev it has been accepted that there can exist rules mutually inconsistent, even mutually exclusive (e.g. “Magic does not exist” and “Magic exists”). The reason for this choice is to provide an open perspective during generation and leave it up to the human evaluator to decide whether the generated story is more interesting despite the potential inconsistencies.

A future option could be including non monotonic reasoning [16], providing default rules, or even developing truth maintenance mechanisms (e.g. “Magic does not exist for muggles”). These approaches are left for later as a future work due to their complexity and importance.

In addition to the above, the use of a domain-specific glossary would serve not only for establishing a proper definition of the knowledge domain, but also for reducing the risk of polysemy. One of the potential issues with CNL is that they are not specifically designed to address word sense disambiguation. The CNL are usually focused on analysing only the key words that are relevant for building the discourse representation structure, so it will be necessary to validate the portability of this representation over the different services.

6 CONCLUSIONS AND FUTURE WORK

Afanasyev is an architectural framework for building story generation systems, not a story generation system itself. The main advantage of the Afanasyev model comes from its modularity. By means of a flexible architectural structure, a common knowledge representation model and a set of services with well-defined interfaces, the proposed framework eases the development of collaborative story generation ecosystems. Different systems can work together in a cooperative story creation process by providing one or more services according the required types of service —plot generator, episode generator, draft reflector and text generation services. Some of these services might take the form of user interfaces to allow human intervention, so it also encourages the development of co-creation models.

In the present version of Afanasyev, for every draft processed in every iteration, there can be generated several continuations that are added to the population of drafts to process during the next iteration. On the generated population, a reflection process is applied by means of the Draft Reflector microservice, and the drafts that it considers already finished are marked as stories. This process continues until all drafts are marked as finished or a limit of iterations is reached (to guarantee completion). In the face of future work, the development of a service that helps to decide what is the most appropriate level of detail in each of the scenes is still pending. This aspect can be provided in a first instance by a human —applying a co-creation model—, but it would be perfectly evolved to introduce a component for automating this task.

In the short term, the next steps are focused on adding the capabilities of different existing storytelling systems such as Charade [34], STellA [31] and PropperWryter [21]. In a first approach, the goal is demonstrating the ability of the framework for reconstructing existing systems and the adequacy of the knowledge representation model for expressing the needs of various existing systems. Next, the objective would be the implementation of a real collaboration between different systems by mixing services from different origins.

ACKNOWLEDGEMENTS

This paper has been partially funded by the projects IDiLyCo: Digital Inclusion, Language and Communication, Grant. No. TIN2015-66655-R (MINECO/FEDER) and InVITAR-IA: Infraestructuras para la Visibilización, Integración y Transferencia de Aplicaciones y Resultados de Inteligencia Artificial, UCM Grant. FEI-EU-17-23.

References

[1] Mongodb official site. `https://www.mongodb.com/`, 2017. [Online; accessed 29-December-2017].

[2] Alexander Afanasyev. Russian fairy tales, tr. norbert guterman, 1945.

[3] Roland Barthes, *S/Z: an essay*, Siglo XXI, 1980.

[4] Seymour Benjamin Chatman, *Story and discourse: Narrative structure in fiction and film*, Cornell University Press, 1980.

[5] Eugenio Concepción, Pablo Gervás, and Gonzalo Méndez, ‘An api-based approach to co-creation in automatic storytelling’, in *6th International Workshop on Computational Creativity, Concept Invention, and General Intelligence. C3GI 2017*, (2017).

[6] Eugenio Concepción, Pablo Gervás, and Gonzalo Méndez, ‘A common model for representing stories in automatic storytelling’, in *6th International Workshop on Computational Creativity, Concept Invention, and General Intelligence. C3GI 2017*, (2017).

[7] Eugenio Concepción, Pablo Gervás, and Gonzalo Méndez, ‘A microservice-based architecture for story generation’, in *Microservices 2017*, (2017).

[8] Eugenio Concepción, Pablo Gervás, Gonzalo Méndez, and Carlos León, ‘Using cnl for knowledge elicitation and exchange across story generation systems’, in *International Workshop on Controlled Natural Language*, pp. 81–91. Springer, (2016).

[9] Eugenio Concepción, Gonzalo Mendez, and Pablo Gervás, ‘Mining knowledge in storytelling systems for narrative generation’, in *Proceedings of the INLG 2016 Workshop on Computational Creativity in Natural Language Generation*, pp. 41–50, (2016).

[10] Julio Cortázar, *Rayuela*, Editorial Sudamericana, Buenos Aires, 1963.

[11] Natlie Dehn, ‘Story generation after tale-spin.’, in *IJCAI*, volume 81, pp. 16–18, (1981).

[12] Thomas Erl, *Service-oriented architecture: a field guide to integrating XML and web services*, Prentice Hall PTR, 2004.

[13] Eric Evans, *Domain-driven design: tackling complexity in the heart of software*, Addison-Wesley Professional, 2004.

[14] Roy Thomas Fielding, *Architectural styles and the design of network-based software architectures*, Ph.D. dissertation, University of California, Irvine, 2000.

[15] Martin Fowler, *Patterns of enterprise application architecture*, Addison-Wesley Longman Publishing Co., Inc., 2002.

[16] Norbert E Fuchs, ‘Reasoning in attempto controlled english: non-monotonicity’, in *International Workshop on Controlled Natural Language*, pp. 13–24. Springer, (2016).

[17] Norbert E Fuchs, Kaarel Kaljurand, and Tobias Kuhn, ‘Attempto controlled english for knowledge representation’, in *Reasoning Web*, 104–124, Springer, (2008).

[18] Norbert E Fuchs, Kaarel Kaljurand, and Gerold Schneider, ‘Attempto controlled english meets the challenges of knowledge representation, reasoning, interoperability and user interfaces.’, in *FLAIRS Conference*, volume 12, pp. 664–669, (2006).

[19] Israel Gat and G Succi, ‘A survey of the api economy’, *Cut. Consort*, (2013).

[20] P. Gervás, ‘Story generator algorithms’, in *The Living Handbook of Narratology*, Hamburg University Press, (2012).

[21] Pablo Gervás, ‘Propp’s morphology of the folk tale as a grammar for generation’, in *OASIcs-OpenAccess Series in Informatics*, volume 32. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, (2013).

[22] Pablo Gervás, Belén Díaz-Agudo, Federico Peinado, and Raquel Hervás, ‘Story plot generation based on cbr’, *Knowledge-Based Systems*, **18**(4), 235–242, (2005).

[23] Joseph Goguen and D Fox Harrell, ‘Style as a choice of blending principles’, *Style and Meaning in Language, Art Music and Design*, 49–56, (2004).

[24] Jing Han, E Haihong, Guan Le, and Jian Du, ‘Survey on nosql database’, in *Pervasive computing and applications (ICPCA), 2011 6th international conference on*, pp. 363–366. IEEE, (2011).

[25] D Fox Harrell, 'Walking blues changes undersea: Imaginative narrative in interactive poetry generation with the griot system', in *AAAI 2006 Workshop in Computational Aesthetics: Artificial Intelligence Approaches to Happiness and Beauty*, pp. 61–69, (2006).

[26] Barbara Hayes-Roth, *The blackboard architecture: A general framework for problem solving?*, Heuristic Programming Project, Computer Science Department, Stanford University, 1983.

[27] Sheldon Klein, 'Automatic novel writer: A status report', *Papers in text analysis and text description*, (1973).

[28] Tobias Kuhn, *Controlled English for knowledge representation*, Ph.D. dissertation, Faculty of Economics, Business Administration and Information Technology of the University of Zurich, 2009.

[29] Michael Lebowitz, 'Creating characters in a story-telling universe', *Poetics*, **13**(3), 171–194, (1984).

[30] Michael Lebowitz, 'Storytelling and generalization', in *Seventh Annual Conference of the Cognitive Science Society*, pp. 100–109, (1985).

[31] Carlos León and Pablo Gervás, 'Creativity in story generation from the ground up: Nondeterministic simulation driven by narrative', in *5th International Conference on Computational Creativity, ICCC*, (2014).

[32] Uri Margolin, Peter Hühn, Jan Christoph Meister, John Pier, and Wolf Schmid, 'The living handbook of narratology', (2013).

[33] James R. Meehan, 'Tale-spin, an interactive program that writes stories', in *In Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, pp. 91–98, (1977).

[34] Gonzalo Méndez, Pablo Gervás, and Carlos León, 'On the use of character affinities for story plot generation', in *Knowledge, Information and Creativity Support Systems*, 211–225, Springer, (2016).

[35] Nick Montfort, 'Curveship's automatic narrative style', in *Proceedings of the 6th International Conference on Foundations of Digital Games*, pp. 211–218. ACM, (2011).

[36] Nick Montfort, Rafael Pérez, D Fox Harrell, and Andrew Campana, 'Slant: A blackboard system to generate plot, figuration, and narrative discourse aspects of stories', in *Proceedings of the fourth international conference on computational creativity*, pp. 168–175, (2013).

[37] Sam Newman, *Building microservices: designing fine-grained systems*, " O'Reilly Media, Inc.", 2015.

[38] Mike P Papazoglou, 'Service-oriented computing: Concepts, characteristics and directions', in *Web Information Systems Engineering, 2003. WISE 2003. Proceedings of the Fourth International Conference on*, pp. 3–12. IEEE, (2003).

[39] R. Perez y Perez, *MEXICA: A Computer Model of Creativity in Writing*, Ph.D. dissertation, The University of Sussex, 1999.

[40] Mark O Riedl and Robert Michael Young, 'Narrative planning: balancing plot and character', *Journal of Artificial Intelligence Research*, **39**(1), 217–268, (2010).

[41] Rolf Schwitter, 'Controlled natural languages for knowledge representation', in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pp. 1113–1121, Stroudsburg, PA, USA, (2010). Association for Computational Linguistics.

[42] Mei Si, Stacy C Marsella, and David V Pynadath, 'Thespian: Modeling socially normative behavior in a decision-theoretic framework', in *Intelligent Virtual Agents*, pp. 369–382. Springer, (2006).

[43] Scott R. Turner, *Minstrel: A Computer Model of Creativity and Storytelling*, Ph.D. dissertation, University of California at Los Angeles, Los Angeles, CA, USA, 1993. UMI Order no. GAX93-19933.

[44] Tony Veale, 'Creativity as a web service: A vision of human and computer creativity in the web era.', in *AAAI Spring Symposium: Creativity and (Early) Cognitive Development*, (2013).

[45] Eberhard Wolff, *Microservices: Flexible Software Architecture*, Addison-Wesley Professional, 2016.

Outcome Inference based on Threat Resources in Suspenseful Scenes

Pablo Delatorre [1] and **Carlos León** [2] and **Alberto Salguero** [3] and **Manuel Palomo-Duarte** [4] and **Pablo Gervás** [5]

Abstract. Suspense is a complex phenomenon and a key narrative issue in terms of emotional gratification. However, despite its complexity, existing automatic storytelling systems based on suspense commonly implement it by restricting the success options of the main character. In order to provide coverage to other components of suspense, in this paper we focus on elements that potentially influence the anticipation of the protagonist's final state. In particular, we present a study of how the threat's resources impact the foreseen outcome of the scene. To achieve this, we collected a list of threat resources and possible outcomes from a set of suspense films. Then suspense evoked by each these resources and outcomes were gathered. These data were analysed and classified to obtain the most common resource-anticipated outcome pairs. An automatic story generation system was adapted to generate plots including these pairs but not show the outcome to the audience. Results evidence that it is possible to omit the outcome in suspenseful automated storytelling if the threat resource is given.

1 INTRODUCTION

Readers of narrative try to enjoy through experimenting different real emotions, varying according to the content and the form [54, p. 12]. Suspense is a common, essential emotion that affects such narrative response. People feel suspense in a Stephen King's story, in the play of *The Woman in Black*, in a Hitchcock's movie and in the video game *Silent Hill*, but suspense can also be found in the Munch works or in a TV advertisement [46].

Together with coherence and thematic complexity, suspense explains 54% of the variance in interest of a narrative, making the single greatest contribution explaining roughly 34% [61, p. 436, 444]. Results support the assumption that suspense is a driver of video game enjoyment too. It predicts that suspense arises from media users' strong, emotion-based preference for how a given situation –e.g., in a thriller or horror movie– should be resolved [39, p. 29].

In the field of computational creativity, the importance of suspense is taken into account in a number of automatic storytellings. However, the review of these systems reveals some limitations in the way they generate suspense. Most of these systems evaluate and implement suspense through a functional simplification, as increasing or decreasing emotional links between characters [55, p. 4] or, more commonly, removing potential paths of success for the protagonists [14, p. 44] [65, p. 767]. While suspense is a complex concept, these strategies are based on a reductionist perspective. Thus, essential features extracted from the different conceptions of the term –as outcome importance, proximity or empathy– are usually addressed by automatic storytelling systems.

Against this background, we support that managing cognitive aspects of suspense helps to build robust suspenseful story generation systems. With this objective in mind, an architecture that tries to address the cognitive aspects of suspense as a whole has been previously proposed [24], providing evidence that general affective responses of the audience to the elements in the scene influence suspense [25]. Furthermore, the influence of affective elements in the generation of suspenseful scenes has been studied and implemented in the automatic storytelling system Stella [42], with positive –but not definitive– results [26].

However, so far only affective terms have been included in a story generation system as part of the overall goal of implementing storytelling system providing coverage to a wider set of components of suspense. This is clearly insufficient for achieving human-level suspense generation. Among others, the literature suggests that suspense involves outcome transcendence [6], outcome valence [57], uncertainty [2] or characters' morality [12].

In this paper we focus on the analysis of the effect of outcome anticipation in the audience. Anticipation is considered essential in the generation of suspense [34]. The inclusion of specific elements in the story may lead to the readers to predict possible outcomes. We hypothesize that while the perception about outcome transcendence directly impacts suspense, elements that lead to expect this outcome influence suspense in a comparable way.

For example, the knife and the mask of Michael Mayers in Carpenter's *Halloween* may evidence the intention of murdering even before his first killing, and presumably in a rougher way than gunfire does. Including a shark in a scene may led the audience to predict a mortal bite, which will be influenced by the size and aspect of the animal. In the context of suspense, ropes may lead to forecast abduction; in horror films, a chainsaw may imply a dismemberment; a disfigured threat usually express madness and painful outcome; likewise, a scalpel may induce expectation of torture, which is an intensely feared outcome [33, p. 24]. In this way, using preemptive resources to managing expectation plays an essential role when triggering suspense by anticipation.

This paper describes an effort to analyse and qualify this relation between specific elements in the scene and expected outcome. We specifically focus on threat resources as one of the most influential items in suspenseful stories.

[1] Universidad de Cádiz, Spain, email: pablo.delatorre@uca.es
[2] ITC, CS Faculty, U. Complutense de Madrid, Spain, email: cleon@ucm.es
[3] Universidad de Cádiz, Spain, email: alberto.salguero@uca.es
[4] Universidad de Cádiz, Spain, email: manuel.palomo@uca.es
[5] ITC, CS Faculty, U. Complutense de Madrid, Spain, email: pgervas@ucm.es

The study is based on the next two hypotheses:

1. *Threat resources influence the expected outcome.*
2. *There is a direct relation between suspense evoked by the expected outcome and suspense evoked by the corresponding threat resources.*

In order to verify these hypotheses, several steps were taken. First, we used popular suspense and horror films to collect different types of threat resources –which the audience can use to infer the outcome– and outcomes. In a second stage, a different set of participants was asked to relate outcomes and threat resources. Third, we used a classifier to obtain the most common resources for each threat, and we statistically analysed the potential relations. Afterwards, we implemented a simple model of *outcome-by-resource*. Finally, we tested the model by providing a small suspense scene to a different set of participants.

The rest of the paper is organised as follows: Section 2 describes the related literature on the elements of the scene and suspense. Section 3 describes the experiment, whose results are detailed in Section 4. Later, we briefly describe the implementation of the model in Section 5, and the results of testing it in Section 6. Finally, Section 7 and Section 8, respectively, discuss and summarize these results.

2 RELATED WORK

Hitchcock talked about suspense as the dramatisation of the narrative material in films, as much as the more intense representation of a dramatic situation [66, p. 11]. According to this, film suspense can be described as an anticipatory emotion, initiated by an event which sets up anticipations about a forthcoming, harmful outcome event for one of the main characters [22, p. 325].

The effect of anticipation due to increasing psychological stress is well documented [50, p. 204], and is outlined by the threat. Lazarus Alfert (1964) define it as the anticipation of something harmful in the future [41]–. Anticipation is considered a critical variable in the production of psychological stress, where a discrimination or interpretation of events had to be made for the threat to be perceived [18, p. 50]. Therefore, suspense primarily arises through the anticipation of how the story will go on or by the hope for a happy ending [71, p. 11].

Among all "expectable" situations, several authors agree that the expected outcome itself is a key factor to evoke suspense. Carroll (1984) introduces suspense as an affective concomitant of an answering scene or event which has two opposed outcomes –morally correct but unlikely versus evil and likely– [11, p. 72]. De Wied (1992) views suspense as a high degree of certainty of a negative outcome [22, p. 325]. Caplin et al. (2001) relate it to the amount that is at stake on the outcome [7, p. 73]. Zillmann (1980) defends that an universal restriction of suspense is that implies the preoccupation with feared probable outcomes threating liked protagonists[6] [73, p. 135]. Without a transcendent outcome *impact* is not possible to feel suspense [62, p. 287]. In other words, from the point of view of the audience, outcome must be significant to the character [6, p. 115]. Suspense is correlated with the audience's ability to generate a plan for the protagonist to avoid an impending negative outcome [53, p. 444].

Therefore, audience makes an own interpretation of the situation, where the series of events can be viewed not as merely accidental outcomes but natural consequences from a more detached perspective, taking into account the possible message in the story [36, p. 193, 202]. As an *anticipatory feeling* [48, p. 54] or *anticipation of misfortune* [63, p. 1], the discourse structure must present a situation leading to a significant result, since the reading of this event will make the reader concerned about the outcome of the event [36, p. 25].

In this regard, suspense may be related to a "perceived risk of victimization" [3, p. 54] in which information has not to be clear to predict the outcome at all. The strategy of hiding information leads the audience to fill the misrepresentation gap to *construct* an expected outcome [56, p. 102] or how to achieve it [52, p. 36]. An understanding of generic patterns of films, combined with the limited evidence offered within the narrative, are enough to have the audience "know" beyond the mere information supplied [19, p. 61]. According to Beecher (2007), stories that incite projections on behalf of characters, in relation to structures or to information gaps, generate suspense, which is presented as the emotional quotient of future prospects calibrated against the current and evolving *status quo* [4, p. 265]. In this way, features as mental diseases [15, p. 5], victims' attributes [16, p. 24] or the kind of weapon [70, p. 571] may induce to fear a certain type of intentionality on the part of the threat. Likewise, Deitz et al. (1984) suggest that subject characteristics, as well as those of the victim and defendant, may be predictive of the outcome [23, p. 277].

Being mainly a fear related to physical pain and/or psychological distress, the emotion evoked by potential outcomes involves characters' expectations about a future, undesirable, event [59, p. 302]. The possibility of the loss of self in a diabolic possession is perceived with terror due to the fear of losing self-control and therefore hurting people and the implied self-degradation [10, p. 18-19]. Mutations and metamorphoses in horror films may be considered to represent the fear of the destruction of the human organic form to the point of unnatural evolutionary insignificance [20, p. 167]. For their part, in addition to the suffering of the attack itself, the effects of being bitten by a vampire, scratched by a werewolf or eaten by zombies not infrequently lead to the fear of "return" in an inhuman state that resembles the threat itself [70, p. 570][47, p. 26].

In their work about sex and violence in slasher films, Sapolsky et al. (2003) examine several acts of violence –including beating, kicking, choking, drowning, burning, electrocuting, poisoning, beheading, dismembering, bludgeoning, hanging, stabbing, and shooting– [60, p. 31], which respectively lead to different types of distress. Clover (1987) seems to unify different types of weapon associated to the killer –knife, sledge hammer, scalpel, gun, machete, hanger, knitting needle, chainsaw– even when it is not clear if the effect in suspense is the same [17, p. 80]. Another interesting approach to outcome effects is studied by Hron (2008), who focuses on torture as an intensely feared outcome, and the instruments of torture –machinery and its sounds– as example of predicting outcome by inference, magnifying the menace [33, p. 24]. Specific instances of instruments associated to suspense genre are: the pendulum in Roger Corman's *The Pit and the Pendulum*; the music/movie contraption in Kubrick's *Clockwork Orange*; the dentist's drill in Schlesinger's *Marathon Man*; the rat cage in Radford's *1984* [33, p. 24]; the chainsaw and the meat hook in Tobe Hooper's *The Texas Chainsaw Massacre* and sequels [31, p. 68]; a sharp end of a tripod in Michael Powell's *Peeping Tom* [27, p. 8]; or –again– a drill and a scalpel in Eli Roth's *Hostel* [44, p. 52], among others.

Potential outcomes derived from these resources –death, mutilation, torture, injury, social debasement– can be categorized as negative outcomes [73, p. 136]. Zillman (1991) defends that the common denominator is the likely suffering of the protagonist. Slightly over-

[6] The work of Niehaus et al. (2012) reveals evidences that the preferences of the audience influence directly in the impact of the story [49].

stated, it thrives on fear [74, p. 284]. This "fear of victimization" has been analyzed in different fields. In particular, criminology and social behaviours study the concept "fear of crime", which may be considered as a general predictor of the victimization. According to Custer Van den Bulck (2017), criminological and psychological research has shown that perceived personal risk of criminal victimization, perceived ability to control crime, and perceived seriousness of crime are important predictors of fear of crime [21, p. 97]. On this matter, a number of studies have compared different hazardous situations in terms of apprehension. In addition, effects of victimisation have been broadly studied in *slasher* sub-genre[7] [60, 30, 70, 43, 9, 69]. According to Oliver Sanders (2004), this is a domain of special interest due to operationalizations of enjoyment of frightening films that focus on aggression and victimization may be best understood as applying to horror films: they generally succeed in increasing arousal or tension by threatening or actually showing graphic, horrifying, violent victimization [51, p. 245, 256]. The situativity of the filmic scene is closely connected with the fact that people deal with a danger. This requires one to instrumentalize the objects of the scenario for the story [72, p. 13].

Despite of the extensive amount of literature about victimization, there is a lack of analysis about measuring the emotional impact of involved resources in a scene. The explicit distinction among the effects of these resources is barely distinguished. This distinction includes not having resources; using a knife, club or gun [68, p. 27]; or mentioning but not using these resources beyond their emotional effects [37, p. 1252].

In the same way as the literature lacks this analysis of the impact of threat resources on the overall perceived suspense, to our best knowledge there is no automatic storytelling system addressing the impact of the resources [55, 67, 13, 52, 64, 58, 8, 1], with the exception of a general "resource of escape" [25, p. 308, 309].

3 EXPERIMENT

The experiment was conducted from March 2017 to June 2017. The first task consisted on gathering a list of threat resources and common outcomes from thriller and horror films. Next, a different group linked resources and outcomes and a third group evaluated the emotional affection evoked by these features.

3.1 Participants

The experiment was announced and those wanting to take part in it voluntary enrolled, counting finally one hundred and seven undergraduate students ($N = 107$), 58 males (54.63%) and 49 females (45.37%), from the University of (hidden for review), with ages ranging from 17 and 33 years ($mean = 20.14$, $stdev = 3.19$). All participants were Spanish native speakers. There was no compensation for participating in the evaluation.

An internal code –from 001 to 107– was assigned to each participant, relating this code with age, genre and contact method. By this way, participants were anonymously distributed in a way that limited the variability of number of participants, age and genre among the different stages of the experiment. The participants were distributed among the groups, by balancing their genre and age. Table 1 shows the distribution of the participants.

stage	**males**	**females**	$\overline{age}$	SD_{age}
1. films' analysis	14 (70.00%)	6 (30.00%)	20.05	3.02
2. features' links	15 (51.72%)	14 (48.28%)	20.32	3.82
3. suspense evaluation	15 (50.00%)	15 (50.00%)	21.03	3.99
4. testing the model	14 (50.00%)	14 (50.00%)	20.47	3.22

Table 1. Participants' distribution among the stages of the experiment

3.2 Stage 1: Analysis of Films

In this stage we collected a subset of the best suspense films. The selection was based on four on-line movie magazines: IMDb [35], MovieLens [32], Rotten Tomatoes$^{\text{TM}}$ [28], and Filmaffinity [29]. All four provide a clear genre division and an active community, with a high number of evaluations.

For each magazine, films tagged as suspense, terror, thriller horror or crime were gathered –depending on the tag name given by the magazine–. A total of five[8] lists of 150 films each where compiled, and the results were ordered by score. We discarded any movie not tagged as suspense, terror, thriller, horror or crime by at least three out of the four magazines. We finally obtained a list of 93 thriller and horror films, ranging from year 1931 (*M*) to year 2015 (*Bridge of Spies*).

Once films were gathered, each movie was randomly assigned to the twenty participants of this stage –see Table 1–. Each film was assigned to two different subjects for a total of nine or ten movies per participant. Participants had to identify suspense scenes, outcomes, and potential threat resources. Later, reported terms were checked to generate and to be assigned into a list of common words based on literature about fear of crime and victimization. Based on these results, classification was obtained as following:

- **Resources to Damage**: blunt weapon, physical power, bomb, claws, club, crusher, cudgel, dog, fire, force of nature, glass, hammer, immobilizer, knife, otherwordly, pencil, pistol, razor, rock, rope, scalpel, sharp weapon, shotgun, smasher, outer space, sword, teeth, venom, and water.
- **Outcomes**: death, torture, physical damage, sexual assault, returning –as ghost, living-dead...–, madness, loss of a limb, confinement for an indefinite period, loss of a loved one, and material losses.

3.3 Stage 2: Connecting Resources and Outcomes

In the second stage, twenty nine participants ($N = 29$) –fourteen women and fifteen men (see Table 1)– were asked to connect resources to their corresponding expected outcomes. For this purpose, a two-dimensional table was supplied to each subject, where columns where identified as the outcomes and rows, threat resources.

For each cell, subjects pointed a number from 1 to 3 on response to the question: *Assuming a movie scene in which a character is under an imminent threat, report the outcomes that you expect –columns– for each threat resource –rows–, with: 1, barely expected; 2, possible; and, 3, highly expected.*

This stage was designed as a paper-and-pencil test, starting with demographic information –gender, age and career–. All rows and columns were shuffled for each table so that each participant would receive a different version.

[7] According to Keisner (2008), slasher films are "those films characterized by a psychotic human or superhuman (i.e., monster, alien, poltergeist) that kills or stalks a succession of people, usually teenagers, and predominantly females" [38, p. 411].

[8] Rotten Tomatoes$^{\text{TM}}$ differences audience score from critics' score, so both evaluations were collected separately.

3.4 Stage 3: Suspense Evaluation

The objective of this task was to obtain the suspense evoked by the collected elements. Thirty participants ($N = 30$), fifteen women and fifteen men –see Table 1–, were queried about the perceived suspense for each concept in a suspenseful context.

This stage was designed as a paper-and-pencil test, starting with demographic information –gender, age and career–.

Once the demographic and personal questions were gathered, we provided the subject with the definition of suspense suggested by de Wied et al. (1992): "a high degree of certainty of a negative outcome" [22, p. 325]. This was done in order to reduce the ambiguity of the concept of suspense. After this the set of terms was presented to the participants. The list was randomly shuffled to avoid a sequence effect.

- *In a movie scene in which a character is about to face an imminent threat, report how much suspense you would feel as spectator by each of the following threat resources.* The terms to be scored were previously collected as the list of resources *Resources to Damage*.
- *In a movie scene in which a character is about to face an imminent outcome, report how much suspense you would feel as spectator.* The terms to be scored were previously collected as the list of resources *Outcomes*.

A 9-point rating scale –from *no suspense* to *extremely suspenseful*– through a pictographic scale based on the SAM model was used [5].

Thirty different surveys with random order were prepared and handed over to the participants for evaluation. This stage was conducted to test the model. Its procedure and results are described in Section 6.

4 RESULTS

This section describes the results obtained from the experiment detailed in Section 3. For all measures, the criteria for statistical significance was set at $\alpha = 0.05$.

In stage 2 (Section 3.3), participants would connect threat resources to expected outcomes. From this data, we conducted a k-means clustering [45] to obtain the strongest resource-outcome links by selecting those from the highest cluster. By applying the elbow method [40, p. 92], the number of clusters was set to 3. All resources in the highest clusters present a high or medium-high reported relation. Selected resources are shown in Table 2.

Results evidence that there are threat's resources which more likely predict specific outcomes. This supports hypothesis 1.

The results of stage 3 –Section 3.4– yield reported suspense means of outcomes and threat resources are listed in Table 3. Most reported scores have a medium to high value of suspense (> 5.00).

Based on the results of stages 2 and 3 we conducted a correlation test between reported suspense for outcomes and for resources. We obtained $r = 0.522, p < 0.000$, which moderately supports hypothesis 2.

A more exhaustive analysis of the results was carried out. Analizing Table 2 and Table 3, it can be observed that high-suspense outcomes were seldom connected to low-suspense threat resources. On the contrary, the lower reported suspense for outcomes, the more variability in reported suspense for the connected resources. This effect suggests that the correlation may increase when it comes to high suspense outcomes –as torture, death or sexual assault–.

	tor	sex	dea	lov	ret	lim	con	mad	phd	mat
blu	●									
bom			●							
cla					●					
clu									●	●
cru						●	●	●		
cud									●	
dog						●	●		●	
fir	●		●	●						●
gla	●	●								
ham						●	●	●		
imm	●	●				●	●			
kni				●				●		●
nat				●						●
phy	●	●			●				●	
pis		●	●	●						●
oth			●	●	●					
out			●	●						
raz	●	●	●	●		●				
roc						●			●	
rop		●					●			●
sca	●		●	●				●		
sha	●	●	●					●		
sho			●	●						
sma						●			●	
tee					●	●		●		
ven			●					●		
wat	●			●	●		●			

Table 2. Stronger outcome-resource pairs after selection by k-means.

	reported suspense	
element	**mean**	**std**
outcomes		
torture (*tor*)	7.60	0.99
sexual assault (*sex*)	7.50	1.57
death (*dea*)	7.50	1.43
loss of a loved one (*lov*)	7.35	1.81
returning (*ret*)	6.15	1.93
loss of a limb (*lim*)	6.00	1.87
confinement (*con*)	5.80	2.46
madness (*mad*)	5.10	1.77
physical damage (*phd*)	4.50	1.64
material losses (*mat*)	1.80	1.15
threat resources		
bomb (*bom*)	7.70	1.69
pistol (*pis*)	7.70	1.49
shotgun (*sho*)	7.50	1.61
outer space (*out*)	7.35	1.69
scalpel (*sca*)	7.05	1.57
fire (*fir*)	7.00	1.17
otherwordly (*oth*)	7.00	1.31
sharp weapon (*sha*)	7.00	1.59
razor (*raz*)	7.00	1.52
glass (*gla*)	6.95	1.76
immobilizer (*imm*)	6.90	1.71
blunt weapon (*blu*)	6.85	1.27
physical power (*phy*)	6.70	1.13
venom (*ven*)	6.65	2.01
water (*wat*)	6.50	1.67
teeth (*tee*)	5.90	2.20
claws (*cla*)	5.80	1.82
force of nature (*nat*)	5.80	1.32
hammer (*ham*)	5.55	1.36
smasher (*sma*)	5.55	2.04
dog (*dog*)	5.50	2.06
rope (*rop*)	5.50	2.16
club (*clu*)	5.35	2.56
cudgel (*cud*)	4.30	2.03
sword (*swo*)	5.15	2.46
crusher (*cru*)	5.00	2.25
knife (*kni*)	4.90	2.25
rock (*roc*)	4.20	2.46
pencil (*pen*)	4.10	2.75

Table 3. Reported suspense for outcomes and threat resources (9-Likert).

In order to support this, the mean of each outcome and the mean of its respective connected threat resources in terms of reported suspense were compared. Results show that differences between suspense by outcome and suspense by linked threat resources trends to decrease as suspense by outcome does. Figure 1 illustrates the trend when this deviation is applied.

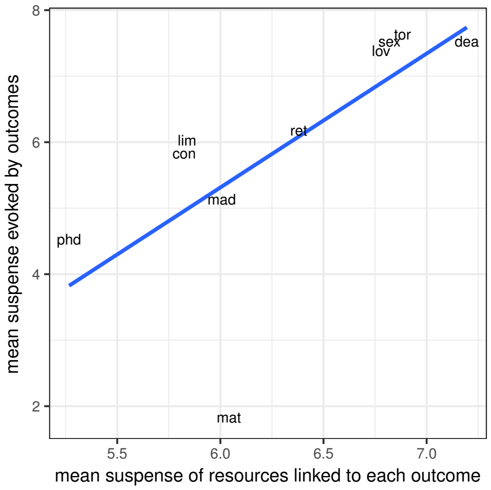


Figure 1. Suspense between each outcome and means of its linked elements.

The plot in Figure 1 evidences that "material loss" is an intrusive outlier due to its comparatively low suspense. Such a low reported suspense (1.8) seems to interfere in the correlation: the correlation analysis without "material loss" yields $r = 0.626, p < 0.000$.

Additionally, an ANOVA analysis shows a relation between weights and outcome suspense ($F_{1,288} = 4.144, p < 0.05$), which implies that discrepancies among participants about the probability of a resource being connected to an outcome depend on the type of outcome. However, it barely depends on its suspense ($r = 0.119, p < 0.05$).

Other aspects like the influence of participant gender and the context of the evaluation of affectivity were also analyzed, and no significant differences where found between reported suspense ($Z = 0.249, p = 0.478$) nor specified weights ($Z = 0.532, p = 0.322$).

5 APPLYING THE RESULTS TO A COMPUTATIONAL MODEL

The results described in Section 4 were applied to a prototype story generation system. We developed a generation model in which both the threat resource and the final outcome are internally computed, but the outcome is not rendered in the final text.

This model is implemented in Stellite, a stripped-down version of the storytelling generation system Stella, based on a hybrid model in which exhaustive, non-deterministic simulation is controlled by a narrative layer [42, 26].

Stella models stories as time-ordered sequences of states. Each state contains a detailed representation of each of the entities that populate it: physical information, emotions, intentions, knowledge about the world, and others. The simulation is carried out non-deterministically. On each generation step, the current state $s^{current}$ is expanded and all its potential next states $\{s_1^{next}, s_2^{next}, \ldots, s_n^{next}\}$ are generated. This means that, for each non-deterministic option for each next value of each attribute for each entity, a new path is created. This produces a vast generative space of stories.

In Stellite, the core generation engine and the knowledge model have been kept. The curves generation and matching engine have been removed, and the generation constraints and objectives have been made simpler. This reduces the chances to find a highly original story, and makes it impossible to generate at a very fine detail, but the generation is faster and the output stories are all coherent. Additionally, Stellite is enriched with the implementation of a computational model to compute the suspense of decorative elements [25].

In this version, Stellite generates short stories in which there is an outcome taken from the list described in Section 3.2 and the threat resources are selected among those which are strongly related to the outcome, according to the values show in Table 2.

6 TESTING OUTCOME INFERENCE THROUGH THE COMPUTATIONAL MODEL

This section describes the experiment that was carried out to verify that the impact of threat resource when inferring corresponding outcomes in suspenseful scenes is stable and that it can be formalized and included in an automatic story generation system.

6.1 Participants

The experiment took place in the College of (hidden for review) of the University of (hidden for review). As referred in Table 1, twenty eight undergraduate students ($N = 28$), 14 males (50.00%) and 14 females (50.00%) participated in this task, in the same conditions as the ones detailed in Section 3.1.

6.2 Generating a Story for Evaluation

In order to test the model, we analysed the response on resource-outcome variations over a fixed plot. To produce this plot, Stellite was parameterized to generate stories about one `character` trying to escape. This character encounters another character who threats him. The number of possible `outcomes` was set to 1 for each story. The threat must have any of the 29 possible `resources` at hand. The world was formed by a simple `map` including two instances of `location`: a corridor and a neutral room with an entrance door. Decorative elements were not included in scene.

By combining these environmental features and the different resources of the threat, Stellite generated 1523 scenes. In its current state, Stellite is not able to guarantee full coherence and the presence of suspense for all generated stories, so in order to sample a story for experimentation, we proceeded as follows: 1) a story was randomly selected among the set of generated stories and 2) five external evaluators secretly wrote down whether they perceived suspense and coherence. Unless there was full consensus and the story was coherent and suspenseful, 3) steps 1 and 2 had to be repeated. The process finished after 3 tries. While this process provided pseudo randomness in the process, it involved human criteria and makes the result not fully automated. However, the current technical limitations make it imposible to fully rely on the generator capabilities. This is discussed in Section 7.

The obtained plot was set as the source template. A new set of 28 stories for each of the 10 outcomes, a total of 280 stories, was created –one set containing 10 stories per participant–. For each story, Stellite automatically computed a resource to be assigned to the threat. The text was generated from the structured representation, being rendered in Spanish with simple text templates. An example of one final cut from the chosen template is shown next[9].

> *Julia ran away across a corridor, trying not to make any noise. She was looking for the way out. When she reached the end of the corridor, she turned right. She went into a room. Her chaser was waiting for her there. Next to her chaser there were a table and a knife.*

In some cases, the nature of the resource forced a manual adaptation to the rendered text. For instance, changing *water* by *bucket of water*. This issue and its relative impact is discussed in Section 7.

6.3 Method

The experiment was run as a *paper-and-pencil* session in one single classroom. A single demographic survey was filled by each participant. The experiment was explained. The content was then presented, handing each evaluator over the 10 versions of the story, each on one single paper sheet. For each outcome, all possible resources from Table 2 had been computed by Stellite. The outcome was not explicitly included in any of the texts.

For each version, the test invited the participant to check one or several of the 10 possible outcomes in response to the question: *Among the possible outcomes, which ones are plausible according to this scene?*

6.4 Results

The analysis supports the results obtained in Section 4. 687 relations between resource and outcome were gathered, with 76 false positives –relations which were not found in the first experiment–. Additionally, 15 false negatives –expected from the first experiment, but not scored in current– were found. However, 10 (66%) of these false negatives involved the "loss of a loved one" outcome, which is –this outcome– clearly not consistent with the semantics of the sampled template –there is no other character in the plot–.

When compared against the previous experiment, a correlation test reveals a similar and moderate correlation ($r = 0.495$, $p < 0.000$) between suspense evoked by threat resources and suspense evoked by outcome.

We found out that the more suspenseful the outcome, the higher the expectation ($r = 0.745$, $p < 0.000$ for the most suspenseful outcomes). It seems that participants would tend to assume more dangerous/harmful endings.

Figure 2 depicts this effect of the higher correlation for the most suspenseful outcomes[10]. An ANOVA analysis supports both findings –resources and outcomes ($F_{1,134} = 17.168$, $p < 0.000$), and occurrences and outcomes ($F_{1,134} = 23.510$, $p < 0.000$), although it does not show a significant impact of both factors together ($F_{3,134} = 1.038$, $p = 0.310$).

7 DISCUSSION

While the results have shed some light on the possibility of using inference for improving the generation of suspenseful scenes, there are a few pending issues that are worth discussing.

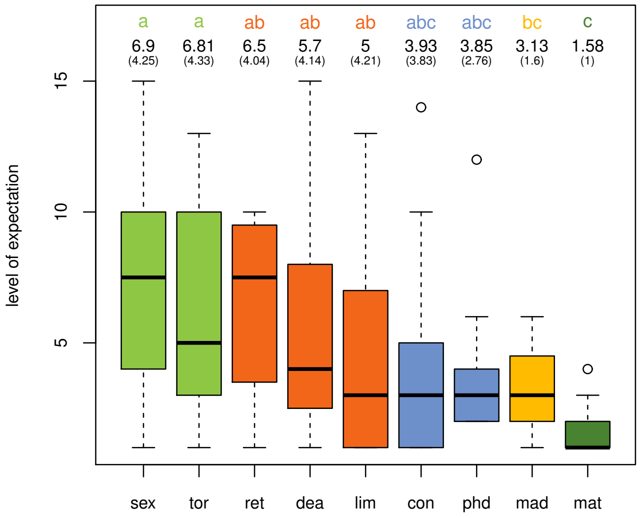


Figure 2. Level of expectation for each outcome.

In order to find out which resources are the most strongly connected to the specific outcomes, we have used a k-means analysis instead of a central tendency measure. This is because this classifier discriminates better when the variance is low. In any case, elements with minimum impact were not included because every used outcome had at least one strongly connected threat resource.

The plot template sampling process is still very dependent on manual intervention. While the story generation system works and is able to produce valid content in most cases, checking actual coherence and suspense is beyond the current state of the art. Among other things, it would require complete automatic understanding and contextualization of the underlying semantics and a model of suspense itself, which is actually what we are trying to achieve. The authors are well aware that full automation must replace this manual influence by proper computational means.

The probability of expecting some outcomes is heavily dependent on the plot elements. For example and as described, losing a loved one was not marked by any participant. This is, again, a limitation of the storytelling system and the chosen depth of the template plot. We expect further versions of both the model and the experimentation to be able to provide more general insight by producing complex stories in which the semantics provide coverage to a wider pairs of threat resource-outcome.

8 CONCLUSIONS AND FUTURE WORK

The paper has reported on an effort to provide evidence on the hypothesis that outcomes can be partially inferred by the threat resources in suspenseful scenes. An initial experimental study provided both a list of classic outcomes and threat resources and their most likely connections. A second experiment in which this information was used to automatically generate stories that show the resource but not the outcome was run. The experiments showed that subjects are certainly able to provide inferences relatively consistent with the initial results.

We therefore conclude that there is evidence of the existence of an influence of the resources on the final outcome, and that this can be applied in computational storytelling for creating suspenseful scenes.

While there is evidence of the resource-outcome dependency, we found out that highly emotional outcomes are inferred when highly

[9] In the original, in Spanish: "*Julia huía por un pasillo tratando de no hacer ruido. Buscaba la salida de la casa. Al llegar al final del pasillo, giró a la derecha. Entró en una habitación. Su perseguidor le esperaba allí. Al lado de su perseguidor había una mesa y un cuchillo.*"

[10] Suspense may be reviewed in Table 3.

emotional resources are present in the scene. However, this relation is not still clear in the case of outcomes with medium-low emotional values in terms of suspense. In any case, all these relations seem clearly dependent on the semantics, which makes it difficult to choose the best resource in some cases. These issues have produced over a 10% of false positives. How to reduce this effect will be studied in further experiments.

ACKNOWLEDGEMENTS

This work has been supported by the Andalusian Government under the University of Cadiz programme for Researching and Innovation in Education 2015/2016 (SOL-201500054211-TRA); by the IDiLyCo project (TIN2015-66655-R) funded by the Spanish Ministry of Economy, Industry and Competitiveness.

The research presented in this paper has been carried out as part of the first author's work in the PhD Programme in Computer Science of Universidad Complutense de Madrid (RD 99/2011) and is part of his PhD dissertation.

REFERENCES

[1] Byung-Chull Bae and R Michael Young, 'A use of flashback and foreshadowing for surprise arousal in narrative using a plan-based approach', in *Interactive Storytelling*, 156–167, Springer, (2008).

[2] Yoav Bar-Anan, Timothy D Wilson, and Daniel T Gilbert, 'The feeling of uncertainty intensifies affective reactions.', *Emotion*, **9**(1), 123, (2009).

[3] Victoria Simpson Beck and Lawrence F Travis, 'Sex offender notification: An exploratory assessment of state variation in notification processes', *Journal of Criminal Justice*, **34**(1), 51–55, (2006).

[4] Donald Beecher, 'Suspense', *Philosophy and Literature*, **31**(2), 255–279, (2007).

[5] Margaret M Bradley and Peter J Lang, 'Measuring emotion: the self-assessment manikin and the semantic differential', *Journal of behavior therapy and experimental psychiatry*, **25**(1), 49–59, (1994).

[6] William F Brewer, 'The nature of narrative suspense and the problem of rereading', *Suspense: Conceptualizations, theoretical analyses, and empirical explorations*, 107–127, (1996).

[7] Andrew Caplin and John Leahy, 'Psychological expected utility theory and anticipatory feelings', *Quarterly Journal of economics*, 55–79, (2001).

[8] Rogelio E Cardona-Rivera, Bradley A Cassell, Stephen G Ware, and R Michael Young, 'Indexter: A computational model of the event-indexing situation model for characterizing narratives', in *The Workshop on Computational Models of Narrative at the Language Resources and Evaluation Conference*, pp. 32–41, (2012).

[9] Enrique Carrasco-Molina, 'Apuntes sobre la percepción subconsciente en el cine. el ejemplo de Alien, el octavo pasajero (1979) y su propuesta orgánica de atracción/repulsión.', *Revista Mediterránea de Comunicación*, **3**(2), 46–82, (2012).

[10] Noel Carroll, 'Nightmare and the horror film: The symbolic biology of fantastic beings', *Film Quarterly*, **34**(3), 16–25, (1981).

[11] Noël Carroll, 'Toward a theory of film suspense', *Persistence of Vision*, **1**(1), 65–89, (1984).

[12] Noël Carroll et al., *The philosophy of horror: Or, paradoxes of the heart*, Routledge (New York), 1990.

[13] Yun-Gyung Cheong and R Michael Young, 'A computational model of narrative generation for suspense', in *Association for the Advancement of Artificial Intelligence (AAAI Journal)*, pp. 1906–1907, (2006).

[14] Yun-Gyung Cheong and R Michael Young, 'Suspenser: A story generation system for suspense', *IEEE Transactions on Computational Intelligence and AI in Games*, **7**(1), 39–52, (2015).

[15] Kristen Elizabeth Chmielewski, *Silver screen slashers and psychopaths: a content analysis of schizophrenia in recent film*, Master of arts, University of Iowa, 2013.

[16] Kyle Christensen, 'The Final Girl versus Wes Craven's A Nightmare on Elm Street: Proposing a Stronger Model of Feminism in Slasher Horror Cinema', in *Studies in Popular Culture*, ed., Ronda V (Gordon College) Wilcox, volume 34, 23–48, Popular Culture Association in the South, (2010).

[17] Carol J Clover, 'Gender in the slasher film', in *Her body, himself*, chapter 6, 91–133, (1987).

[18] Paul Comisky and Jennings Bryant, 'Factors involved in generating suspense', *Human Communication Research*, **9**(1), 49–58, (1982).

[19] David R Coon, 'Building suspense: Spaces, boundaries, and drama in hitchchock's rear window and psycho', *Polymath: An Interdisciplinary Arts and Sciences Journal*, **2**(3), (2012).

[20] Ronald Allan Lopez Cruz, 'Mutations and metamorphoses: Body horror is biological horror', *Journal of Popular Film and Television*, **40**(4), 160–168, (2012).

[21] Kathleen Custers and Jan Van den Bulck, 'The Association between Soap Opera and Music Video Viewing and Fear of Crime in Adolescents: Exploring a Mediated Fear Model', *Communication Research*, **44**(1), 96–116, (2017).

[22] Minet de Wied, Ed SH Tan, and Nico Henry Frijda, 'Duration experience under conditions of suspense in films', *NATO ASI series. Time, action and cognition: Towards bridging the gap*, 325–336, (1992).

[23] Sheila R Deitz, Madeleine Littman, and Brenda J Bentley, 'Attribution of responsibility for rape: The influence of observer empathy, victim resistance, and victim attractiveness', *Sex Roles*, **10**(3-4), 261–280, (1984).

[24] Pablo Delatorre, Barbara Arfè, Pablo Gervás, and Manuel Palomo-Duarte, 'A component-based architecture for suspense modelling', in *Proceedings of AISB 2016's Third International Symposium on Computational Creativity (CC2016)*, pp. 32–39, (2016). http://hdl.handle.net/10498/18328.

[25] Pablo Delatorre, Carlos León, Pablo Gervás, and Manuel Palomo-Duarte, 'A computational model of the cognitive impact of decorative elements on the perception of suspense', *Connection Science*, **29**(4), 295–331, (2017).

[26] Pablo Delatorre, Carlos León, Manuel Palomo-Duarte, and Pablo Gervás, 'Adding suspense to a story generation system through a cognitive model of the impact of affective terms', in *Proceedings of 6th International Workshop on Computational Creativity, Concept Invention, and General Intelligence (C3GI)*, Madrid (Spain), (2017). (In Press).

[27] Shyla Fairfax, *Women in Slashers Then and Now: Survival, Trauma, and the Diminishing Power of the Close-Up*, Master of film studies, Carleton University, 2014.

[28] Fandango. Rotten Tomatoes$^{\mathrm{TM}}$. `http://www.rottentomatoes.com/`, 2016. Accessed on 2017-02-14.

[29] Filmaffinity S.L. Filmffinity. `http://www.filmaffinity.com`, 2017. Accessed on 2017-02-14.

[30] Cynthia A Freeland, 'Feminist frameworks for horror films', *Film theory & criticism*, 627–648, (1996).

[31] Craig Frost, 'Erasing the b out of bad cinema: Remaking identity in the texas chainsaw massacre', *COLLOQUY text theory critique*, **18**, (2009).

[32] GroupLens Research. MovieLens. `https://movielens.org/`, 2017. Accessed on 2017-02-15.

[33] Madelaine Hron, 'Torture goes pop!', *Peace Review*, **20**(1), 22–30, (2008).

[34] Mitchell Alexander Ian, *Reading again for the first time: Rereading for closure in Interactive Stories*, Msc, National University of Singapore, 2012.

[35] IMDb.com, Inc. IMDb. `http://www.imdb.com/`, 2017. Accessed on 2017-02-15.

[36] Yumiko Iwata, *Creating Suspense and Surprise in Short Literary Fiction: A stylistic and narratological approach*, Ph.D. dissertation, University of Birmingham, 2009.

[37] Hyunseok Jang, Ji Hyon Kang, Rick Dierenfeldt, and Greg Lindsteadt, 'Weapon possession among college students: a study from a midwestern university', *International journal of offender therapy and comparative criminology*, **59**(11), 1239–1259, (2015).

[38] Jody Keisner, 'Do you want to watch? a study of the visual rhetoric of the postmodern horror film', *Women's Studies*, **37**(4), 411–427, (2008).

[39] Christoph Klimmt, Albert Rizzo, Peter Vorderer, Jan Koch, and Till Fischer, 'Experimental evidence for suspense as determinant of video game enjoyment', *CyberPsychology & Behavior*, **12**(1), 29–31, (2009).

[40] Trupti M Kodinariya and Prashant R Makwana, 'Review on determining number of cluster in k-means clustering', *International Journal*, **1**(6), 90–95, (2013).

[41] Richard S Lazarus and Elizabeth Alfert, 'Short-circuiting of threat by experimentally altering cognitive appraisal.', *The Journal of Abnormal*

and Social Psychology, **69**(2), 195–205, (1964).

[42] Carlos León and Pablo Gervás, 'Creativity in story generation from the ground up: Nondeterministic simulation driven by narrative', in *5th International Conference on Computational Creativity, ICCC*, (2014).

[43] Daniel Linz and Edward Donnerstein, 'Sex and violence in slasher films: a reinterpretation', *Journal of Broadcasting and Electronic Media*, 243–246, (1994).

[44] Adam Lowenstein, 'Spectacle horror and hostel: why 'torture porn'does not exist', *Critical Quarterly*, **53**(1), 42–60, (2011).

[45] James MacQueen et al., 'Some methods for classification and analysis of multivariate observations', in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pp. 281–297. Oakland, CA, USA., (1967).

[46] Robert Madrigal and Colleen Bee, 'Suspense as an experience of mixed emotions: Feelings of hope and fear while watching suspenseful commercials.', *Advances in consumer research*, **32**(1), (2005).

[47] James McFarland, 'Philosophy of the living dead: At the origin of the zombie-image', *Cultural Critique*, **90**(1), 22–63, (2015).

[48] David S Miall, *Literary reading: empirical & theoretical studies*, Peter Lang, 2006.

[49] James Niehaus, Victoria Romero, Jonathan Pfautz, Scott Neal Reilly, Richard Gerrig, and Peter Wayhrauch, 'Towards a computational model of narrative persuasion: a broad perspective', p. 182, (2012).

[50] Markellos S Nomikos, Edward Opton Jr, and James R Averill, 'Surprise versus suspense in the production of stress reaction.', *Journal of Personality and Social Psychology*, **8**(2), 204–208, (1968).

[51] Mary Beth Oliver and Meghan Sanders, 'The appeal of horror and suspense', *The horror film*, 242–260, (2004).

[52] Brian O'Neill, *A computational model of suspense for the augmentation of intelligent story generation*, Ph.D. dissertation, Georgia Institute of Technology, 2013.

[53] Brian O'Neill and Mark Riedl, 'Dramatis: A computational model of suspense.', in *Proceedings of the 28th Conference on Artificial Intelligence (AAAI)*, volume 2, pp. 944–950, (2014).

[54] Federico Peinado, *Un armazón para el desarrollo de aplicaciones de narración automática basado en componentes ontológicos reutilizables*, Phd, University Complutense of Madrid, 2008.

[55] Rafael Pérez y Pérez, 'Employing emotions to drive plot generation in a computer-based storyteller', *Cognitive Systems Research*, **8**(2), 89–109, (2007).

[56] Juan A Prieto-Pablos, 'The paradox of suspense', *Poetics*, **26**(2), 99–113, (1998).

[57] Arthur A Raney, 'Moral judgment as a predictor of enjoyment of crime drama', *Media Psychology*, **4**(4), 305–322, (2002).

[58] Mark O Riedl and R Michael Young, 'Narrative planning: balancing plot and character', *Journal of Artificial Intelligence Research*, **39**(1), 217–268, (2010).

[59] Andrew Salway and Mike Graham, 'Extracting information about emotions in films', in *Proceedings of the eleventh ACM international conference on Multimedia*, pp. 299–302. ACM, (2003).

[60] Burry S Sapolsky, Fred Molitor, and Sarah Luque, 'Sex and violence in slasher films: Re-examining the assumptions', *Journalism & Mass Communication Quarterly*, **80**(1), 28–38, (2003).

[61] Gregory Schraw, Terri Flowerday, and Stephen Lehman, 'Increasing situational interest in the classroom', *Educational Psychology Review*, **13**(3), 211–224, (2001).

[62] Aaron Smuts, 'The desire-frustration theory of suspense', *The Journal of Aesthetics and Art Criticism*, **66**(3), 281–290, (2008).

[63] Sirish Kumar Somanchi, 'A computational model of suspense in virtual worlds', *Technical Report Number 03-002*, (2003).

[64] Nicolas Szilas, 'IDtension: a narrative engine for interactive drama', in *Proceedings of the Technologies for Interactive Digital Storytelling and Entertainment (TIDSE) Conference*, volume 3, pp. 187–203, (2003).

[65] Nicolas Szilas, 'A computational model of an intelligent narrator for interactive narratives', *Applied Artificial Intelligence*, **21**(8), 753–801, (2007).

[66] François Truffaut and Helen Scott, *El cine según Hitchcock*, Alianza Editorial (Madrid), 1974. Reedition of Le Cinéma selon Hitchcock, Robert Laffont Éditions (Paris, 1966).

[67] Scott R. Turner, *The Creative Process: A Computer Model of Storytelling and Creativity*, Taylor & Francis, 2014.

[68] Mark Warr, 'Fear of victimization', *The Public Perspective*, **November/December**, 25–28, (1993).

[69] James B Weaver III, 'Are "slasher" horror films sexually violent? A content analysis', *Journal of Broadcasting & Electronic Media*, **35**(3), 385–392, (1991).

[70] Linda Williams, 'When the woman looks', in *Film Genres*, 561–577, (1984).

[71] Werner Wirth and Holger Schramm, 'Media and emotions', *Communication research trends*, **24**(3), 2–44, (2005).

[72] Hans J Wulff, 'Suspense and the influence of cataphora on viewers' expectations', in *Suspense: Conceptualizations, theoretical analyses, and empirical explorations*, eds., Peter Vorderer, Hans J Wulff, and Mike Friedrichsen, 1–17, Mahwah, New Jersey: Lawrence Erlbaum Associates, (1996).

[73] Dolf Zillmann, 'Anatomy of suspense', in *The entertainment functions of television*, 133–161, Psycology Press, (1980).

[74] Dolf Zillmann, 'The logic of suspense and mystery', in *Responding to the screen. Reception and reaction processes*, 281–303, Lawrence Erlbaum Associates, (1991).

Automatic Detection of Narrative Structure for High-Level Story Representation

Maximilian Droog-Hayes and **Geraint Wiggins** and **Matthew Purver** [1]

Abstract. Automatic summarization is dominated by approaches which focus on the selection and concatenation of material in a text. What can be achieved by such approaches is intrinsically limited and far below what can be achieved by human summarizers. There is evidence that successfully creating a rich representation of text, including details of its narrative structure, would help to create more human-like summaries. This paper describes a part of our ongoing work on a cognitively inspired, creative approach to summarization. Here we detail our work on the detection of narrative structure in order to help build rich interpretations of a text and help give rise to a creative approach to summarization. In particular we consider the domain of Russian folktales. Using Vladimir Propp's thorough description of the interrelations between the narrative elements of such tales, we pose this task as a constraint satisfaction problem. While we only consider this small domain, our approach can be applied to any domain of text on which enough constraints can be placed.

1 Introduction

The field of automatic summarization has seen no major improvements in recent times. We suggest that this is partly due to disputed [30] but still widely used evaluation metrics such as ROUGE [24]. Although such metrics have been used to train summarization systems, they cannot distinguish between summaries of obviously differing qualities [28]. Another key factor is that no fundamentally different approaches to computational summarization have been investigated. While new techniques are being developed, they all essentially remain extractive; content is selected and concatenated to form a summary.

There is evidence that understanding the structure of a discourse and its logical organisation can aid the recognition of key material. As such, it would be be beneficial to the field of automatic summarization if the structure of a discourse could be modelled computationally. This would lead to the creation of richer semantic models of text and, we believe, aid progress toward more human-like capabilities of text summarization.

In this paper we present our approach to determining the narrative structure of a text, which fits into a larger body of work on automatic text summarization. We create high-level interpretations of Russian folktales as well as obtaining the roles of the key characters in a story. This process can produce many different, mutually exclusive representations for a single story, leading to creative interpretations of a text. We use the thorough analysis of the morphology of Russian folktales by Vladimir Propp [27] to build a system which creates a representation of the key narrative events in a tale at the conceptual level. We believe this abstraction away from the surface text of a tale will prove useful in the detection of key summary-worthy events and allow for more human-like summaries to be produced. Propp's morphology describes the narrative structure of a very specific domain of text, however our approach can be applied to any domain for which a detailed enough account of the narrative structure is available.

We treat the problem of annotating folktales in the style of Propp as one of constraint satisfaction. First, the roles of characters must be decided and potential instances of each narrative unit are determined. This gives us a conceptual space of sequences of Propp's character functions for a given tale. We explore this space to identify all valid sequences of character functions that conform to the constraints and interrelations between the elements described by Propp. These are not entirely restrictive, which allows for the production of multiple alternative sequences of character functions to represent a single tale, and creative interpretations to arise.

As this current work is highly entwined with a careful analysis of Propp's work, we first provide a description of Propp's morphology in enough detail so as to give context for our work and implementation decisions. This is followed by a discussion of work in summarization, how it has ties with computational creativity, and the work in narrative comprehension which provides motivation for our task. We then give a brief overview of existing research making use of Propp's Morphology of the Folktale, and subsequently discuss constraint logic programming before giving an explanation of our system and discuss our results and future work.

2 Propp's Morphology

In his book, *Morphology of the Folktale*, Vladimir Propp [27] analysed a subset of the corpus of Russian folk tales compiled by Afanasyev. Over a set of 100 tales, Propp identified five categories of elements which he claims define a tale as a whole.

1. Character Functions - a sequence of 31 character-based functions; the narrative units of a tale which are performed by the *dramatis personae*.
2. Conjunctive Elements - when successive functions are performed by different characters, the latter character must somehow be informed of everything that has occurred up until that point. This may for example occur when characters act *ex machina*, are all knowing, or overhear a dialogue between others.
3. Character Motivations - the goals and aims of characters, which drive their actions.
4. Character Appearance - the forms in which characters first enter the story, for example an accidental encounter, or sudden arrival.
5. Attributive Elements - the specific qualities belonging to each

[1] Queen Mary University of London, United Kingdom, email: {m.droog-hayes, geraint.wiggins, m.purver}@qmul.ac.uk

character, for example their age or peculiarities of their appearance.

The key category amongst these is the sequence of 31 character-based narrative units, or narratemes, that make up the actions of the story[2]. These each provide a generalized description of a key event in a tale. Propp repeatedly stresses the importance of these character functions over other elements of the tale such as the characters who perform them. Table 1 gives the canonical ordering of these functions, with their designation and brief definition.

Table 1. The strict ordering of Propp's character based functions.

β absentation
γ interdiction
δ violation
ϵ reconnaissance
ζ delivery
η trickery
θ complicity
A/a villainy/lack
B mediation
C beginning counteraction
$\uparrow$ departure
D donor tests hero
E hero reacts
F receipt of magical agent
G transference
H struggle
J branding
I victory
K liquidation
$\downarrow$ return
Pr pursuit
Rs rescue
O unrecognized arrival
L unfounded claims
M difficult task
N solution
Q recognition
Ex exposure
T transfiguration
U punishment
W reward

While there is a canonical order to these functions, any given function does not have to be present in a particular instance of a tale. In addition, many of these functions are paired, such as *struggle/victory* and *difficult task/solution*. Propp does however make one key restriction on what *must* be present within a tale; a tale necessarily has to include either an instance of *villainy* or *lack* which provides the motivation for the subsequent actions of the protagonist.

A single tale could hold multiple sequences of these character functions, either sequentially or embedded within one another. Propp defines a *move* as "any development from villainy (A) or a lack (a), through intermediary functions to marriage (W*), or to other functions employed as a denouement." With this definition, every new instance of villainy or lack creates a new move. In this paper, we only consider 'single-move tales' in order to avoid complicating the task and keep the sequential ordering of functions.

Propp describes these functions with a series of short examples, often providing the indicator words for the presence of a function. That is, most of Propp's functions are detected via long and highly varied lists of cue words. Though this is not the case for all of the character functions. For instance Propp's third character function, the violation of an interdiction, is highly dependent on the form that preceding interdiction function takes.

Each character function has multiple subtypes corresponding to the specific forms that it may take. To take the testing of the hero by a donor as an example (*donor tests hero*), Propp describes 10 fine-grained ways in which this function might occur. These cover forms such as: *the testing of the hero*, *requesting mercy* or *requesting the division of property*.

The final element of Propp's morphology requiring discussion here is that of the roles of the dramatis personae. Propp concluded that every character in a tale could be resolved into one of seven types according to their purpose. The Hero (who defeats the villain or resolves the lack), the Villain (the character who creates the main obstacle for the hero), the Donor (the character who may test the hero and gives them a magical agent), the Dispatcher (the character to send the hero on their quest), the Helper (an often magical agent who assists the hero), the Princess/Prize (the marriage to this character is often the goal of the hero), and the False Hero (who takes credit for the hero's actions and seeks the reward for themselves). Certain character functions are logically connected and grouped into *spheres of action*. Each of these spheres corresponds to one of the character types, and specifies the character functions that a character of a given type is involved in. However, one character in a tale may fulfil the role and the actions performed by several character types. For instance, a situation is described whereby the villain character unwillingly fulfils the role of the donor too, accidentally leaving a magical agent behind only to be found by the hero.

3 Related Work

3.1 Summarization

Research on automatic text summarization began with the work of Luhn [25]. This work involved automatically creating literature abstracts by sentence extraction based on scoring sentences according to the proximity of frequently occurring words. It is interesting to note that the author discusses machine summarization with the idea that it may eliminate human bias from the 'abstracter's product', where nearly sixty years on many summarization systems are trained on datasets of human summaries in order to generate summaries that follow a similar style.

Since the work of Luhn [25] nearly 60 years ago, much of the research into automatic summarization has remained *extractive*, selecting content in a document to paste together into a summary. This is in contrast to *abstractive* summarization which involves understanding and representing the contents of a document before generating a summary from this intermediate representation in a concise and original way. Abstractive summarization is desirable in order to reach more human levels of summarization performance. Extractive summarization has been called the low-hanging fruit of summarization, being technological rather than fundamental [31]. In addition it has been shown that not only does human sentence extraction perform poorly in comparison to regular abstractive summarization, but that automatic extractive systems from several years ago were already approaching the ceiling of what can be achieved by human extractors [14].

To reach more human-like levels of summarization it is necessary to have a richer computational representation of a text. One way to achieve this, which we attempt here, is to first examine the structure of a text, in order to subsequently aid the recognition of key material. This idea is not new. Over 30 years ago, Lehnert [23] stated the necessity of a high level analysis of a story in order to create a summary. To achieve this Lehnert proposed *plot units*. These units are comprised of affect states representing positive, neutral or negative events, or mental states. The aim was to use these graph-like conceptual structures to represent the plot of a story. Subsequent work by Goyal et al. [17, 18] attempted to model plot units without vast amounts of knowledge engineering, via a "variety of sentiment-related and general purpose language resources" [18, p. 2]. The modest results indicated the difficulty of the task, but also its feasibility.

[2] Henceforth we shall refer to this set of character-based narrative units as *character functions*.

3.2 Summarization and Creativity

We view summarization as a creative task, involving exploration of the conceptual space of texts that summarise a particular text. The notation for the elements of our conceptual space is a rich graphical meaning representation of a given text, inspired by research into the cognitive representations created by humans as they read. We direct the exploration of the space by various heuristics, to select the content which should be kept in order to form a summary *representation* and allow for subsequent summary generation. In a similar way, Gervás [16] describes his work on generating instances of Russian folktales based on the constraints given by Propp.

Unlike story generation, summarization has not traditionally been considered a computationally creative task. However, just as a human writer would draw on their knowledge of other stories, so do many story generation systems. Systems such as MEXICA [26] use a store of previous stories in long term memory in order to aid the generation of new stories. In this regard our approach is a special case of a story generation task; with an input story as background knowledge, we generate a new, necessarily shorter story, that aims to express the same key events as the input. In addition, TALE-SPIN [29] works with a list of pre-defined actions as prior knowledge, treating story generation as a problem solving task. This is comparable to our work of finding plausible interpretations of a story that fit a set of constraints.

3.3 Discourse Structure

Understanding the structure of discourse helps to explain how a cohesive text is formed. A cohesive text is one where the content of a text is semantically and logically connected, in order to convey meaning to the reader. Detection of a text's structure can potentially aid the recognition of key material in a text or indicate an area of importance.

Grosz and Sidner [20] give a theory of discourse structure that is made of three interrelated components; the linguistic structure, intentional structure, and attentional state. The linguistic structure describes the sequences of clauses in text, which aggregate into discourse segments. The intentional structure is used to explain the discourse relevant purposes and their interrelations. The authors state here that cue phrases are the most distinguished linguistic means that a speaker has for both indicating the boundary of a discourse segment and to convey information about the purpose of a discourse segment. The third component of this model, the attentional state, represents the focus of a participant (a reader in the case of written text). This is a dynamic record of the properties and relations of objects currently in focus for a participant. The authors find that intentions are key to explaining the structure of a discourse and its coherence, but also that they are the most difficult aspect to identify. This is in part due to the fact that surface text alone may not provide enough indicators, and that extra-linguistic features present in spoken discourse are required.

In Propp's morphology, the idea of cohesion is present in the form of constraints. Certain character functions are logically connected and have requirements which must be met in order for them to occur. For example, if the *pursuit* function is present, then the *rescue* function must necessarily occur.

3.4 Narrative Comprehension

Narrative comprehension involves the creation of a meaning representation which goes beyond what is present in the surface text of a document. There is a general agreement that humans construct a network-like mental representation of narrative during reading, where events are connected by causal relations [4, 32]. In addition successful comprehension of narrative gives readers the ability to generate good summaries [19].

There are many existing techniques that can be exploited in order to examine and link the contents of a text, and make progress towards a deeper computational representation. For example, coreference resolution is the task of grouping together expressions which refer to the same entities in a text. In addition, authors often express the same concept in a variety of ways through the use of synonymous words. Lexical chains [2] capture sequences of semantically related words in such a manner, which can be used to go beyond simple frequency measures of importance. Furthermore, the sentiment of words, or even their connotation [7] can be used to gain additional information about the emotions conveyed in a text.

3.5 Applications of Propp

Propp's description of a folktale as being comprised of a sequence of simple units, coupled with a description of how these units can be composed to create new tales, has led to its use in the field of computational story generation. Gervás [15, 16] gives an overview of Propp's semi-formal generation procedure as well as discussing a computational approach to generate instances of Russian folktales. Differing generation options are also considered here, as well as their evaluation by metrics inspired by Propp's morphology and his available annotations. In addition, there has been other work in story generation either inspired by Propp's work [33] or using it combination with other methods such as case-based reasoning [5].

Previous work [36] has also attempted the identification of character roles according to Propp's morphology. This approach primarily considers the identification of characters fulfilling the roles of the hero and the villain, leaving the other roles unspecified. Using manually annotated data, Valls-Vargas et al. create role-action matrices, which specify the characters who are the subject and object for each verb in the text. These are used with a genetic algorithm to learn the actions that each type of character typically perform, as well as the actions that different types of character perform to each other. Only considering the assignment of the hero and the villain (the remaining character roles are grouped into a category for 'other'), this method achieves 78.99% classification accuracy over a small dataset of 8 Russian folktales.

Bod et al. [3] present empirical work on the annotation of three single-move Russian folktales by a group of human annotators. The aim of this was to examine the objectivity and reproducibility of Propp's morphology. With limited training, participants of the first study were asked to assign character roles to the story characters as well as annotating three stories with character functions. Then in a subsequent study a different set of participants were given the roles of the dramatis personae and asked to annotate the same three tales. The authors found that providing the character role assignments had a large impact on assignment of character functions to a tale, but that there was low inter-annotator agreement in both studies. In addition to low agreement between participants, the authors found that none of the human annotations matched Propp's own. They claim that this is in part due to the vagueness of some of Propp's function descriptions. We would also consider the limited training that participants received to be an important factor, and that participants of the first study were trained on an example constructed by the authors, rather than an existing folktale. This research was continued by Fisseni et al. [11] which showed that with significantly more training, inter-annotator agreement was much higher and that annotators could reproduce Propp's own function annotations.

Finlayson [10] describes a process of learning Propp's functions from a corpus of deeply annotated Russian folktales [8]. Using a model merging algorithm created with merge rules derived from Propp's morphology, Finlayson demonstrates the feasibility of computationally learning a theory of narrative structure. The outlined method accurately learns to capture events corresponding to some of the key character functions in Propp's morphology, most notably *vilainy/lack*, *struggle/victory* and *reward*.

4 Constraint Logic Programming

Constraint Logic Programming [21] is a form of programming whereby the relations between variables are expressed as constraints. Such programs can then be queried about the provability of a goal. The constraints provide conditions which must be satisfied, either relating the value of one variable to another, or placing restrictions on the values which a variable may take. We consider in particular the Prolog implementation of Constraint Logic Programming over Finite Domains (CLPFD). CLPFD allows for reasoning about variables which have integer values and, among other things, provides a set of arithmetical and membership constraints.

Constraint Logic Programming is appropriate for tasks where there are multiple variables and a solution is required which fits all constraints placed over the set of variables. For the task posed in this paper, we consider the labelling of sentences in a tale with Propp's character functions. Here, each sentence in a tale represents a distinct variable which must take a value from the domain of all character functions, or a zero-value to indicate that the given sentence does not represent a character function. Character functions are given a unique integer identifier, meaning that each sentence must be labelled with an integer value from 0 to 33.[3] Propp describes how "The storyteller is constrained" in several areas, including the "overall sequence of functions, the series of which develops according to the above indicated scheme" (his function scheme) [27, p. 112]. These constraints and restrictions between functions make this problem well-suited to Constraint Logic Programming.

5 System Description

Here we will describe each of the components necessary to pose the assignment of Propp's character functions to a folktale as a constraint satisfaction problem. This primarily involves the detection of potential instances of each character function, and the identification of character roles. This information is used to create the domain of values (character functions) for each sentence, from which a single value must be chosen.

5.1 Preprocessing

In order to facilitate the detection of Propp's character functions and aid the process of identifying the hero and villain of a tale, we perform some preliminary steps to enrich the representation of a tale. We choose to use the Abstract Meaning Representation (AMR) to represent the semantics of our data. The AMR Bank is a manually constructed set of thousands of English sentences paired with their semantic representations [1, 22]. These representations are rooted, directed and labelled graph structures which each correspond to a single sentence. Nodes represent entities in contrast to words, which is the case for dependency parsing and semantic role labelling. AMR parses do not represent tense or arity, however this is desirable for our current work as it simplifies the task of comparing words to a set of cues. In addition, the meta-data for each parse contains word alignment data; the graph fragments to which each span of text corresponds. We use the open source AMR parser, JAMR [13, 12], to perform this step.

To get an indication about the relative importance of the characters in a tale, as well as their role in the story, it is necessary to obtain coreference information. That is, we need to know each of the noun phrases referring to each character in a tale. This is the only step in our process that we currently perform manually. While automatic coreference resolution systems exist, they have issues in their current form which make them unsuitable for tasks such as these [6].

He seized her and he dragged her to his lair.

```
(a / and
        :op1 (s / seize-01
                :ARG0 (h / he)
                :ARG1 (s2 / she))
        :op2 (d / drag-01
                :ARG0 (h2 / he)
                :ARG1 (s3 / she)
                :ARG2 (l / lair
                        :poss (h3 / he))))
```

Figure 1. Example AMR parse with corresponding sentence.

Figure 1 shows an example AMR parse with its corresponding sentence. For parses such as these, we mark each noun phrase with the character (or characters) to which it refers. In this instance, *he* refers to a villainous dragon and *she* refers to a princess.

5.2 Detection of character functions

Before the structure of a tale can be determined, all potential character function assignments must be identified. Our first step in this process is the initial detection of all possible instances of each character function. This allows us to subsequently generate plausible assignments of functions to tales. The majority of the functions can be considered by the presence of cue words. While an instance of a character function, such as *villainy* may in fact span several sentences in the text, a single keyword such as 'attack' is often enough to indicate this. Although on the surface the presence of cue words appear to be a simplistic approach, they are the most prominent linguistic means for conveying the purpose of a discourse segment [20].

For such character functions, we obtain a seed-list of cue words based on the examples discussed in The Morphology of The Folktale and the detailed annotations provided in the data of Finlayson [9]. In order to expand these lists of cue words, with the aim of making them applicable to a wide range of folk tales, we use the lexical resources of WordNet [35] and FrameNet [34] to obtain sets of synonymous

[3] Propp only defines 31 character functions, however we make two additions. The first comes from Propp's description of an *Initial Situation* prior to his first function definition. The second comes from splitting Propp's function VIII and VIIIa (*Villainy/Lack*) into two separate functions in order to ameliorate their detection and obtain a more informative representation of the tale.

words. For each character function, We first find all of the synsets in WordNet and FrameNet which includes at least one of our initial cue words. We then obtain the member words for each of these synsets and find their union, to obtain our expanded list of cue words. However we are careful to omit from our cue word lists what Finlayson [10] calls 'generic events'. These verbs, such as *go*, can be used as an indicator for the majority of character functions, or *say* which is an indicator for every character function. In other words, any of Propp's character functions can occur via an act of speech. In the case of a character function such as *villainy*, this process leads to a wide array of cue words covering acts such as 'exasperate' and 'immolate'.

For each sentence in a tale, we use its corresponding AMR parse to determine all character functions which it could potentially be represented by. That is, for each individual sentence in a tale, we build up a list of character functions which could possibly represent it. In terms of constraint satisfaction, it is in this stage that we construct the domain of values for each individual sentence. Each sentence must be labelled either with one of these potential character functions or a null value to indicate that it does not represent any character function. Evidently, the majority of sentences in a tale will be assigned a null value; the number of sentences in a tale often far exceeds the number of character functions. In addition, Propp assigns at most 12 character functions to a single-move tale in his annotations. For most character functions, detecting their presence is a case of comparing the node-text of an AMR parse with the expanded list of cue words. This is performed with AMR parses rather than the surface text of a tale as the lists of cue words only provide the base form of each word, without all of their possible inflections. We also make use of the parsed text for the detection of non cue-based character functions. For example, Propp describes an *interdiction* function, whereby a command is addressed to the hero. This may either take the form of an order to do something, or a command not to do something. In both cases, this function is generally expressed in direct speech, as an imperative statement. We detect possible occurrences of this function based on the presence of AMR graph fragments of a verb node and its children, where the main verb is missing a subject argument.

As an example, consider again Figure 1. This sentence is initially tagged with the potential of being an instance of: villainy, the hero acquiring the use of a magical agent, a struggle between the hero and villain, a liquidation of lack, and the punishment of the villain. Acquisition of a magical agent and the liquidation of a lack are indicated by *seize*, as these character functions can occur without the consent of other characters. However, with the additional information about the role of each character mentioned, the majority of these options are subsequently ruled out.

In the detection of functions, we are only considering the presence of an overall function type, such as the defeat of the villain (*victory*) rather than distinguishing between its multiple subtypes, such as the villain losing in a contest, being beaten in open combat, being killed etc. Propp does somewhat discuss the pairing of specific forms of of functions; for example, a struggle in an open field (*struggle*) is specifically paired with victory in an open field (*victory*). However these fine-grained pairings and their initial detection is highly dependent on the linguistic choices of the author, an aspect of the tale in which Propp says the storyteller has freedom. The surface text of a tale may state "they fought in an open field", but this does not guarantee an explicit statement to the effect of "victory in an open field". While the act of victory may be explicit, a restatement of the location is superfluous and so may not be present. As such we believe that it would be detrimental to our task to attempt the recognition and assignment of these fine-grained function subtypes. Even ignoring the plethora of conceivable ways by which these subtypes could be expressed in natural language, their successful detection could lead to an excess of allowable function assignments to a tale, giving the tale an excess of plausible interpretations. The only place in which this distinction is made is in the separation of the *villainy* and *lack* function, which Propp defines as functions VIII and VIIIa. These are separated both for detection purposes and for constraint satisfaction; Propp explains that one of these two key events must be present in a tale.

5.3 Character Role Identification

As discussed in Section 2, the narratemes of Propp's morphology are described in terms of the types of character who perform the actions. As such, it is desirable to know the mapping between the actors in a tale and the character roles of Propp's morphology. At this time, we only consider the assignment of characters to the roles of *hero* and *villain*. As stated in the work on character role identification by Valls-Vargas et al. [36], some of Propp's other character roles are unclear, while other roles (*donor* and *helper*) are often performed by the same character. The assignment of multiple character roles to a single character is another reason why we only consider the detection of the hero and the villain; intuitively (for this genre), and in our corpus of Propp's tales, the roles of the hero and the villain are never performed by the same character. Furthermore, we seek to obtain information regarding the character roles for the purpose of reducing the number of character functions which each sentence could plausibly represent. The majority of character functions require the involvement of at least the hero or villain, but do not necessarily require the other character roles. As can be seen from Table 2, there is little gain from the identification of other character roles. This information regarding the number of functions in which each type of character must appear is obtained directly from Propp's Morphology. In his enumeration of narrative units, Propp describes when a certain type of character must necessarily be involved in a given action. Some types of character, such as 'Prize', are not essential to any given character function. The Prize may be present in units such as 'Reward', but the reward could be, for instance, monetary and not require a character.

Table 2. The number of character functions each type of character must necessarily be involved in.

Role	# Character Functions
Hero	22
Villain	8
Donor	3
Dispatcher	0
Helper	0
Prize	0
False Hero	2

In order to determine the hero of a tale, we define a metric of character importance. This metric considers the number of times a character is mentioned, over the space of the tale in which they perform actions. That is, the product of the number of times a character is mentioned and the distance (number of sentences) between their first and final mentions, divided by the distance of the first mention from the end of the text. This requires accurate coreference resolution information, which is another reason why we opt to perform this manually.

Character Importance =

$$\frac{\#mentions * mention\_range}{pos.\_of\_first\_mention}$$

This metric aims to capture the importance of characters who are only mentioned over a short span of text, or are only introduced near the end of a tale, and so cannot have a high total number mentions. It is evident that a character introduced near the denouement of a tale must serve a purpose, however this importance is not captured by simply counting the number of times a character is mentioned.

We assign the role of *hero* to the character which has the highest score according to this metric. We shall discuss the tale *Nikita the Tanner*[4] to illustrate this. The hero, Nikita, is only introduced a third of the way through the tale, while the dragon (the villain) is introduced in the very first sentence. On our AMR-parsed text, both of the aforementioned characters have the same number of referents, however Nikita only has the opportunity to perform actions over a shorter span of text, and so obtains a higher score.

The role of the *villain* is subsequently assigned to the character who has at least one direct interaction with the hero and is involved in the greatest number of verb predicates with negative connotation. We use the connotation lexicon created by Feng et al. [7] in order to identify verb structures with negative connotations. Looking at verb connotations helps to fill in for the common sense knowledge of a reader, to recognise that certain actions may indicate villainous behaviour. The condition of direct interaction is enforced by only considering characters who are involved in at least one AMR verb predicate which also has an argument referring to the hero. The role assignments are performed in this order as the hero of the story is often also an argument of a large number of verbs with negative connotations; being involved in struggles against the villain.

While we find that this approach generally works for this genre, we do make one concession. In some forms of these Russian folktales there are multiple heroes or multiple villains. For example in the tale *The Crystal Mountain*, the hero first defeats a three-headed dragon, then a six-headed and finally twelve-headed dragon. In such instances there is no single antagonist. In this instance the three separate dragons should ideally share the role of villain.

In the assignment of characters to the roles of hero and villain, we make the assumption that these two roles must be filled by two distinct entities. While it is possible to envisage the situation where a single character acts as both the hero and the villain in other genres, we believe this to be highly unlikely in folktales. Propp discusses the roles of several dramatis personae being filled by one story character, however his examples cover the possibilities of combinations such as a donor-helper, donor-villain, or donor-dispatcher character.

5.4 Application of Constraints

The preprocessing of a tale identifies the domain of each sentence (the values of the character functions which each individual sentence could plausibly represent) and the characters taking the roles of the hero and the villain. These terms are summarized below for clarity.

This leads to a search space of narrative structures which could represent the tale. We apply constraints over this space to find valid solutions to our task. We consider this task of assigning a string of character functions to a tale to be one of sentence labelling. Before searching for valid solutions, we go through a process of domain reduction in order to reduce the size of the search space.

Value	A unique integer ID representing a character function.
Domain	The set of values which could be used to represent a given sentence.
Variable	Each variable corresponds to a unique sentence, which must take a value from its associated domain.

5.4.1 Domain reduction

With coreference and character role information, the domain of each sentence can be greatly reduced. As the name implies, Propp's character functions all require the involvement of characters. We trivially reduce the domain of any sentence that does not involve any characters to only contain the zero value. In addition, Table 2 shows that most functions require the involvement of either the hero, the villain, or both. The domain of each sentence is examined to see if it contains any values which require the presence of the hero and/or villain. Such values are removed from the domain of a sentence if the sentence does not contain referents to the required characters. This reflects what Propp terms the *sphere of action*. That is, certain characters have actions which they must necessarily be involved in. For example, Propp describes the sphere of action of the villain as constituting the act of villainy, struggle with the hero, and pursuit.

Considering again the tale of *Nikita the Tanner*, the hero of this tale is only mentioned for the first time in the eleventh sentence. However the fifth sentence is a valid candidate for *departure*, the character function representing the departure of the hero, as this is often indicated by verbs of motion. With coreference information and character role assignment we are able to automatically detect that the fifth sentence makes no mention of the hero and so in fact cannot represent an instance of *departure*. In practice, the process of character role identification allows us to greatly reduce the number of plausible function assignments available to each sentence.

We further make the intuitive decision that the first thirteen functions, up to and including the hero's departure, must occur in the first half of the text. Propp terms the initial functions, those leading up to the lack or act of villainy, as the preparatory functions. The character functions from *lack/villainy* to *departure* represent what Propp calls the complication. We remove the values representing these functions from the domains of sentences occurring in the latter half of a tale. As the functions have a sequential ordering, this removes assignments of character functions whereby all of the values are bunched-up at the end of a tale, and and represent a tale whose first half carries no meaning. This again greatly reduces the size of the search space and the number of valid outputs.

5.4.2 Propp's constraints

In the implementation of constraints we follow what Propp describes in theory, as opposed to what is present in his available annotations of stories. We observe that what Propp describes is more tightly constrained that what his annotations show in practice. In particular, Propp described pairs of functions which must either be both present or both absent from the annotation of a tale. However not all of his annotations strictly conform to these rules.

Propp's sequence of character functions have a canonical order, which may be repeated in separate *moves*. As this current work only considers single-move tales, our implementation imposes a strict ordering over character functions. Each sentence must either be la-

[4] In *Nikita the Tanner*, a dragon has been devouring maidens from the city of Kiev. After the dragon kidnaps a princess and weds her, the king and queen approach the eponymous hero for aid. He reluctantly agrees to help and eventually defeats the dragon, subsequently returning to his work of tanning hides.

belled with a higher value than the previous non-zero label, or itself be labelled with a zero-value. Each non-zero value can only be assigned to one sentence at most. The work of Gervás [16] allows some character functions to swap positions. We do not consider that possibility here, as it would significantly increase the size of our search space.

As is described by Propp, we impose a constraint to ensure that the representation of a tale must include exactly one instance of either an act of villainy or an expression of lack. We impose additional constraints over the values that sentences can be labelled with according to the function pairings that Propp discusses. Some of these are equivalent to logical implications; regarding an interdiction and its violation, Propp states that "the second half can sometimes exist without the first" [27, p. 27]. Other constraints require that a pair of functions are either both present or both absent, villainy and its subsequent liquidation form such a pair [27, p. 53]. Gervás [16] discusses these constraints in the context of story generation as the long range dependencies between certain character functions.

Although not expressly stated by Propp, we place constraints to ensure that a representation of a tale expresses a beginning, middle and end. The motivation for this is to allow the constraint solver to reject strings of character functions which do not represent complete tales. Our decision is reinforced by the findings of Gervás [16] that in the generation of tales according to Propp's morphology, a relatively low percentage of tales had satisfactory endings. Requiring an instance of either villainy or lack provides a beginning element to the tale. To represent the end of a tale, we constrain the assignments of functions contain an instance of either a rescue from pursuit (*rescue*), the hero's return (*return*) or the rewarding of the hero (*reward*). While the return or rewarding of the hero provide a natural end to the tale, Propp states that "A great many tales end on the note of rescue from pursuit" [27, p. 58]. The forms that the midsection of a tale can take are, as is to be expected, exceptionally varied. We aim to keep this variability, while ensuring that some form of action is represented, by placing a loose restriction over the function assignments to make certain that one of the following four events is present: the testing of the hero by the donor, the spatial transference of the hero to the object of a search, the struggle between hero and villain, the pursuit of the hero.

While we have discussed several diversions from a strict adherence to Propp's morphology, these modifications limit the number of outputs from what would otherwise be an infeasible search space. The intuitive restrictions we have imposed merely act to omit the nonsensical outputs, which would not represent a complete tale in any case.

6 Discussion of Outputs

Our approach returns all sequences of character functions which could represent a given tale while satisfying the constraints that we have imposed. The result of this is that we can obtain a large number of interpretations for even short tales, which are only in the order of tens of sentences long. Below is the shorthand[5] for ideal sequence of character functions for the tale *Nikita the Tanner*, as annotated by Propp[6].

$$A\ B\ C \uparrow H\ I\ K \downarrow W$$

Our system does correctly identify this sequence of functions as one potential representation of *Nikita the Tanner*. But it also identifies a further 231 other possible function sequences for this tale which are also valid according to Propp's morphology. These range in length, representing the tale with between four and eleven functions. With our constraints no strings of character functions shorter than this are possible, as these would not represent a complete tale. Propp makes no mention of a minimum number of functions that should be used to represent a tale, however in his annotations he assigns no less than six functions to a single-move tale. Bod et al. [3] make some observations in this manner about the number of functions that human annotators typically assigned to a story in comparison to Propp's own annotations.

While we obtain a significant number of interpretations for such a short tale, it represents only a tiny portion of the original search space, where each of 34 sentences could essentially be labelled with a 0 or an increasing value between 1 and 33. Such a high number of interpretations may help to explain the low inter-annotator agreement that has been found in studies on the replication of Propp's annotations. It also indicates that Propp's morphology is under-constrained for the unambiguous annotation of stories in this way.

We have observed that our approach can give significantly different, but still meaningful interpretations to a single tale. One tale analyzed by Propp, *The Witch*, tells the story of a boy who is kidnapped by a witch while he is out fishing, but eventually manages to escape with the aid of some geese and return home to his parents. Propp's annotations mark this as a tale about an act of villainy, with the subsequent pursuit and rescue of the hero. Our approach is able to detect this interpretation of the story, however it produces significantly different, but arguably correct interpretations in addition to this. One of these interpretations marks this as a tale about the boy's lack of fish and his wish to go fishing, which is granted by his parents. Although this may not be the intended interpretation of the story, the generated sequence of character functions which represent this does conform to Propp's Morphology and is a credible interpretation.

This work sits within our larger body of ongoing work on a creative and cognitively inspired approach to summarization. Our position is that summarization is a creative process. Summarizing a document depends upon an individual's interpretation of a text, and it can be performed to different degrees of abstraction. This uncertainty can allow for a variety of different outputs to be generated from a single story. As such, we view the ability to produce multiple interpretations of a narrative to be a benefit, and allow for the creative generation of summaries. In addition, with knowledge about the narrative structure of a story, we are able to recognise that long passages of text can sometimes be condensed into a much smaller and more abstract concept. This allows us to recognise that passages of text may be expressing more abstract concepts such as *villainy, lack, struggle* or *victory* and generate summaries that are less dependent on the specifics of the surface text of an input.

7 Conclusion and Future Work

In this paper we have provided motivation for our approach to obtaining the narrative structure of a text, by suggesting how it can aid automatic summarization. There is evidence that creating a richer meaning representation of a text, along with an understanding of its structure, should lead to more human-like abstractive summarization. We have described our ongoing work on obtaining the narrative structure for text in a specific domain, that of Russian folktales.

Although Vladimir Propp goes into great detail in his analysis of

[5] These shorthand function names correspond to the function designations given in Table 1.

[6] Propp's annotations do not include any of the Preliminary functions.

Russian folktales, it does not allow for the unambiguous annotation of text; the placement of some character functions is unclear, and some tales appear to be far different from what Propp's annotations would suggest [8]. In addition, the difficulties associated with obtaining accurate human annotations of Propp's morphology have been observed in empirical studies [3, 11]. The issues involved with training human annotators to successfully carry out this task makes it desirable to perform it automatically and consistently. Here we have described a process to obtain an interpretation of a tale, providing a higher level structural representation of the key events. We treat this as a step towards the creative generation of summaries which are more abstractive and closer to those which can be produced by humans.

In our future work we aim to evaluate both our assignment of strings of character functions selected to represent a tale, and the assignment of character roles in comparison to [36]. The evaluation of the character functions used to represent a tale is non-trivial; this is a somewhat subjective task as has been observed from studies on purely human annotations.

Many similarities can be drawn between the Russian folktales analyzed by Propp and tales of other types from around the world. Going forwards, we shall consider producing a generalisation of Propp's morphology which can be applied to a wider range of tales. While the work outlined here considers only a narrow domain of text, the methods described can be applied to any genre for which a sufficiently detailed set of rules and constraints can be specified. With the availability of such rules, high-level interpretations of a text can be produced and aid, what we believe, is a creative approach to summarization.

REFERENCES

[1] Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider, 'Abstract meaning representation for sembanking', in *Proceedings of the Linguistic Annotation Workshop*, (2013).

[2] Regina Barzilay and Michael Elhadad, 'Using lexical chains for text summarization', *Advances in automatic text summarization*, 111–121, (1999).

[3] Rens Bod, Bernhard Fisseni, Aadil Kurji, and Benedikt Löwe, 'Objectivity and reproducibility of proppian narrative annotations', in *Proceedings of the Third Workshop on Computational Models of Narrative. Ed. by Mark Alan Finlayson*, pp. 17–21, (2012).

[4] Gordon H Bower and Daniel G Morrow, 'Mental models in narrative comprehension', *Science*, **247**(4938), 44–48, (1990).

[5] Belén Díaz-Agudo, Pablo Gervás, and Federico Peinado, 'A case based reasoning approach to story plot generation', *Advances in Case-Based Reasoning*, 142–156, (2004).

[6] Maximilian Droog-Hayes, 'The effect of poor coreference resolution on document understanding', *European Summer School in Logic, Language and Information (ESSLLI) 2017 Student Session*, 209–220.

[7] Song Feng, Jun Sak Kang, Polina Kuznetsova, and Yejin Choi, 'Connotation lexicon: A dash of sentiment beneath the surface meaning', in *Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Sofia, Bulgaria, (August 2013). Association for Computational Linguistics.

[8] Mark A Finlayson, 'ProppLearner: Deeply annotating a corpus of russian folktales to enable the machine learning of a russian formalist theory', *Digital Scholarship in the Humanities*, **32**(2), 284–300, (2015).

[9] Mark A Finlayson et al., 'Supplementary materials for "ProppLearner: Deeply annotating a corpus of russian folktales to enable the machine learning of a russian formalist theory"', (2015).

[10] Mark Alan Finlayson, 'Inferring propp's functions from semantically annotated text', *Journal of American Folklore*, **129**(511), 55–77, (2016).

[11] Bernhard Fisseni, Aadil Kurji, and Benedikt Löwe, 'Annotating with propps morphology of the folktale: reproducibility and trainability', *Literary and Linguistic Computing*, **29**(4), 488–510, (2014).

[12] Jeffrey Flanigan, Chris Dyer, Noah A Smith, and Jaime G Carbonell, 'CMU at semeval-2016 task 8: Graph-based amr parsing with infinite ramp loss.', in *SemEval@ NAACL-HLT*, pp. 1202–1206, (2016).

[13] Jeffrey Flanigan, Sam Thomson, Jaime G Carbonell, Chris Dyer, and Noah A Smith, 'A discriminative graph-based parser for the abstract meaning representation', (2014).

[14] Pierre-Etienne Genest and Guy Lapalme, 'Absum: a knowledge-based abstractive summarizer', *Génération de résumés par abstraction*, **25**, (2013).

[15] Pablo Gervás, 'Propp's morphology of the folk tale as a grammar for generation', in *OASIcs-OpenAccess Series in Informatics*, volume 32. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, (2013).

[16] Pablo Gervás, 'Reviewing propps story generation procedure in the light of computational creativity', in *AISB Symposium on Computational Creativity, AISB-2014, April 1-4 2014*, (2014).

[17] Amit Goyal, Ellen Riloff, and Hal Daumé III, 'Automatically producing plot unit representations for narrative text', in *Proceedings of the 2010 Conference on EMNLP*, pp. 77–86. ACL, (2010).

[18] Amit Goyal, Ellen Riloff, et al., 'A computational model for plot units', *Computational Intelligence*, **29**(3), 466–488, (2013).

[19] Arthur C Graesser, Keith K Millis, and Rolf A Zwaan, 'Discourse comprehension', *Annual review of psychology*, **48**(1), 163–189, (1997).

[20] Barbara J Grosz and Candace L Sidner, 'Attention, intentions, and the structure of discourse', *Computational linguistics*, **12**(3), 175–204, (1986).

[21] Joxan Jaffar and J-L Lassez, 'Constraint logic programming', in *Proceedings of the 14th ACM SIGACT-SIGPLAN symposium on Principles of programming languages*, pp. 111–119. ACM, (1987).

[22] Kevin Knight. Abstract meaning representation (amr). http://amr.isi.edu/, 2015. Accessed November 10, 2015.

[23] Wendy G Lehnert, 'Plot units and narrative summarization', *Cognitive Science*, **5**(4), 293–331, (1981).

[24] Chin-Yew Lin, 'Rouge: A package for automatic evaluation of summaries', in *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8. Barcelona, Spain, (2004).

[25] Hans Peter Luhn, 'The automatic creation of literature abstracts', *IBM Journal of research and development*, **2**(2), 159–165, (1958).

[26] Rafael Pérez Ý Pérez and Mike Sharples, 'Mexica: A computer model of a cognitive account of creative writing', *Journal of Experimental & Theoretical Artificial Intelligence*, **13**(2), 119–139, (2001).

[27] Vladimir Propp. Morphology of the folktale. 1928, 1968.

[28] Peter Rankel, John M Conroy, Eric V Slud, and Dianne P O'Leary, 'Ranking human and machine summarization systems', in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 467–473. Association for Computational Linguistics, (2011).

[29] Roger C Schank and Christopher K Riesbeck, *Inside computer understanding: Five programs plus miniatures*, Psychology Press, 2013.

[30] Natalie Schluter, 'The limits of automatic summarisation according to rouge', in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 41–45. Association for Computational Linguistics, (2017).

[31] Karen Spärck Jones, 'Automatic summarising: The state of the art', *Information Processing & Management*, **43**(6), 1449–1481, (2007).

[32] Isabelle Tapiero, Paul van den Broek, and Marie-Pilar Quintana, 'The mental representation of narrative texts as networks: The role of necessity and sufficiency in the detection of different types of causal relations', *Discourse Processes*, **34**(3), 237–258, (2002).

[33] Scott R Turner, 'Minstrel: a computer model of creativity and storytelling', (1993).

[34] Berkeley University. About framenet. https://framenet.icsi.berkeley.edu/fndrupal/, 2017.

[35] Princeton University. About wordnet. http://wordnet.princeton.edu, 2010.

[36] Josep Valls-Vargas, Santiago Ontanón, and Jichen Zhu, 'Toward character role assignment for natural language stories', in *Proceedings of the Ninth Artificial Intelligence and Interactive Digital Entertainment Conference*, pp. 101–104, (2013).

Storifying Observed Events: Could I Dress This Up as a Story?

Pablo Gervás [1]

Abstract. The format preferred by people to receive reports on events that have been observed is a story. Sometimes real life events inspire a story, but either lack the structure or the clear motivation for the characters that one would expect in a story. When this happens, a process of "fictionalising" these real life events can be applied. This process usually creates a discourse in which the real life events may have been filtered, adapted or extended, possibly with additional material added, and which presents the type of causally connected structure we are used to observing in a story. We call this process *storifying* the events. The present paper postulates one possible computational model of how this process is carried out. Based on the record of piece movements for a chess game, and a set of schemas for plot, the model selects narrative threads for particular pieces (based on the concept of pieces having a restricted view of the whole board), finds the portions of those threads that match plot schemas, and uses them to instantiate the schemas into stories.

1 Introduction

Narrative plays a significant role in human communication as the vehicle generally employed by individuals to convey to others information about events that have been observed. Yet the mechanisms by which such narratives are constructed from the basic building blocks of a set of events are badly understood. In a world where technological advances are progressively making it possible to extract basic information about events from multiple sources (surveillance videos, postings in social networks, records of change of location of specific devices, sensors, smart objects), there is a need for solutions that can model the ability of humans to sift through large amounts of event descriptions acquired in this fashion and automatically carry out the task of selecting and combining a subset of these into pseudo-narrative formats that can act as adequate renderings of the part of what has happened that is worth reporting.

To achieve this, it would be extremely useful to have accurate models of how humans construct narratives from observed experience, and how these processes address the task of selecting particular events to mention while omitting others, how they postulate particular connections among these events to provide a body for the narrative, and how they choose to arrange the selected material into the linear sequence of statements that constitute a narrative.

Beyond basic reporting of observed events, where faithful rendering of fact is fundamental, humans have developed a more elaborate form of storytelling, in which departure from accurate fact is allowed (even encouraged) if it achieves certain desired (literary?) effects. These effects may take different forms, including making the stories easier to remember, conveying a particular message in a subtle way, or providing pure entertainment value. To achieve such effects human storytellers operating from an inspiring set of facts apply a number of operations. Again, while most of us have seen these operations applied in film or literature to repackage episodes from reality in fictional form, very little is known about them from the computational point of view. Yet endowing a computer with the ability to so enhance bare-bones information to make it easier to remember or simply more entertaining might go a long way towards reducing the feeling of dry-fact presentation that one gets from computer generated material.

The present paper addresses these problems by presenting a computational model of how input data that record patterns of movement and interaction between basic agents is mined for possible pseudo-narratives that present a significant subset of the observed events packaged into a sequence of statements that exhibits desirable properties that make it resemble narratives as preferred by humans.

2 Related Work

A number of academic disciplines – narratology, psychology, artificial intelligence – have focused on narrative from various points of view. Yet the ability to build story-like discourses from conceptual records of experience has very rarely been addressed, as it lies much at the gaps between disciplines – too elementary to be considered by literary studies, more elaborate than other yet to be understood abilities to be addressed in experimental psychology, and side-lined by artificial intelligence as less glamorous than *ex novo* generation of stories.

2.1 Narrative

Narrative has been considered as an elementary cognitive ability relevant for human beings [49, 7, 29]. Yet the process by which a particular experience of reality gets transformed into a narrative in the classic sequential sense that we consider a "story" is poorly understood. In recent years there has been a significant effort to relate narrative to the study of human cognition [28, 30]. It is clear that this line of research constitutes a major challenge, given the levels of complexity involved in both narrative and human cognition. The picture to be considered is complex and full of open questions.

Insights on narrative may also be obtained from a number of related disciplines, such as narratology, psychology, cognitive science and creative writing.

An important obstacle that faces this challenge is the fact that humans are notoriously poor at identifying the processes that they apply

[1] Instituto de Tecnología del Conocimiento, Universidad Complutense de Madrid, email: pgervas@ucm.es

in processing reality [40]. As a result, we are faced with the task of postulating the underlying latent processes from the observation of their external manifestation. Observable manifestations are the actual *narratives as literary works*, which are studied by narratology, or the *processes by which humans produce narratives*, which are studied from different points of view by cognitive science and creative writing. Another possible approach is to consider the *role of simulation* in the understanding of how narrative works.

2.1.1 Narratives as Products

Relevant concepts from the field of narratology [1] are the distinction between *fabula* – the set of events behind a story, independently of how they are rendered – and *discourse* – the particular way of rendering a given fabula as a sequence of statements – and *focalization* [11] – the way in which a story is told from the view point of particular characters, switching between them at need to cover what happens elsewhere.

Existing narratives can very rarely be paired with alternative records of the experience that led to them, or even the events that are represented in them. This is a significant obstacle for applying a data-driven approach to model narrative construction computationally, as these approaches require instances of both the input that lead to the communication impulse, the narrative that arose from it, and possibly representations of intermediate design decisions.

2.1.2 Narrative Construction as a Process

Two different processes on narrative have been studied by cognitive science: comprehension and writing.

Narrative comprehension involves progressive enrichment of the mental representation of a text beyond its surface form by adding information obtained via inference, until a situation model (representation of the fragment of the world that the story is about) is constructed [55]. Trabasso et al [53] postulate comprehension as the construction of a causal network by the provision by the user of causal relations between the different events of a story. This network representation determines the overall unity and coherence of the story. These insights need to be considered in the identification of the relevant aspects to be represented for a fabula.

Cognitive scientists have proposed models of the *writing task*. Flower and Hayes [10] define a cognitive model of writing in terms of three basic process: planning, translating these ideas into text, and reviewing the result with a view to improving it. These three processes are framed by what Flower and Hayes consider “the rhetorical problem”, constituted by the rhetorical situation, the audience and the writer’s goals. This corresponds to the contextual parameters considered in the present proposal. Sharples [50] presents a description of writing understood as a problem-solving process where the typical writer alternates between the simple task of drafting possible additions to his text and the more complex task of reflecting on how the text matches his goals to review what to do next. This type of feedback loop based on satisfaction of the stated goals needs to be considered both in fabulation and discourse composition processes.

Creative writing emerged as a specific discipline to obtain insights into the processes that lead to the production of narrative. The difference in purpose with traditional treament of literature in the humanities has been identified as an open question that needs solving [31].

2.1.3 The Role of Simulation

Disciplines such as social psychology have long accepted the role of computer simulation as a useful tool for addressing research problems that are difficult to represent linguistically or mathematically [42]. Computational modeling of processes of narrative construction allows us to study how they replicate observed human behaviour as well as how they operate internally. This has a potential for yielding insights on how humans address the same tasks.

2.2 Automated Story Telling

Storytelling efforts in AI have focused on two different tasks: building fictional plots from scratch and structuring appropriate discourse for conveying a given plot. Solutions to *build fictional plots* [12] rely on different techniques, such as grammars [33, 6, 34]– to build stories according to a particular structure –, planning [38, 36, 48] – to build stories that reach particular given goals –, reuse [54, 44, 18, 47, 41] – to build stories that resemble previous instances of valid stories –, or simulation of world dynamics [52] – to build stories that emerge from the interactions of modelled characters. Solutions to *build a discourse that renders a given plot* have been developed for the logs of a social simulation system [27] and for constructing cinematic visual discourse [3, 32]. My own work [13, 14, 15] pioneered the task of first identifying valuable stories from the record of a chess game and then generating natural language renderings of them. However, the narratives resulting from this effort lacked a clear concept of plot, which is a central focus of the present paper.

Following a general trend in computational creativity to develop generative systems that are capable of carrying out some evaluation of their outputs – as human creators do – there has been considerable progress in the development of *metrics for automatically generated narratives* [43, 16, 51] .

Different storytelling systems tend to focus on the representation and manipulation of a particular subset of the *possible relevant aspects* [21], whereas full-fledged solutions to the problem are unlikely to succeed unless they provide sufficient coverage of the complete range of relevant features.

Existing story generation systems often rely on extremely simple solutions for rendering their results as text [8], far removed from the state of the art in natural language generation. This disconnect – between the set of events that can be generated and the well-structured discourse plan that an NLG system needs to produce adequate prose to narrate them – may partly be resolved by the consideration of storytelling as a form of data to text generation. The present paper proposes a possible avenue in which to address this issue.

2.3 Natural Language Generation

Natural language generation (NLG) studies the automated construction of text documents from input data [46]. It is traditionally considered in terms of three different phases: *content planning* – deciding what to say and how to organize it into a structured set of sentences, or a *discourse plan* –, *sentence planning* – deciding how to structure each of those sentences internally –, and *surface realization* – deciding how to convey each sentence as text. Academic efforts in the recent past have shown a tendency to focus on sentence planning and surface realization, partly due to the fact that content planning tends to be very dependent on the particular domains of application, and scientific work on content planning faces a strong requirement of having access to appropriate input data for the domain in a machine-readable format.

Content planning is usually considered in terms of two different operations. *Content determination* is the task of identifying which facts from the input data are to be included in the intended message. *Discourse planning* is the task of establishing a particular ordering and structuring for the discourse created to convey a particular message. Existing efforts to model these tasks have focused on construction of texts to report sporting events [2, 5, 35], or generation of elaborate narrative variations for sequences of user actions in interactive fiction [39].

The present proposal addresses the task of constructing a story to match a set of input data in terms of a specific stage of content planning based on matching the input data with known narratives schemas, and using the match to drive both the selection (content determination) and the organisation (discourse planning) of the content to be conveyed. The complete transcription of the planned discourse to text is not considered in this proposal, as state of the art solutions exist that could be applied to solve that task [8].

2.4 Computational Narratology

Emerging in recent times at the joining point of computer science and narratology, computational narratology [37] focuses on the algorithmic processes involved in creating and interpreting narratives, modeling narrative structure in terms of formal, computable representations. Much of the work carried out in artificial intelligence could be considered computational narratology, as the borders are considerable blurred.

Originally based on accounts of narrative structure in narratology, recent advances have proposed formal computable representations for plot [22], an enriched vocabulary of representational abstractions of narrative content [20], procedures for generating plot structures [19, 26] and procedures for composing narrative discourse from an input set of data [23, 24, 15, 25].

3 Storifying

A computational model of the task of storifying a set of observed events must address a number of tasks. First, it needs to be able to see the events from the point of view of the participating agents. This is the process known in narratology as focalisation, and it partitions experience into narrative threads centred on particular characters. Second, it needs a representation of the structure expected for a story. Existing accounts of archetypal plots will be of use here. Third, it needs to establish mappings between the narrative thread for some character and some instance of archetypal plot. This is the key to the process. The mapping should provide the information required to instantiate the plot with the characters from the observed events. Metrics must be provided to measure the degree to which the mapping respects the information in the observed events used as inspiration. Finally, it would need to generate a readable version of the resulting discourse.

The solution for storification described here has been implemented as an application named *StoryFire*.

3.1 Focalised Representation of Events

The task of addressing computationally the partition of experience into narrative threads centred on particular characters had already been addressed in [23, 15]. We adopt here the solution proposed there, based on the establishment of a range of perception for each agent which determines how much of the reality around her she perceives at any given moment in time. This requires explicit representation of space and explicit encoding of the location of both events and observing agents. The simplest way of achieving this is by relying on a simple two-dimensional grid. By applying this constraint, a representation can be obtained for the narrative thread for each character by compiling into a linear thread all the events that fall within the range of perception of the agent over time. In this way, a *fibre* is a sequence of events that either involve or are seen by a given character. It represents a focalized perception of the world.

The task of *heckling* involves establishing the range of perception, tracking the set of all possible characters involved in the events to be narrated, and for each character constructing a fibre representation that includes descriptions of all the event that the character initiates, suffers or perceives. These descriptions will appear in the fibre in chronological order.

A short example of a fibre – extracted from the application to telling stories from a chess game developed in section 3.5 – is given in Table 1. It describes what the focalising character can see at a given point in time, separated into a descriptive section that accounts for static information and a narrative section that accounts for changes occurring at this particular point in time. This is verbose to guarantee that all relevant information is registered. When actually rendering this information, whatever has not changed from a previous stage is omitted.

```
START-FIBRE for : lwr
[
Focalizer: lwr
Time: 7
Date: 7
a 1 /
Perception Range: 2

DESCRIPTIVE:
  is_at(wp1, a 2)

  is_at(lwk, b 1)

  is_at(wp2, b 2)

  is_at(lwb, c 1)

  is_at(wp3, c 2)

NARRATIVE:
  leaves_from(wp3, c 2)

]

(...)

END-FIBRE
```

Table 1. Example of a short fibre focalised on chess piece `lwr`, includes snapshot of the fabula at time 7, in which the focaliser, at a point where it can see pieces `wp1`, `lwk`, `wp2`, `lwb` and `wp3` around it, notices piece `wp3` leave

3.2 Representing Archetypal Plots

The hypothesis on which we base our current approach to storifying is that the storifier applies to the observed set of events a set of pre-existing frames for stories, and selects the best pairing between a subset of the observed event and one of the possible storytelling frames. Other approaches are possible, but this seemed a plausible baseline to start the research.

As a computational approximation of this type of pre-existing storytelling frame we turn to existing work on formal computable representations for plot. Existing solutions rely on a representation of plot as a succession of labels that represent units of abstraction of plot-relevant actions by the characters, along the lines of Propp's concept of a *character function* [45]. Such representations have been used to build a set of narrative schemas for plot [22] and even to develop a case-based solution for generating plots in terms of them [19]. However this type of representation restricted to flat labels does not hold enough data to inform a subsequent process of instantiation with knowledge from real life. A plot as a storytelling frame is tied together by relations that need to hold between the elements that compose it, such as who the hero and the villain are, and what relative roles they play in the elements used to build the plot line.

For this reason, in the present paper we rely on an enriched representation of plot. A *plot frame* has a basic skeleton that is indeed as sequence of labels for character-function-like elements (referred as *plot elements*), but holds additional information to indicate what roles are relevant to the plot (hero, villain, victim,...) and who the protagonist of the plot is. The roles used for this purpose were originally based on Propp's concept of the *dramatis personae* of a Russian folk tale but had to be extended to account for other types of stories. The need for explicit indication of who the protagonist arose from the observation made in [22] that archetypal plots for *Overcoming the monster* and *Tragedy* were very similar in structure, and only differentiated by who the protagonist is (the hero in one, the villain in the other). Each plot element has a more specific set of roles that describe how the characters take part in it. For instance, an *Abduction* involves an *abductor* and an *abductee*. In most instances of plot, the abductor is the villain, but this need not always be the case. For this reason, each instance of a plot element occurring within a plot explicitly provides a mapping between the narrative roles for the plot and the specific roles for the plot element.

The plot frames considered in the present paper are instantiations of the seven basic plots defined by Booker [4]. These are not considered to be exhaustive but constitute a good set for the initial trials. Extension of the set of plot frames will be considered as further work. An example of a short plot frame is given in Table 2.

```
PLOT FRAME = Comedy-UnrelentingGuardian
PLOT PROTAGONIST = hero
PLOT ROLES = hero love-interest obstacle

PLOT-START

PLOT ELEMENT NAME = CoupleWantsToMarry
ROLE-DATA
lover hero
beloved love-interest

PLOT ELEMENT NAME = UnrelentingGuardian
ROLE-DATA
lover hero
beloved love-interest
guardian obstacle

PLOT ELEMENT NAME = HighStatusRevealed
ROLE-DATA
lover hero
beloved love-interest
guardian obstacle

PLOT ELEMENT NAME = Wedding
ROLE-DATA
lover hero
beloved love-interest

PLOT-END
```

Table 2. Archetypal Plot for Unrelenting Guardian Comedy Plot

3.3 Storifying: matching an observed thread of events to a known plot frame

The establishment of mappings between the narrative thread for some character and some instance of archetypal plot would ideally consider all available information about what the character does in the thread and what it is expected to do in the archetypal plot. For this paper, we will consider only a first approximation of how basic mappings may be established. More elaborate solution may be considered later once the overall feasibility of the approach has been tested.

A *mapping* between a thread and a plot frame involves an alignment between a subset of the events in a thread and the sequence of plot elements in a plot frame, and a correspondence between the characters present in the thread and the plot roles in the plot frame. An example of such a mapping is given in Table 3.

```
BEGIN thread to plot frame match
Thread lwk
PlotFrame Comedy-UnrelentingGuardian
Score 83

alignMENT
9 [0]
11 [1]
16 [2]
17 [3]
MAPPING
bp4=love-interest
rwb=obstacle
lwk=hero
END thread to plot frame match
```

Table 3. Mapping between thread and plot frame

The basic constraint to be satisfied when matching a thread and a plot frame is that, at each point where an event is aligned with a plot element, the plot roles for the plot element are appropriately instantiated with characters present in the event according to the correspondence established in the mapping.

We propose a baseline algorithm in two stages. The first stage establishes a set of possible correspondences between the set of characters in the thread and the set of roles in the plot frame. This is done by assigning the focaliser of the thread to the protagonist of the plot frame, and considering all other possible assignments of remaining characters in the thread to the remaining roles in the plot frame. The second stage identifies the best possible alignment between thread events and plot elements in the plot frame. For each of the possible correspondences between characters and roles, an assignment of roles is made to the characters in the thread. the sequence of plot elements in the plot frame is then traversed, trying to match the set of roles involved in the current plot element with the set of roles now assigned to the characters present in the next event in the thread. If the set of roles assigned to the characters present in the event matches at least 50% of the set of roles involved in the plot element, they are considered aligned, if not that event is skipped. A valid alignment results if the end of the sequence of plot elements in the plot frame is reached before the events in the thread run out.

For each valid alignment a score is computed as the average of the percentage of satisfaction of set of roles involved in the plot element by roles assigned to the characters present in the event, over the whole set of plot elements. This constitutes an acceptable baseline metric to measure the degree to which the mapping respects the information in the observed events used as inspiration.

3.4 Instantiating a Plot Frame

A plot frame is an abstract representation of a possible story structure. As such, it needs to be instantiated into a story by providing the additional detail that has been omitted during the process of abstraction. The task of instantiating abstract representations of stories had already been addressed in [17] for the case of Russian folk tales. In this paper we rely on an updated version of that procedure to instantiate plot frames into conceptual descriptions of stories. To account for the broader range of stories covered by the archetypal plots considered, we have expanded the original vocabulary of story actions to those considered by the Propper Wryter system [20], which was used to generate the plot of the *Beyond the Fence* musical [9].

3.5 Storifying Partial Views of a Chess Game

To provide a preliminary benchmark for the various intuitions outlined in the rest of the paper the simplest approximation to a case study that could be conceived is described in this section. This is done by considering a chess game as a very simple model of a formalised set of events susceptible of story-like interpretations. Chess provides a finite set of characters (pieces), a schematical representation of space (the board) and time (progressive turns), and a very restricted set of possible actions. Operating on simple representations of a chess game in algebraic notation, exploratory solutions for the tasks of content selection and content planning are explored based on a fitness function that aims to reflect some of the qualities that humans may value on a discourse representation of a story.

1. d4 d5	11. Bc2 h6
2. Nf3 Nf6	12. b3 b6
3. e3 c6	13. Bb2 Bb7
4. c4 e6	14. Qd3 g6
5. Nc3 Nbd7	15. Rae1 Nh5
6. Bd3 Bd6	16. Bc1 Kg7
7. O-O O-O	17. Rxe6 Nf6
8. e4 dxe4	18. Ne5 c5
9. Nxe4 Nxe4	19. Bxh6+ Kxh6
10. Bxe4 Nf6	20. Nxf7+ 1-0

Table 4. Algebraic notation for an example chess game

Each individual chess piece taking part in the game is considered a character. Perception range is defined as the small space of N x N squares of the board that constitutes that immediate surroundings of each piece at any given moment.

Events are triggered by pieces moves. Whenever a piece moves, this constitutes an event for the piece itself, for any other piece captured during the move, and for any other piece that sees either the full move, the start of the move or the conclusion of the move.

Fibres for each of the pieces are built by collecting event descriptions for those moves that they are involved in or they see. The same event may get described differently in different fibres depending on the extent to which the corresponding focalizer is involved in it.

An example of how the storification process applies to the chess game given in Table 4 is shown in Figure 5. The figure shows the partial views of the game as seen by the focaliser (in this case, the left white knight) for the events of his thread that have been aligned with the UnRelenting Guardian plot frame shown in Table 2. This corresponds to the best scoring mapping found for pairing the plot frame with threads from the game. Further examples of storification of other threads from the game are shown in Appendix A.

The process of rendering the conceptual description of a story as text introduces in itself a number of compacting solutions (aggregation, ommission, replacement of nouns with anaphoric pronouns...) that somewhat obfuscate the data to which it is applied. In order to allow the reader to evaluate directly how well the results of the described storification process respect the input data, and how much additional material has been introduced in each case, the examples given below include the conceptual representation of the resulting story rather than its text rendering.

4 Discussion

The process of storification as described takes data on observed movement of characters and superimposes on them a layer of possible motivation for their actions. The information on such motivation cannot normally be observed and has to be inferred by viewers. Humans are very good at this task, and much of the information they obtain about the events they observe results from such processes of inference. The procedure proposed in this paper replicates such functionality at a very basic level.

When humans carry out these processes to interpret reality, their purpose is usually to compile information on the observed characters with a view to predict future behaviour. However, in cases of storification, departure from truth is generally accepted as a tool of the craft. To make the result interesting the storyfier can introduce conflicts that were not apparent, or take sides for one of the characters, and from that point on minimise references to their shortcomings and maximise those of their rivals. In some cases, characters may be introduced to play the role of rivals if none were available in the observed events.

The procedure described here relies on these allowances to provide a baseline storification process that produces acceptable simple stories that respect the observed relevant features of the events they are based on. In doing so, some events from the thread are omitted if they are not considered relevant to the plot frame under consideration. Some characters present in the scene may be omitted from the story if they play no relevant role in the plot being told. These solutions conform to acceptable practice when telling a story.

Formal evaluation of this type of storification presents several important difficulties. The most relevant is that the storification of a given set of events is, by definition, subjective. Given sequence of snapshots of a game – as the one shown in Figure 5 – human judgements on the plausibility of a given storification for it, or on the entertainment value of the resulting story, may be collected. However, only very extreme negative values would be damaging for the validity of the process.

5 Conclusions

Storification of observed events can be modelled computationally with very basic baseline solutions for the intervening steps. Whereas there may not be an immediate practical application of this type of process, we believe it to be a fundamental ingredient of the human storytelling capacity. As such, computational models of it are useful *per se* as accounts of how the task may be carried out. In the process of developing the one reported in the present paper, important insights on the nature of plot – such as the need to represent explicitly protagonism, narrative roles, and mapping of narrative roles to specific plot elements – and the process of content determination – how the requirement of a successful alignment between observed event and intended plot frame forces selection or ommission of particular

Move: 9

	a	b	c	d	e	f	g	h
1	■	■	■	■	■	■	■	■
2	■	■	■	■	■	■	■	■
3	■	■	■	■	■	■	■	■
4				p		■	■	■
5			P	**P**		■	■	■
6			**N**		P	■	■	■
7	P	P				■	■	■
8	R		B	Q	K	■	■	■

character lwk (**N**)
character wp4 (**P**)
mutual_love lwk wp4
want_to_marry lwk wp4

The left white knight and the fourth white pawn are in love. They want to get married.

Move: 11

	a	b	c	d	e	f	g	h
1	■	■	■	■	■	■	■	■
2	■	■	■	■	■	■	■	■
3	■	■	■	■	■	■	■	■
4				p		■	■	■
5			P	**P**		■	■	■
6			**N**	**B**	P	■	■	■
7	P	P				■	■	■
8	R		B	Q	K	■	■	■

character rwb (**B**)
(guardian rwb wp4)
opposed_to_plan rwb
sundered lwk wp4

The right white bishop is the guardian of the fourth white pawn. The right white bishop is opposed to their union.

Move: 16

	a	b	c	d	e	f	g	h
1	■	■	■	■	■	■	■	■
2	■	■	■	■	■	■	■	■
3	■	■	■	■	■	■	■	■
4						■	■	■
5			P	**P**	p	■	■	■
6			**N**	**B**		■	■	■
7	P	P				■	■	■
8	R		B	Q	K	■	■	■

(different_class lwk wp4)
high_status_revealed lwk
¬ sundered lwk wp4

The high status of the left white knight is unexpectedly revealed. The right white bishop relents in his opposition.

Move: 17

	a	b	c	d	e	f	g	h
1	■	■	■	■	■	■	■	■
2	■	■	■	■	■	■	■	■
3	■	■	p	b	p	n		■
4	■	■						■
5	■	■	P	**P**	**N**			■
6	■	■		**B**		N		■
7	■	■				P	P	■
8	■	■	■	■	■	■	■	■

marry lwk wp4

The left white knight and the fourth white pawn get married.

Table 5. Storification as a Comedy of the thread for the left white knight (lwk, represented in the diagrams as **N**) in terms of his romance with the fourth white pawn (wp4, represented in the diagrams as **P**) in the face of opposition of is guardian the right white bishop (rwb, represented in the diagrams as **B**).

events or characters – have emerged. In addition, they may provide useful tools to enhance existing storytelling solutions.

Many refinements of the proposed procedure are possible. At present, baseline decision making has been applied at all the relevant stages. Detection of co-location of characters required as a prerrequisite for interaction is currently based on co-presence of both within the perception range of one another. Proximity may be introduced as a further refinement. The establishment of mappings between characters and roles is currently done by exhaustive testing of all possible combinations. Informed procedures at this point may lead to more efficient implementations. The metric for satisfactory alignment is currently opaque to the semantics of the events and the plot elements. In all these cases, the fact that the baselines solutions employed lead to acceptable results suggest that investing effort in exploring more refined solutions would be worthwhile.

ACKNOWLEDGEMENTS

This paper has been partially supported by the IDiLyCo project (TIN2015-66655-R) funded by the Spanish Ministry of Economy, Industry and Competitiveness and the FEI-EU-17-23 project InViTAR-IA: Infraestructuras para la Visibilización, Integración y Transferencia de Aplicaciones y Resultados de Inteligencia Artificial.

REFERENCES

[1] H.P. Abbott, *The Cambridge Introduction to Narrative*, Cambridge Introductions to Literature, Cambridge University Press, 2008.

[2] N. D. Allen, J. R. Templon, P.S. McNally, L. Birnbaum, and K.Hammond, 'Statsmonkey: A data-driven sports narrative writer', in *Computational Models of Narrative: AAAI Fall Symposium 2010*, (2010).

[3] B.-C. Bae and R. M. Young, 'A use of flashback and foreshadowing for surprise arousal in narrative using a plan-based approach', in *Proc. ICIDS 2008*, (2008).

[4] C. Booker, *The Seven Basic Plots: Why We Tell Stories*, The Seven Basic Plots: Why We Tell Stories, Continuum, 2004.

[5] N. Bouayad-Agha, G.Casamayor, and L. Wanner, 'Content selection from an ontology-based knowledge base for the generation of football summaries', in *Proc. ENLG 2011*, pp. 72–81, (2011).

[6] S. Bringsjord and D. A. Ferrucci, *Artificial Intelligence and Literary Creativity: Inside the Mind of BRUTUS, a Storytelling Machine*, Lawrence Erlbaum Associates, 1999.

[7] J. Bruner, 'The narrative construction of reality', *Critical inquiry*, 1–21, (1991).

[8] C. B. Callaway and J. C. Lester, 'Narrative prose generation', *Artif. Intell.*, **139**(2), 213–252, (August 2002).

[9] Simon Colton, Maria Teresa Llano, Rose Hepworth, John Charnley, Catherine V. Gale, Archie Baron, François Pachet, Pierre Roy, Pablo Gervás, Nick Collins, Bob Sturm, Tillman Weyde, Daniel Wolff, and James Robert Lloyd, 'The beyond the fence musical and computer says show documentary', in *Seventh International Conference on Computational Creativity (ICCC 2016)*, Paris, France, (06/2016 2016). Sony CSL, Sony CSL.

[10] L. Flower and J.R. Hayes, 'A cognitive process theory of writing', *College Composition and Communication*, **32**(4), 365–387, (1981).

[11] G. Genette, *Narrative discourse : an essay in method*, Cornell University Press, 1980.

[12] P. Gervás, 'Computational approaches to storytelling and creativity', *AI Magazine*, **30**(3), 49–62, (2009).

[13] P. Gervás, 'From the fleece of fact to narrative yarns: a computational model of narrative composition', in *Proc. Workshop on Computational Models of Narrative 2012*, (2012).

[14] P. Gervás, 'Stories from games: Content and focalization selection in narrative composition', in *First Spanish Symposium on Digital Entertainment, SEED 2013*, (2013).

[15] P. Gervás, 'Composing narrative discourse for stories of many characters: a case study over a chess game', *Literary and Linguistic Computing*, **29**(4), (08/14 2014).

[16] P. Gervás, 'Metrics for desired structural features for narrative renderings of game logs', *Journal of Entertainment Computing*, (08/14 2014).

[17] P. Gervás, 'Computational Drafting of Plot Structures for Russian Folk Tales', *Cognitive Computation*, (07/2015 2015).

[18] P. Gervás, B. Díaz-Agudo, F. Peinado, and R. Hervás, 'Story Plot Generation Based on CBR', *Knowledge-Based Systems. Special Issue: AI-2004*, **18**, 235–242, (2005).

[19] P. Gervás, R. Hervás, and C. León, ‘Generating plots for a given query using a case-base of narrative schemas’, in *Creativity and Experience Workshop, International Conference on Case-Based Reasoning*, Bad Homburg, Frankfurt, Germany, (09/2015 2015).

[20] P. Gervás, R. Hervás, C. León, and C. V Gale, ‘Annotating musical theatre plots on narrative structure and emotional content’, in *Seventh International Workshop on Computational Models of Narrative*, Kravov, Poland, (07/2016 2016). OpenAccess Series in Informatics, OpenAccess Series in Informatics.

[21] P. Gervás and C. León, ‘The need for multi-aspectual representation of narratives in modelling their creative process’, in *2014 Workshop on Computational Models of Narrative*, Quebec City, Canada, (07/2014 2014). Scholoss Dagstuhl OpenAccess Series in Informatics (OASIcs), Scholoss Dagstuhl OpenAccess Series in Informatics (OASIcs).

[22] P. Gervás, C. León, and G. Méndez, ‘Schemas for narrative generation mined from existing descriptions of plot’, in *Computational Models of Narrative*, Atlanta, Georgia, USA, (05/2015 2015). Scholoss Dagstuhl OpenAccess Series in Informatics (OASIcs), Scholoss Dagstuhl OpenAccess Series in Informatics (OASIcs).

[23] Pablo Gervás, ‘From the fleece of fact to narrative yarns: a computational model of composition’, in *Workshop on Computational Models of Narrative, 2012 Language Resources and Evaluation Conference (LREC’2012)*, Istambul, Turkey, (05/2012 2012).

[24] Pablo Gervás, ‘Stories from games: Content and focalization selection in narrative composition’, in *I Spanish Symposium on Entertainment Computing*, Universidad Complutense de Madrid, Madrid, Spain, (09/2013 2013).

[25] Pablo Gervás, ‘Empirical determination of basic heuristics for narrative content planning’, in *Proceedings of Computational Creativity and Natural Language Generation Workshop, International Conference on Natural Language Generation (INLG 2016)*, Edimburgh, Scotland, (2016).

[26] Pablo Gervás, ‘Comparative evaluation of elementary plot generation procedures’, in *6th International Workshop on Computational Creativity, Concept Invention, and General Intelligence*, Madrid, Spain, (12/2017 2017).

[27] S. Hassan, C. León, P. Gervás, and R. Hervás, ‘A computer model that generates biography-like narratives’, in *International Joint Workshop on Computational Creativity. London*, (2007).

[28] D. Herman, *Narrative Theory and the Cognitive Sciences*, CSLI Publications, CSLI Publications, 2003.

[29] D. Herman, *Story Logic: Problems and Possibilities of Narrative*, Frontiers of narrative, University of Nebraska Press, 2004.

[30] D. Herman, *Storytelling and the Sciences of Mind*, Cambridge, MA, 2013.

[31] P. Howarth, ‘Creative writing and schiller’s aesthetic education’, *The Journal of Aesthetic Education*, **41**(3), 41 – 58, (2007).

[32] A. Jhala and R. M. Young, ‘Cinematic visual discourse: Representation, generation, and evaluation’, *IEEE Trans. on Comp. Int. and AI in Games*, **2**(2), 69–81, (2010).

[33] S. Klein, J. F. Aeschliman, D. F. Balsiger, S. L. Converse, C. Court, M. Foster, R. Lao, J. D. Oakley, and J. Smith, ‘Automatic novel writing: A status report’, Technical Report 186, Computer Science Department, The University of Wisconsin, Madison, Wisconsin, (December 1973).

[34] R. Raymond Lang, ‘A declarative model for simple narratives’, in *Proceedings of the AAAI Fall Symposium on Narrative Intelligence*, pp. 134–141. AAAI Press, (1999).

[35] F. Lareau, M. Dras, and R. Dale, ‘Detecting interesting event sequences for sports reporting’, in *Proc. ENLG 2011*, pp. 200–205, (2011).

[36] M. Lebowitz, ‘Story-telling as planning and learning’, in *Proc. IJCAI 1983*, volume 1, (1983).

[37] Inderjeet Mani, ‘Computational narratology’, *Handbook of narratology*, 84–92, (2014).

[38] J. R. Meehan, ‘TALE-SPIN, An Interactive Program that Writes Stories’, in *Proc. IJCAI 1977*, pp. 91–98, (1977).

[39] N. Montfort, *Generating narrative variation in interactive fiction*, Ph.D. dissertation, University of Pennsylvania, Philadelphia, PA, USA, 2007.

[40] Richard E. Nisbett and TImothy Wilson, ‘Telling More Than We Can Know: Verbal Reports on Mental Processes’, *Psychological Review*, **84**(3), 231–259, (1977).

[41] S. Ontañón and J. Zhu, ‘On the role of domain knowledge in analogy-based story generation’, in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two, IJCAI’11*, pp. 1717–1722. AAAI Press, (2011).

[42] T. M Ostrom, ‘Computer simulation: The third symbol system’, *Journal of Experimental Social Psychology*, **24**(5), 381 – 392, (1988).

[43] F. Peinado, V. Francisco, R. Hervás, and P. Gervás, ‘Assessing the novelty of computer-generated narratives using empirical metrics’, *Minds and Machines*, **20**(4), 588, (10/2010 2010).

[44] R. Prez y Prez, *MEXICA: A Computer Model of Creativity in Writing*, Ph.D. dissertation, The University of Sussex, 1999.

[45] Vladimir Propp, *Morphology of the Folk Tale*, Akademija, Leningrad, 1928.

[46] E. Reiter and R. Dale, *Building Natural Language Generation Systems*, Cambridge University Press, New York, NY, USA, 2000.

[47] M. Riedl and C. León, ‘Toward vignette-based story generation for drama management systems’, in *Workshop on Integrating Technologies for Interactive Stories - 2nd International Conference on Intelligent Technologies for Interactive Entertainment*, (2008).

[48] M. Riedl and R. M. Young, ‘Narrative planning: Balancing plot and character’, *J. Artif. Intell. Res. (JAIR)*, **39**, 217–268, (2010).

[49] R. Schank and R. Abelson, *Scripts, Plans, Goals and Understanding: an Inquiry into Human Knowledge Structures*, L. Erlbaum, Hillsdale, NJ, 1977.

[50] M. Sharples, ‘An account of writing as creative design’, in *The Science of Writing: Theories, Methods, Individual Differences, and Applications*, eds., C. M. Levy and S. Ransdell. Lawrence Erlbaum Associates, (1996).

[51] A. Tapscott, J. Gómez, C. León, J. Smailovic, M. Znidarsic, and P. Gervás, ‘Empirical Evidence of the Limits of Automatic Assessment of Fictional Ideation’, in *5th International Workshop on Computational Creativity, Concept Invention, and General Intelligence, C3GI at ESSLLI 2016*, Bozen-Bolzano, Italy, (08/2016 2016).

[52] M. Theune, E. Faas, A. Nijholt, and D. Heylen, ‘The Virtual Storyteller: Story creation by intelligent agents’, in *Proceedings of the Technologies for Interactive Digital Storytelling and Entertainment (TIDSE) Conference*, pp. 204–215, (2003).

[53] T. Trabasso, P. vand den Broek, and S.Y. Suh, ‘Logical necessity and transitivity of causal relations in stories’, *Discourse Processes*, **12**, 1–25, (1989).

[54] S. R. Turner, *Minstrel: a computer model of creativity and storytelling*, Ph.D. dissertation, University of California at Los Angeles, Los Angeles, CA, USA, 1993.

[55] T. A. van Dijk and W. Kintsch, *Strategies of Discourse Comprehension*, New York: Academic Press, 1983.

A Further Storification Examples

Further examples of storification by the StoryFire system are shown in Figures 6 and 7. The system does produce instances of a number of additional plot frames. However, most of these are longer than the examples shown, which makes it impractical to include instances of them in a paper of this length.

The hero in the Tragedy in Figure 6 often absents himself from the story. This is acceptable because the protagonist of the tragedy is the villain. The punishment meted to the villain appears to be one of banishment.

The Comedy in Figure 7 involves some of the same characters in the tragedy in Figure 6, but storyfied differently (with a different plot frame and a different selection of events and characters). It also provides a different instantiation of the plot frame used in Figure 5. The movement of the pieces in Figure 6 seems to scenify the sundering of the lovers, and a conference between the suitor and the guardian.

In the example in Figure 5, the guardian seems to interpose himself between the lovers in the second frame of the story, and the suitor jumps over the guardian to stand next to his lover in the last frame.

All these surprising interpretations of the actual movements by the pieces arise in the present version by serendipity. The possibility of examining these serendipitous behaviours to be incorporated as features of an improved system will be considered as further work.

Move: 23

	a	b	c	d	e	f	g	h
1	■	■	■	■	■	■	■	■
2	■	■	■	■	■	■	■	■
3	■	■	■	■	■	■	■	■
4	■	■	■	■	■	■	■	■
5	■	■	■	■	■	■	■	■
6		**P**				■	■	■
7	P		B			■	■	■
8	R		**B**	Q		■	■	■

0 character lwb
0 kidnap lwb wp2
0 character wp2
0 misbehaved lwb
0 abductor lwb
0 abducted wp2

Move: 25

	a	b	c	d	e	f	g	h
1	■	■	■	■	■	■	■	■
2	■	■	■	■	■	■	■	■
3	■	■	■	■	■	■	■	■
4	■	■	■	■	■	■	■	■
5			P	P	■	■	■	■
6		**P**			■	■	■	■
7	P	**B**	B		■	■	■	■
8	**R**			**Q**	■	■	■	■

1 character lwr
1 character wq
1 orders wq lwr
1 called lwr

Move: 29

	a	b	c	d	e	f	g	h
1	■	■	■	■	■	■	■	■
2	■	■	■	■	■	■	■	■
3	■	■	■	■	■	■	■	■
4	■	■	■	■	■	■	■	■
5			P	P	■	■	■	■
6		**P**		**Q**	■	■	■	■
7	P	**B**	B		■	■	■	■
8					■	■	■	■

2 sets_out lwr
2 traveller lwr

Move: 31

	a	b	c	d	e	f	g	h
1	■	■	■	■	■	■	■	■
2	■	■	■	■	■	■	■	■
3	■	■	■	■	■	■	■	■
4	■	■	■	■	■	■	■	■
5	■	■	■	■	■	■	■	■
6		**P**		**Q**		■	■	■
7	P		B			■	■	■
8			**B**		**R**	■	■	■

3 fight lwr lwb
3 confrontation lwr lwb
3 enemies lwr lwb
3 attacker lwr
3 defender lwb

Move: 33

	a	b	c	d	e	f	g	h
1	■	■	■	■	■	■	■	■
2	■	■	■	■	■	■	■	■
3	■	■	■	■	■	■	■	■
4	■	■	■	■	■	■	■	■
5	■	■	■	■	■	■	■	■
6		**P**		**Q**		■	■	■
7	P		B			■	■	■
8			**B**			■	■	■

4 wins lwr
4 winner lwr

4 looser lwb

Move: 37

	a	b	c	d	e	f	g	h
1	■	■	■	■	■	r		
2	■	■	■	■	■	p	k	
3	■	■	■	■	■	n	p	**B**
4	■	■	■	■	■			
5	■	■	■	■	■			
6	■	■	■	■	■	■	■	■
7	■	■	■	■	■	■	■	■
8	■	■	■	■	■	■	■	■

5 punished lwb

Table 6. Storification as a Tragedy of the thread for the left white bishop (lwb, represented in the diagrams as **B**) who kidnaps the second white pawn (wp2, represented in the diagrams as **P**) and is finally defeated by the left white rook (lwr, represented in the diagrams as **R**) sent to the rescue by the white queen (wq, represented in the diagrams as **Q**).

Move: 27

	a	b	c	d	e	f	g	h
1	■	■	■	■	■	■	■	■
2	■	■	■	■	■	■	■	■
3	■	■	■	■	■	■	■	■
4	■						■	■
5	■		P	P			■	■
6	■	P		**Q**		**N**	■	■
7	■	B	B			P	■	■
8	■					R	■	■

0 character wq
0 character rwk
0 mutual_love wq rwk
0 want_to_marry wq rwk
0 lover wq
0 beloved rwk

Move: 35

	a	b	c	d	e	f	g	h
1	■	■	■	■	■	■	■	■
2	■	■	■	■	■	■	■	■
3	■	■	■	■	■	■	■	■
4	■				**N**		■	■
5	■		P	P			■	■
6	■	P		**Q**			■	■
7	■		B			P	■	■
8	■		B			R	■	■

(1 character bp3)
(1 guardian bp3 rwk)
1 opposed_to_plan bp3
1 sundered wq rwk

Move: 36

	a	b	c	d	e	f	g	h
1	■	■	■	■	■	■	■	■
2	■	■	■	■	■	■	■	■
3	■	■	■	■	■	■	■	■
4	■		**p**		**N**		■	■
5	■		P	P			■	■
6	■	P		**Q**			■	■
7	■		B			P	■	■
8	■		B			R	■	■

(2 different_class wq rwk)
2 high_status_revealed rwk
2 ¬ sundered wq rwk

Move: 39

	a	b	c	d	e	f	g	h
1	■	■	■	■	■	■	■	■
2	■	■	■	■	■	■	■	■
3	■	■	■	■	■	■	■	■
4	■		**p**				■	■
5	■		P	P			■	■
6	■	P		**Q**			■	■
7	■		B			P	■	■
8	■					R	■	■

3 marry wq rwk

Table 7. Storification as a Comedy of the thread for the white queen (wq, represented in the diagrams as **Q**) in terms of her romance with the right white knight (rwk, represented in the diagrams as **N**) in the face of opposition of her guardian the third black pawn (bp3, represented in the diagrams as **p**).

Creativity vs quality: why the distinction matters when evaluating computational creativity systems

Anna Jordanous[1]

Abstract. The evaluation of computational creativity systems is increasingly becoming part of standard practice in computational creativity research, particularly with recent development in evaluation tools. One matter that can cause confusion, however, is in distinguishing between the concepts of creativity and quality/value. These two concepts are highly interrelated, to the point that it is difficult (and perhaps inappropriate) to define creativity without incorporating quality judgements into that definition. Several examples exist, however, where creativity evaluation has been confused with quality judgments, leading to less grounded evaluative results. Many computational creativity projects aim to produce high quality results; this is a worthy research aim. If, however, the aim of a computational creativity research project is to make as *creative* a system as possible, then a more careful approach is needed that acknowledges and understands the differences - and also the overlaps - between creativity and quality. This paper critically investigates the concepts of creativity and quality (and how they are related). It offers warning examples showing the dangers of conflating the two concepts. These are followed by practical examples of how to incorporate value judgements into the evaluation of creativity of software, to further our overall pursuit of building more creative computational systems.

1 introduction

There is a distinction to be drawn between the aim of evaluating creativity or evaluating quality (Section 3); as the survey of evaluative practice in Section 3.1 showed, these aims have become blurred to some extent.

How does one evaluate the creativity of a computer system? It can be attractive to sidestep this issue somewhat, evaluating the *quality* of a system's output rather than *creativity*.[2] As a result, though, system development progresses towards more successful output, but not necessarily more creative output. Perhaps this is what is desired? But **if** the aim of a computational creativity research project is to make a *creative* system, then a more careful approach is needed that acknowledges and understands the differences (and the overlaps) between creativity and quality.

This paper specifically tackles the above aim, investigating at how researchers can navigate the distinction between creativity and quality, in pursuit of building more creative computational systems.

Section 2 investigates the meaning of 'quality', the meaning of 'creativity' and the ways in which these two concepts are interconnected. These investigations are carried out looking at both human and computational creativity. Section 3 tackles the question of how these concepts should be handled when performing evaluation of computational creativity systems, reflecting on relevant 'good, bad and ugly' previous evaluative practice. The paper concludes with the take-home message that while quality forms a key part of creativity, it is not a direct substitute. To evaluate creativity, we must incorporate evaluation of quality, but not treat evaluation of quality as a sufficient proxy for evaluation of creativity.

2 Quality, creativity and the connections between them

2.1 Quality is...

The concept of quality, as treated in this paper, is highly related to its synonymous concepts of value as well as its near-synonyms of utility, usefulness, appropriateness, correctness, fit, relevance, and effectiveness. During this paper I will occasionally use quality and value interchangeably to represent the overarching concept of something having some worth.

Quality judgements represent the value that something has to at least one observer. As in [1], the word 'value' can be treated either as a noun or as a verb, i.e. the action of valuing; and value is not restricted to commercial or quantitative measurements, but also to cultural and qualitative assessments of value. Such judgements can be affected by societal contexts and influences, as discussed for example by Wiggins et al. in their considerations of how value is manifested in creative contexts:

> 'we treat value as a relation between an artefact, its creator and its observers and the context in which creation and observation take place.' [38, p. 2]

It is difficult to find domain-independent heuristics to follow when ascertaining the value of products. Usefulness is relative; what is considered useful in products of one domain is not necessarily reproduced in the other and may not apply equally across that individual domain. Wiggins et al [38] note in particular how Western perspectives may differ from non-Western perspectives.

To recognise the usefulness of a creative product, one might know the product's domain well enough to appreciate value, or one might have access to the opinions of people who are experts in that domain. To exemplify this point, and to begin to link quality with creativity: the greatest contributor to creativity in musical improvisation has been found to be the social communication and interaction that happens between musicians, or between performer(s) and audience during the creative process of improvising [10, 7]. Specifically for creativity, improvisers prioritise this over the quality or 'correctness' of the music produced during improvisation. In mathematical proof

[1] University of Kent, Medway, UK, email: A.K.Jordanous@kent.ac.uk

[2] The distinction between the evaluative aims of creativity and quality is raised in Section 3.1.

derivation systems, however, accuracy (and hence quality, which is strongly related to accuracy in this domain) is vital.

Zongker's paper entitled *Chicken Chicken Chicken: Chicken Chicken* [39] demonstrates how the perception of quality in a particular domain is not always consistent across all examples of creativity in a domain. *Chicken Chicken Chicken* shows quality, in a domain that emphasises content correctness (scientific research papers), because of the extreme absence of any scientifically useful and correct content. Instead the quality of *Chicken Chicken Chicken: Chicken Chicken* comes from its value as an ironic and humorous reflection on academic publications.

2.1.1 Evaluating quality achieved by computational creativity systems

Various approaches have been used to evaluate quality; ranging from relatively simple quantitative metrics of the validity or correctness of products (as discussed below), to those more tricky evaluative scenarios when qualitative, multiple, complex or non-objective metrics are required to judge quality, as discussed for example in the [9] assessment of the cultural value of electronic musicians' creative work.

In a 2011 survey of evaluation practice in computational creativity [6, 7] (see also Section 3.1, evaluations of quality of the surveyed systems[3] were typically based on aspects of the end product(s) rather than any of the other Four Ps: process, person/producer or press (see [8]). While many examples were found of empirical measurements of value or quality, as described below, several systems were assessed for quality through user evaluations. Evaluation data was either directly provided by the user or provided indirectly through studies, such as through audience reactions and feedback at exhibitions or through qualitative tests with target users for usability and effectiveness of the system. Feedback about the appeal of systems' products and personal preferences about the products was also provided through user evaluations.

Many systems were evaluated by the correctness and validity of their products, such as calculating the percentage of material produced during runtime that can actually be used, or statistical tests for validity. Some systems were measured in terms of how interesting or novel their products were, for example seeing if the products performed at a level above a given threshold for novelty and originality in the Wundt curve function [29] or using variables representing domain-specific interest or complexity measurements.

The usefulness of a system's products could also be quantified, through the percentage of a user query which is satisfied by system output [22], or the percentage of results that are valid. Human ratings of usefulness were also used. Usefulness ratings were not all quantitative, with use of post-implementation discussions on usefulness or the interpretation of value as serving an intended purpose. Other definitions of quality were highly tuned to domain-specific metrics for value, making them less generally applicable across several types of creative system or for a more general discussion.

2.2 Creativity is...

It is difficult to define creativity without bringing quality into the definition: the concept of quality is heavily used when defining what creativity is. Psychology research has settled on a slightly controversial but now fairly commonly accepted 'standard definition of creativity' [28]:

> 'The standard definition is bipartite: Creativity requires both originality and effectiveness. ... Originality is vital for creativity but is not sufficient. ... Original things must be effective to be creative. ' [28, p. 92]

Here the word 'effectiveness' is used to represent the concept referred to in this paper as quality or value, as explained by Runco and Jaeger during the discussions in [28].

Prior to the publication of [28], the *quality* (and related concepts: value usefulness, appropriateness, relevance) and *novelty* (and related concepts: originality, newness) of creative products have often been identified as the two main aspects of creativity. Creativity was being defined in computational creativity research as 'how to create something new and useful at the same time.' [21, p. 290] Similar definitions were widely adopted e.g. in [20, 21, 27] in computational creativity, and e.g. [14, 30, 3, 23] in psychological research into creativity. Mayer [14] refers to this combination as the 'basic definition of creativity' [14, p. 450]. Table 22.1 of [14], reproduced here in Table 1, summarises the 'Two Defining Features of Creativity' [14, p. 450] as used in [30].

In a 2004[4] survey of 34 definitions of creativity used in creativity research [24], the survey found that:

> 'The most common characteristics of explicit definitions were uniqueness (n = 24) and usefulness (n = 17). Of interest, all 17 articles that included usefulness in their definition also mentioned uniqueness or novelty.' [24, p. 88]

Table 1. Mayer's summary of how novelty and value (or highly related concepts) are used to define creativity by different authors in various chapters of Robert J. Sternberg's influential *Handbook of Creativity* [14, (Table 22.1, p. 450)], in [30].

Author (Chapter)	Feature 1: Originality	Feature 2: Usefulness
Gruber & Wallace (5)	novelty	value
Martindale (7)	original	appropriate
Lumsden (8)	new	significant
Feist (13)	novel	adaptive
Lubart (16)	novel	appropriate
Boden (17)	novel	valuable
Nickerson (19)	novelty	utility

It is questionable whether the combination of novelty and value is enough to understand creativity [12]. This reductionist approach provides two tangible attributes with which to evaluate creativity. Work in computational creativity has produced countless systems that produce novel results that have value; but still the notion that computers can be creative is resisted. This undefinable part of creativity is reflected in Weiley's coining of creativity as 'novelty, value and "x" ' [32]. As argued in [11], there is much more to consider in terms of what creativity is, that the combination of novelty and value alone does not incorporate.

In dictionary definitions of creativity, the word 'quality' is one of the more frequent words used, as is 'new' (excluding common-use English words such as 'the', 'and', and so on) ([7], see also the word cloud in Figure 1). However this word cloud reveals many other

[3] Here I consider quality of a system to be treated pragmatically based on how it performs, but acknowledge that software quality in its own right, encompassing software engineering and code quality, would also be an alternative interpretation of the title of this section.

[4] This 2004 survey by [24] predates the above-mentioned work by Runco et al [28] defining their 'standard definition' of creativity.

words relating to creativity other than 'quality' and 'new'. Jordanous and Keller [11] empirically identified 14 key components of creativity through the analysis of multi-disciplinary discussions of the nature of creativity. These components do include Value, as well as Originality, but also components such as Active Involvement & Persistence, or Spontaneity & Subconscious Processing. Nonetheless, it is almost universally agreed that the concept of quality, incorporating the notions of value and usefulness, is a necessary component of creativity.

Before concluding this section, I briefly acknowledge an incidental point that connects quality and creativity in the scientific study of creativity (of which computational creativity forms a part). An interesting subjective objection to the scientific study of creativity is whether it may have a detrimental effect on our sense of the 'marvelling', 'awe and delight' of creativity:[5]

> 'Forget computers, for the moment: the conviction is that *any* scientific account of creativity would lessen it irredeemably. ... [There is a] widespread feeling that science, in general, drives out wonder. Wonder is intimately connected with creativity. All creative ideas, by definition, are valued in some way. Many make us gasp with awe and delight. ... To stop us marvelling at the creativity of Bach, Newton, or Shakespeare would be almost as bad as denying it altogether. Many people, then, regard the scientific understanding of creativity more as a threat than a promise.' [3, pp. 277-278]

3 Evaluative aims: creativity or quality?

An issue that researchers often face when evaluating their system is:

> Should systems be evaluated solely on the value and correctness of their output, or should there be some assessment of the creativity demonstrated by the system (which incorporates quality judgements on the output)?

Both are important, though the quality of output is often easier to define and test for, especially in the absence of a standard definition or creativity evaluation methodology. Particularly for computational *creativity* research, though, it is as important to consider to what extent a computational creativity system can actually be considered creative [7, 5].

It is becoming easier for computational creativity researchers to specifically target evaluation of the creativity of their systems, due to the development of evaluation tools. Creativity evaluation methods such as SPECS [7], Creative Tripod [5] or Ritchie's criteria [27] are starting to become more widely used in practice in computational creativity.

No one methodology has yet been adopted as standard, however. Historically, a 2011 survey of practice in computational creativity evaluation [6, 7] revealed issues in conflating judgements of creativity and quality during evaluation which did not follow these evaluation methods. This survey, of which the most relevant parts are reported below, investigated various questions about evaluation practice of creative systems, including these questions:[6]

- Evaluation details:
  - Is system evaluation mentioned at all in the paper?
  - Has a system evaluation been performed and described in the paper?
  - Do the authors state the aims of their evaluation and/or their evaluative criteria?
  - Is the main aim of evaluation to assess creativity (including quality of output/system) or (just) quality of output/system?
  - Brief description of evaluation done.

From the 75 surveyed creative systems in Section 3.1, only 35% of systems were evaluated according to how creative they were; the rest of the systems were evaluated solely by the quality of the system's performance. Two systems [25, 4] were described as being assessed for creativity but were actually assessed only on the accuracy of the system.

Of the 18 papers making practical use of creativity evaluation methodologies such as Ritchie's criteria [27] or Colton's Creative Tripod [5], only 10 papers used the methodologies for creativity evaluation, with the rest adapting the methodologies to evaluate the quality of their system output.

This shows some confusion about the distinction between creativity and quality; as this paper investigates, our interpretation of creativity includes reflections on quality but encapsulates more than just how correct or valuable the creative output is. A pertinent example of such confusion can be found in [31]: Ventura aimed to critically analyse creativity evaluation methodologies via a thought experiment. but actually addressed quality (or 'recognisability') evaluation only.

3.1 Survey findings

75 systems were reviewed during this comprehensive survey of computational creativity literature at the time. Results relevant to this paper are summarised in [6, 7]. Looking at the 75 surveyed systems for information relevant to this current paper:

- Of the 75 programs presented as creative systems, 26 systems (35%) were critically discussed in terms of how *creative* they were.
- 32 systems (43%) were evaluated based on the *quality* or accuracy of system performance compared to a human performing that task. This set of 32 systems includes 3 systems which were described as being assessed on how creative the systems were, but which were actually assessed by the quality of the system's performance.
- 1 paper evaluated its system in terms of knowledge gained for future research.
- The remaining 16 papers did not include evaluation of the system.

The survey also revealed interesting details from 18 papers that applied recognised creativity evaluation methodologies [5, 27, 2] or creativity models [3, 36], or that proposed new metrics to evaluate their creative systems. Of the 18 papers that applied recognised creativity evaluation methodologies:

- 10 papers used the methodologies to measure how creative their systems were.
- 6 papers adapted the chosen method to measure the quality of the systems.
- 2 papers used 'creativity' methodologies that actually measured quality.

[5] Perhaps appropriately, it is difficult to debate this point scientifically but it is worth being aware of this point, as scientific researchers studying creativity

[6] For the full list of questions included in the survey, see [7].

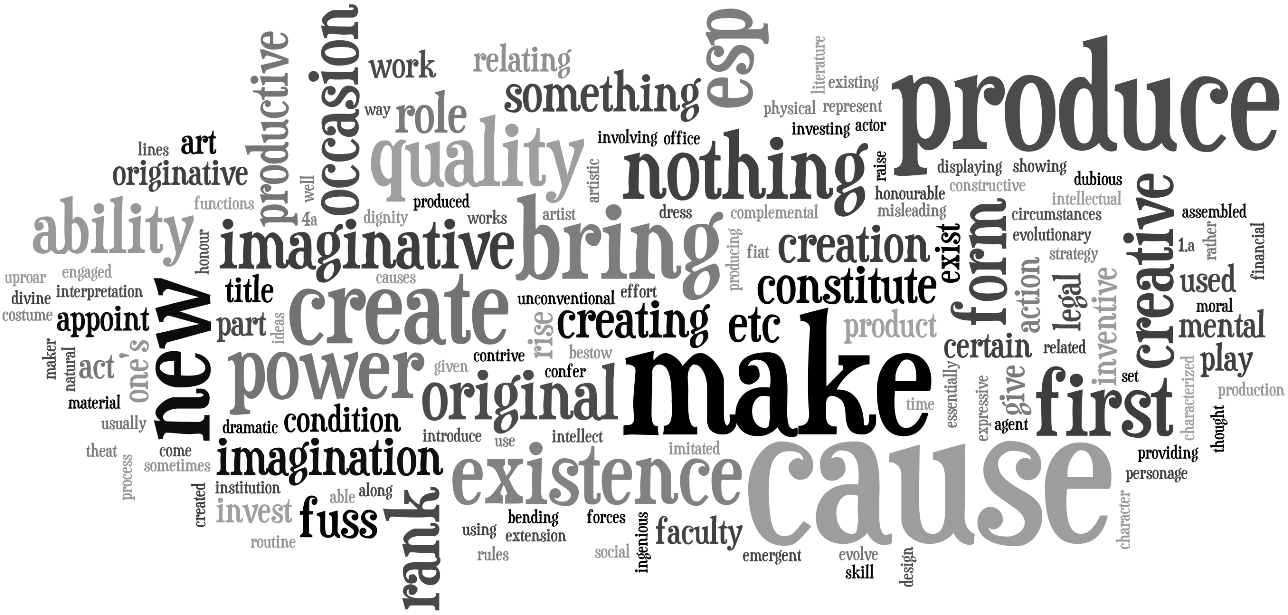


Figure 1. Words used in various dictionary definitions of creativity, as analysed in [7]. The font size of a word is relative to the frequency with which that word occurs in the collection of dictionary definitions; the larger the word, the more it appears. The word 'quality' appears fairly prominently, but does not overly dominate the diagram.

4 Using quality judgements as part of an evaluation of computational creativity

Often, as seen in the above-mentioned survey, creative systems have often been evaluated with regard to the quality of the output and this has been used to justify that system being described as creative by the authors. The discussions below first look at exemplar scenarios where such evaluation has been done due to a confusion between the two concepts of quality and creativity, then turn to discussing how quality judgements have been consciously and justifiably incorporated into evaluation of creativity.

4.1 Confusion between creativity and quality in computational creativity evaluation

The overlap between creativity and quality can sometimes cause confusion as to how to evaluate creativity; this is perhaps unsurprising given the many issues manifest in evaluating creativity of computational systems [7]. One representative example, taken from the systems surveyed above, sees a so-called creativity metric proposed which actually evaluates quality, but which is used to generate evidence justifying a system being labelled a creative system.

Collins et al. [4] employ the Wilcoxon's two-sample statistical test on their music harmonisation generator. The metric examines similarity between generated output and the system's knowledge base. Despite Collins et al. describing their test as a creativity metric, they actually measure how closely the system can replicate the test set (similar to the approach in [34]). In other words, in [4] what is actually proposed is a correctness metric rather than a creativity metric, a distinction which they briefly acknowledge as they admit a lack of conviction in describing their system as creative:

> 'This paper has presented a metric for evaluating the creativity of a music-generating system. Until further evaluation has been conducted (by human listeners rather than just by the creativity metric), we are cautious about labelling our overall system as creative.' [4, p. 9]

Another example of this confusion is clear when we ask people to evaluate the creativity of computational systems; expressions of the difficulty of this task is often acknowledged by recognition that the evaluators do not know where a creativity judgement is distinct from a value judgement. For example, in the case studies reported in [7] where several systems were evaluated on how creative they were, quotes from respondents included:

> 'I liked this one better than the other ones, but am really struggling to distinguish between "like" or "approve" and "think it's creative".'
>
> 'I kept going with my gut instinct which was basically to rate it on how much I *liked* it... but I don't think that really equates to how creative it was... but I'm not even sure a computer *can* be creative, which is why I had to just keep reverting to "like".'
>
> 'I think it depends on the definition of creativity - is it just creating something? or creating something that makes the appropriate amount of "sense", for want of a better word, for people to appreciate? I'm using the latter definition!'

There are deeper issues afoot here than mere confusion, to do with people's perception of computational creativity.[7] One evaluator in the [7] evaluation case studies explained how they struggled with applying the concept of creativity to computers when they saw creativity as 'a uniquely human thing'. Thus, they instead resorted to a conceptually easier measure of aesthetic, even though they were aware of the difference, concluding:

> 'I preferred these samples to the previous ones, but that isn't really a measure of creativity!'

[7] This thorny topic will not be explored in this current paper, but is explored to some extent in [7] as well as in [15, 13, 16].

4.2 Conscious incorporation of quality evaluation in creativity evaluation

The blurring of evaluative aims, between assessing quality and creativity, is a theme that is detectable not only in [4], but also in several computational creativity system evaluations.[8] Several evaluation tools in computational creativity, however, consciously include quality evaluation as part of a creativity evaluation for creative systems. This is in keeping with the view that quality or value is a fundamental component of creativity [28, 11].

4.2.1 Ritchie's empirical criteria for computational creativity

Graeme Ritchie proposed a set of formal empirical criteria for creativity [27]. The criteria are situated in an overall framework describing the design and implementation of a creative computational system in set-theoretic form. Ritchie advocates post-hoc analysis of artefacts generated by the system, disregarding the process by which they were created. For systems that produce abstract rather than concrete results (Ritchie gives the example of analogies), Ritchie's approach is not applicable.

The criteria collectively describe aspects of the *typicality* and *quality* of the output of the creative system (and indirectly, the novelty of the system output). Two key mappings are used to separate out the concepts of typicality and novelty:

typ - a rating of how typical the output is in the intended domain

> 'To what extent is the produced item an example of the artefact class in question?' [27, p. 73]

val - a rating of how valuable the output is

> 'To what extent is the produced item a high quality example of its genre?' [27, p. 73]

Ritchie emphasises the importance of assessing computer-generated artefacts both in terms of how typical an example they are of items in the target domain and in terms of atypicality. Further to this, an artefact may be typical of the domain but not be a good example, so the value rating is introduced to assess the quality of that artefact.

> 'If a person produces a painting which is radically different from previous work ... and which is definitely a good painting, then that will usually be deemed creative. ' [26, p. 4]

The formal definitions of the 18 criteria can be found in [27]. Here, the criteria are deliberately presented informally, with descriptors such as 'suitable' and 'high' substituted for the parameters left unspecified in [27]. It is hoped that any subsequent loss in formal semantics is balanced by a more immediate understanding of each criterion. We can see that while Ritchie allows for both typical and atypical results to be recognised by the criteria, criteria involving value judgements are always required to find high levels of value if that criterion is to be satisfied.

1. On average, the system should produce suitably typical output.
2. A decent proportion of the output should be suitably typical.
3. On average, the system should produce highly valued output.
4. A decent proportion of the output should be highly valued.
5. A decent proportion of the output should be both suitably typical and highly valued.
6. A decent proportion of the output is suitably atypical and highly valued.
7. A decent proportion of the atypical output is highly valued.
8. A decent proportion of the valuable output is suitably atypical.
9. The system can replicate many of the example artefacts that guided construction of the system (the *inspiring set*).
10. Much of the output of the system is not in the inspiring set, so is novel to the system.
11. Novel output of the system (i.e. not in the inspiring set) should be suitably typical.
12. Novel output of the system (i.e. not in the inspiring set) should be highly valued.
13. A decent proportion of the output should be suitably typical items that are novel.
14. A decent proportion of the output should be highly valued items that are novel.
15. A decent proportion of the novel output of the system should be suitably typical.
16. A decent proportion of the novel output of the system should be highly valued.
17. A decent proportion of the novel output of the system should be suitably typical and highly valued.
18. A decent proportion of the novel output of the system should be suitably atypical and highly valued.

4.2.2 Pease et al.'s tests on the input, output and process of a system

[20] proposed a combination of evaluative tests for creativity, based on:

- The input provided to the system.
- The output produced by the system.
- The process(es) employed by the computational system.

The tests for the output produced by the system are categorised by Pease et al. as either *Novelty Measures* or *Quality Measures*. The latter set of quality measures consists of:

- Quality Measures.
  - *Emotional Response Measure:* human judges evaluate to what degree an item has affected them positively or negatively; the responses are used to categorise items according to the intensity of the response.
  - *Pragmatic Measure:* using unspecified (domain-specific) 'marking criteria' [20, p. 6] to judge to what extent an item meets an aim.

The tests for the process(es) employed by the computational system are divided into two sets: tests of generative processes and tests of evaluative processes within the systems. The evaluation set of tests includes a test based around a quality judgement:

- *Evaluation of Process Measure:* comparing the quality measures from above on two comparable sets of output items. One set is produced by methods which can be transformed internally during program run-time and one by methods which cannot. The quality of the first set should exceed the quality of the second.

[8] In fact, encountering such examples during peer review duties was one of the prompts to write this paper.

As in Section 4.2.1, these tests are summarised in informal language above.[9]

4.2.3 *Wiggins' framework for categorising creative systems*

Wiggins proposed a framework for categorising creative systems [37] inspired by Boden's proposals on creativity [3]. Strictly speaking, this framework is for formal description and classification of different aspects of creative system, rather than evaluation of creativity, but has been used in computational creativity evaluation.

Though the framework is intended to be used to 'analyse, evaluate and compare creative systems' [36, p. 1], Wiggins carefully states that he does not contribute to the debate on creative evaluation:

> 'I am making no attempt here to discuss or assess the value of any concepts discovered: while this issue is clearly fundamentally important [citing [3, 26, 17]], it can safely be left for another time.' [37, p. 453]

The framework describes system details formally according to seven formal rule sets and functions relating to the system's conceptual space (i.e. the set of all possible items that could conceivably be output by the system). One of these rulesets, $\mathcal{E}$, is the set of rules used to evaluate items in the conceptual space. This set, as with the others, is left to be populated as the framework is applied for categorising creative systems.

Looking at Wiggins' immediate subsequent work as a guide for how to populate this set, Wiggins and colleagues have tended to focus on quality evaluation rather than creativity evaluation [34, 19, 33]. In particular [19]'s melody generation system was evaluated using a variation of Amabile's Consensual Assessment Technique (CAT) [2], intended by Amabile for evaluating the creativity demonstrated by humans in a quantitative way by expert judges. CAT was adapted slightly in [19] to assess the quality of output ('stylistic success') from their system rather than the creativity of the system itself. This decision was perhaps influenced by the authors' substantial background in musical quality evaluation e.g. [35, 17, 36, 18, 37].

5 CONCLUSION

Distinguishing between creativity and quality is a tricky task to negotiate when we are evaluating the creativity of computational creativity systems. The two concepts overlap considerably; in fact it is generally accepted that creativity cannot be defined without incorporating the concept of quality into that definition. The two concepts are however not to be confused; an evaluation of value is distinct from an evaluation of creativity.

Above, examples have been presented where such confusion in evaluation has led to less-than-solid conclusions about the results of such evaluation. We have also seen, however, that creativity evaluation can (and should) incorporate evaluation of quality as part of that overall evaluation. Quality is a necessary part of creativity; but it is not sufficient for creativity.

ACKNOWLEDGEMENTS

These thoughts have been shaped over conversations with several people over the past years, most recently a discussion point raised by Oded Ben-Tal at the 2017 Computational Simulation of Musical Creativity conference on whether creative music was the same as high quality music.

REFERENCES

[1] Daniel Allington, Byron Dueck, and Anna Jordanous. Networks of Value in Electronic Dance Music: SoundCloud, London, and the Importance of Place, 2015.

[2] Teresa M Amabile, *Creativity in context*, Westview Press, Boulder, Colorado, 1996.

[3] Margaret A. Boden, *The creative mind: Myths and mechanisms*, Routledge, London, UK, 2nd edn., 2004.

[4] Tom Collins, Robin Laney, Alistair Willis, and Paul Garthwaite, 'Music: Patterns and Harmony Using Discovered, Polyphonic Patterns to Filter Computer-generated Music', in *Proceedings of the International Conference on Computational Creativity*, pp. 1–10, Lisbon, Portugal, (2010).

[5] Simon Colton, 'Creativity versus the Perception of Creativity in Computational Systems', in *Proceedings of AAAI Spring Symposium on Creative Intelligent Systems*, pp. 14–20, Stanford. CA, (2008). AAAI Press.

[6] Anna Jordanous, 'Evaluating Evaluation: Assessing Progress in Computational Creativity Research', in *Proceedings of the Second International Conference on Computational Creativity (ICCC-11)*, Mexico City, Mexico, (2011).

[7] Anna Jordanous, *Evaluating Computational Creativity: A Standardised Procedure for Evaluating Creative Systems and its Application*, Ph.D. dissertation, University of Sussex, Brighton, UK, sep 2012.

[8] Anna Jordanous, 'Four PPPPerspectives on computational creativity in theory and in practice', *Connection Science*, **28**(2), 194–216, (2016).

[9] Anna Jordanous, Daniel Allington, and Byron Dueck, 'Measuring cultural value using social network analysis: A case study on valuing electronic musicians.', in *ICCC*, pp. 110–117, Park City, UT, (2015).

[10] Anna Jordanous and Bill Keller, 'What makes musical improvisation creative?', *Journal of Interdisciplinary Music Studies*, **6**(2), 151–175, (2012).

[11] Anna Jordanous and Bill Keller, 'Modelling creativity: identifying key components through a corpus-based approach', *PloS ONE*, **11**(10), e0162959, (2016).

[12] James C Kaufman, *Creativity 101*, The Psych 101 series, Springer, New York, 2009.

[13] Carolyn Lamb, Daniel G Brown, and Charles LA Clarke, 'Human competence in creativity evaluation.', in *ICCC*, pp. 102–109, (2015).

[14] Richard E Mayer, 'Fifty Years of Creativity Research', in *Handbook of Creativity*, ed., Robert J Sternberg, chapter 22, 449–460, Cambridge University Press, Cambridge, UK, (1999).

[15] David C Moffat and Martin Kelly, 'An investigation into people's bias against computational creativity in music composition', in *The Third Joint Workshop on Computational Creativity*, Riva del Garda, Italy, (2006).

[16] Philippe Pasquier, Adam Burnett, Nicolas Gonzalez Thomas, James B Maxwell, Arne Eigenfeldt, and Tom Loughin, 'Investigating listener bias against musical metacreativity', in *Proceedings of the Seventh International Conference on Computational Creativity*, (2016).

[17] Marcus Pearce and Geraint Wiggins, 'Towards a framework for the evaluation of machine compositions', in *Proceedings of the AISB Symposium on AI and Creativity in Arts and Science*, York, UK, (2001).

[18] Marcus T Pearce, *The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition*, Ph.D. dissertation, Department of Computing, City University, London, UK, 2005.

[19] Marcus T Pearce and Geraint A Wiggins. Evaluating Cognitive Models of Musical Composition, 2007.

[20] Alison Pease, Daniel Winterstein, and Simon Colton, 'Evaluating Machine Creativity', in *Proceedings of the ICCBR'01 Workshop on Creative Systems*, pp. 129–137, Vancouver, Canada, (2001).

[21] F Peinado and P Gervas, 'Evaluation of automatic generation of basic stories', *New Generation Computing*, **24**(3), 289–302, (2006).

[22] Francisco C Pereira and Amílcar Cardoso, 'Experiments with free concept generation in Divago', *Knowledge-Based Systems*, **19**(7), 459–470, (2006).

[23] Jonathan A Plucker and Ronald A Beghetto, 'Why Creativity is Domain General, Why it Looks Domain Specific, and why the Distinc-

[9] For more formal descriptions of the tests, the interested reader can refer to [20].

tion Doesn't Matter', in *Creativity: From Potential to Realization*, eds., Robert J Sternberg, Elena L Grigorenko, and Jerome L Singer, chapter 9, 153–167, American Psychological Association, Washington, DC, (2004).

[24] Jonathan A Plucker, Ronald A Beghetto, and Gayle T Dow, 'Why Isn't Creativity More Important to Educational Psychologists? Potentials, Pitfalls, and Future Directions in Creativity Research', *Educational Psychologist*, **39**(2), 83–96, (2004).

[25] Mark O Riedl, 'Vignette-Based Story Planning: Creativity Through Exploration and Retrieval', in *Proceedings of the 5th International Joint Workshop on Computational Creativity*, pp. 41–50, Madrid, Spain, (2008).

[26] Graeme Ritchie, 'Assessing Creativity', in *Proceedings of the AISB Symposium on AI and Creativity in Arts and Science*, pp. 3–11, York, UK, (2001).

[27] Graeme Ritchie, 'Some Empirical Criteria for Attributing Creativity to a Computer Program', *Minds and Machines*, **17**, 67–99, (2007).

[28] Mark A. Runco and Garrett J. Jaeger, 'The standard definition of creativity', *Creativity Research Journal*, **24**(1), 92–96, (2012).

[29] Rob Saunders, Petra Gemeinboeck, Adrian Lombard, Dan Bourke, and Baki Kocaballi, 'Curious Whispers: An Embodied Artificial Creative System', in *Proceedings of the International Conference on Computational Creativity*, pp. 100–109, Lisbon, Portugal, (2010).

[30] *Handbook of Creativity*, ed., Robert J Sternberg, Cambridge University Press, Cambridge, UK, 1999.

[31] D Ventura, 'A Reductio Ad Absurdum Experiment in Sufficiency for Evaluating (Computational) Creative Systems', in *Proceedings of the 5th International Joint Workshop on Computational Creativity*, pp. 11–19, Madrid, Spain, (2008).

[32] Viveka Weiley, 'Remixing realities: distributed studios for collaborative creativity', in *Proceedings of the Seventh ACM Conference on Creativity and Cognition*, pp. 345–346, Berkeley, California, (2009). ACM.

[33] Raymond Whorley, Geraint Wiggins, Christophe Rhodes, and Marcus Pearce, 'Development of Techniques for the Computational Modelling of Harmony', in *Proceedings of the International Conference on Computational Creativity*, pp. 11–15, Lisbon, Portugal, (2010).

[34] Raymond P Whorley, Geraint A Wiggins, and Marcus T Pearce, 'Systematic Evaluation and Improvement of Statistical Models of Harmony', in *Proceedings of the 4th International Joint Workshop on Computational Creativity*, pp. 81–88, London, UK, (2007).

[35] Geraint Wiggins, Eduardo Miranda, Alan Smaill, and Mitch Harris, 'A Framework for the Evaluation of Music Representation Systems', *Computer Music Journal*, **17**(3), 31–42, (1993).

[36] Geraint A Wiggins, 'Categorising creative systems', in *IJCAI*, Acapulco, Mexico, (2003). IJCAI.

[37] Geraint A Wiggins, 'A preliminary framework for description, analysis and comparison of creative systems', *Knowledge-Based Systems*, **19**(7), 449–458, (2006).

[38] Geraint A. Wiggins, Peter Tyack, Constance Scharff, and Martin Rohmeier, 'The evolutionary roots of creativity: mechanisms and motivations', *Philosophical Transactions of the Royal Society B*, **370**, 20140099, (2015).

[39] D Zongker, 'Chicken Chicken Chicken: Chicken Chicken', *Annals of Improbable Research*, **12**(5), 16–21, (2006).

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

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

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

1 INTRODUCTION

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

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

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

1.1 From embodied symbols to abstract ideas

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

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

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

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

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

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

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

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

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

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

2 FROM PICTURES TO BODIES

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

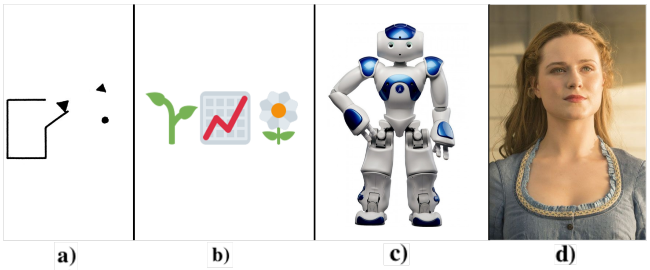


Figure 1. Examples of different renderings to tell a story from most abstract (a) to most distinct embodiment (d): **a)** Geometrical objects from the Heider and Simmel [21] experimental study of apparent behavior. Subjects were found to interpret the animation of these geometrical objects and shapes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

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

3 GESTURES: EXPRESSING THE INTERNAL

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

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

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

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

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

4 OF MEN AND MACHINES

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

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

4.1 Previous Work

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

Most relevant is the approach of Pelachaud *et al.* who designed an expressive gesture model for a storytelling Nao robot [44, 30, 12]. Their approach offers a unified framework that formalizes gestures previously used for a virtual avatar. As such, these gestures are rendered in a Function Markup Language and a Behavior Markup Language. This results in a reusable database of approx. 500 annotated gestures. They use a subset of these for a version of the robot that reads stories to children. Their evaluation in [31] confirms that the gestures are perceived as appropriate to their objectives while scoring poorly for naturalness. They highlight that their approach tries to blend this instantiation of storytelling with a common framework that also allows it to be applied for other robots. This is true for the gesture database, which has been annotated with admirable detail about the gesture space by dissecting each gesture into preparation, stroke and retraction phases. While an adaptation for the Nao robot requires two additional databases that were not available on request, we shall draw as many from insights from this study as we can.

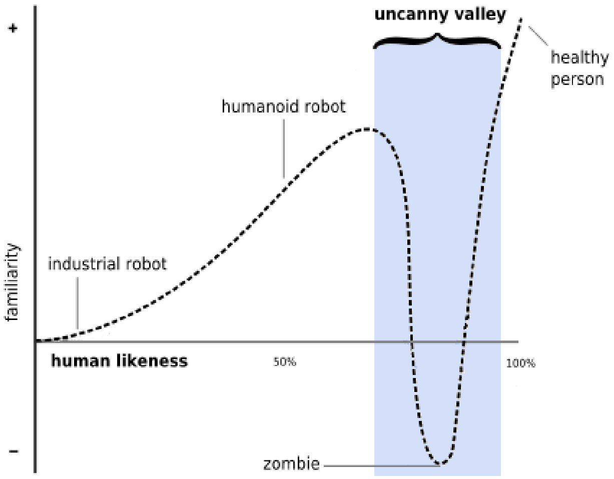


Figure 2. This graph depicts the theoretical perception of familiarity on a scale from industrial robot to healthy person. The area in blue marks the uncanny valley. Adapted from author: Karl MacDorman [33].

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

With respect to multi-modal uses of the Nao, studies by Jokinen, Wilcock et al. [9, 37, 55] are worthy of mention. Their system, which is half question-answering system and half spoken-dialog system, uses Wikipedia as a knowledge source and renders the retrieved content in a conversational manner [55]. In [9] these authors discuss the different modalities of face detection, tactile sensors, non-verbal cues and gestures. They use the Nao's inbuilt face recognition software, as well as sonar sensors and speech direction detection to start the conversation, and empirically determine that the best communication distance is 0.9 meters. They implemented a small set of six gestures to signal discourse-level details, hyperlinks or to manage turn-taking with human interlocutors. Some insights about speech and gesture synchronization are especially noteworthy. For example, their animation software did not accurately reflect the timing of gestures when performed by the actual robot. Each gesture was parametrized using Python code but the Nao's speech recognizer does not allow for a sudden interruption by the user. These authors also split each gesture into preparation, stroke and retraction phases to align the pitch of the spoken sentence with the stroke of the gesture.

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

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

5 THE FRAMEWORK

5.1 Modalities

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

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

5.2 Technical Solutions

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

5.2.1 Speech Recognition

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

5.2.2 Text-To-Speech

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

5.2.3 Creaky Joints

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

5.2.4 *Autonomous Behaviour and Eye Color*

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

5.3 Gestures

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

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

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

6 TELLING AN INTERACTIVE STORY

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

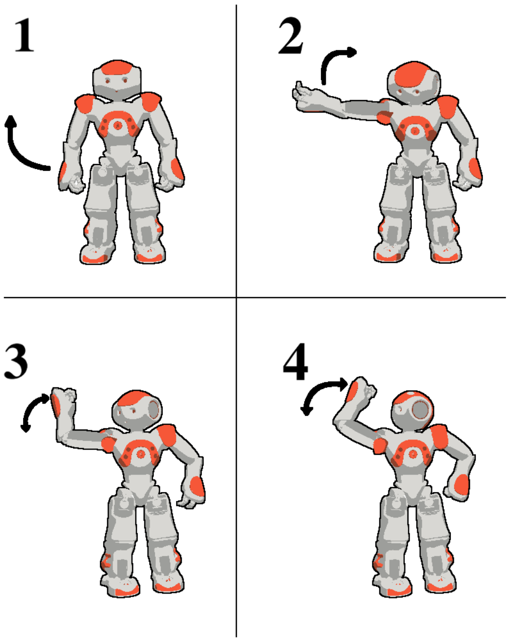


Figure 3. Example of a Nao gesture in four frames. This gesture has been annotated to strongly associate with the action *train*. First frame is the resting position, followed by a raising of the arm in the second frame. The third and fourth frame are alternating a few times. This gesture is a *show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

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

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

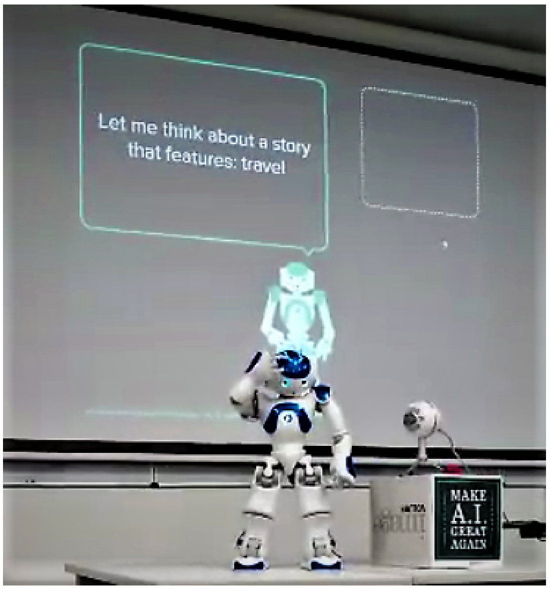


Figure 4. Demonstration of the preliminary storytelling framework at the UCD School of Computer Science Opening Evening 2017.

6.1 Digging for Stories

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

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

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

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

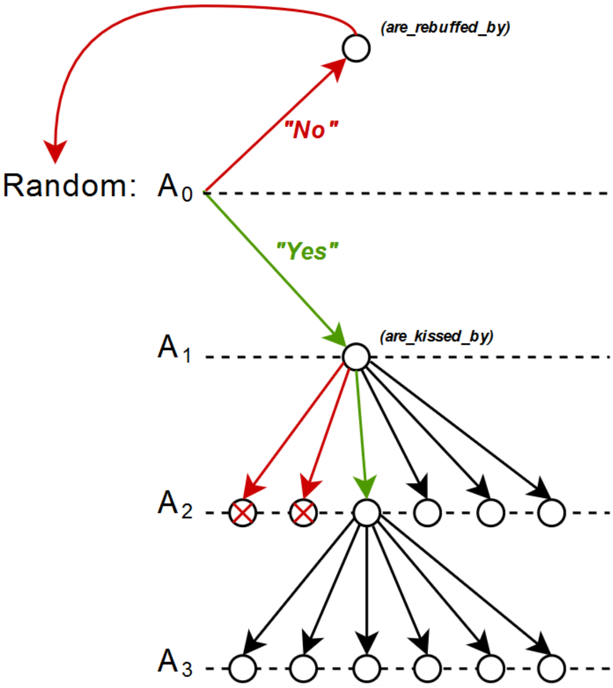


Figure 5. Example of the knowledge acquisition process in a tree diagram. Red arrows indicate a negative response from the user and green a positive. Black arrows have not been evaluated.

6.2 Enactment

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

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

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

7 FUTURE WORK

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

8 CONCLUSION

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

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

ACKNOWLEDGEMENTS

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

REFERENCES

[1] G. Austin, *Chironomia; or, a treatise on rhetorical delivery*, T. Cadell and W. Davies, 1806.

[2] B. Bergen, S. Narayan, and J. Feldman, 'Embodied verbal semantics: Evidence from an image-verb matching task', in *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society*, pp. 139–144, (2003).

[3] M. Bucholtz and K. Hall, 'Embodied sociolinguistics', *Sociolinguistics: Theoretical debates*, (2016).

[4] J. Bulwer, *Chirologia, or the natural language of the hand.*, Thom. Harper and Henry Twyford, 1975.

[5] A. Cienki and C. Müller, 'Metaphor, gesture, and thought', *The Cambridge handbook of metaphor and thought*, 483–501, (2008).

[6] H. H. Clark, *Using language*, Cambridge university press, 1996.

[7] N. Cocks, G. Morgan, and S. Kita, 'Iconic gesture and speech integration in younger and older adults', *Gesture*, **11**(1), 24–39, (2011).

[8] E. Concepción, P. Gervás, and G. Méndez, 'A common model for representing stories in automatic storytelling', in *6th International Workshop on Computational Creativity, Concept Invention, and General Intelligence. C3GI*, (2017).

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

[10] E. Fischer-Lichte, *Ästhetik des Performativen*, Suhrkamp Verlag, 2012.

[11] A. Flinker, A. Korzeniewska, A.Y. Shestyuk, P.J. Franaszczuk, N.F. Dronkers, R.T. Knight, and N.E. Crone, 'Redefining the role of brocas area in speech', *Proceedings of the National Academy of Sciences*, **112**(9), 2871–2875, (2015).

[12] R. Gelin, C. d'Alessandro, Q. A. Le, O. Deroo, D. Doukhan, J.C. Martin, C. Pelachaud, A. Rilliard, and S. Rosset, 'Towards a storytelling humanoid robot.', in *AAAI Fall Symposium: Dialog with Robots*, (2010).

[13] P. Gervás, 'Computational drafting of plot structures for russian folk tales', *Cognitive computation*, **8**(2), 187–203, (2016).

[14] P. Gervás, B. Díaz-Agudo, F. Peinado, and R. Hervás, 'Story plot generation based on cbr', *Knowledge-Based Systems*, **18**(4), 235–242, (2005).

[15] C. Goddard and A. Wierzbicka, *Semantic and lexical universals: Theory and empirical findings*, volume 25, John Benjamins Publishing, 1994.

[16] D. Gouaillier, V. Hugel, P. Blazevic, C. Kilner, J. Monceaux, P. Lafourcade, B. Marnier, J. Serre, and B. Maisonnier, 'Mechatronic design of nao humanoid', in *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pp. 769–774. IEEE, (2009).

[17] B. Habets, S. Kita, Z. Shao, A. Özyurek, and P. Hagoort, 'The role of synchrony and ambiguity in speech–gesture integration during comprehension', *Journal of Cognitive Neuroscience*, **23**(8), 1845–1854, (2011).

[18] J. Ham, R. Bokhorst, R. Cuijpers, D. van der Pol, and J.J. Cabibihan, 'Making robots persuasive: the influence of combining persuasive strategies (gazing and gestures) by a storytelling robot on its persuasive power', in *International conference on social robotics*, pp. 71–83. Springer, (2011).

[19] M. Häring, N. Bee, and E. André, 'Creation and evaluation of emotion expression with body movement, sound and eye color for humanoid robots', in *Ro-Man, 2011 IEEE*, pp. 204–209. IEEE, (2011).

[20] O. Hauk, I. Johnsrude, and F. Pulvermüller, 'Somatotopic representation of action words in human motor and premotor cortex', *Neuron*, **41**(2), 301–307, (2004).

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

[22] T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro, 'Interactive robots as social partners and peer tutors for children: A field trial', *Human-computer interaction*, **19**(1), 61–84, (2004).

[23] S. D Kelly, D. J. Barr, R. B. Church, and K. Lynch, 'Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory', *Journal of memory and Language*, **40**(4), 577–592, (1999).

[24] S. D Kelly, J. M. Iverson, J. Terranova, J. Niego, M. Hopkins, and L. Goldsmith, 'Putting language back in the body: Speech and gesture on three time frames', *Developmental neuropsychology*, **22**(1), 323–349, (2002).

[25] A. Kendon, 'Gesticulation and speech: Two aspects of the process of utterance', *The relationship of verbal and nonverbal communication*, **25**(1980), 207–227, (1980).

[26] S. Kettebekov and R. Sharma, 'Toward natural gesture/speech control of a large display', *Engineering for human-computer interaction*, 221–234, (2001).

[27] S. Kettebekov, M. Yeasin, and R. Sharma, 'Prosody based audiovisual coanalysis for coverbal gesture recognition', *IEEE transactions on multimedia*, **7**(2), 234–242, (2005).

[28] K. L. Koay, D. S. Syrdal, M. L. Walters, and K. Dautenhahn, 'Living with robots: Investigating the habituation effect in participants' preferences during a longitudinal human-robot interaction study', in *Robot and Human interactive Communication, 2007. RO-MAN 2007. The 16th IEEE International Symposium on*, pp. 564–569. IEEE, (2007).

[29] R.W. Langacker, 'Nouns and verbs', *Language*, 53–94, (1987).

[30] Q. A. Le, S. Hanoune, and C. Pelachaud, 'Design and implementation of an expressive gesture model for a humanoid robot', in *Humanoid*

Robots (Humanoids), 2011 11th IEEE-RAS International Conference on, pp. 134–140. IEEE, (2011).

[31] Q. A. Le and C. Pelachaud, 'Evaluating an expressive gesture model for a humanoid robot: Experimental results', in *Submitted to 8th ACM/IEEE International Conference on Human-Robot Interaction*, (2012).

[32] B. Li, S. Lee-Urban, G. Johnston, and M. Riedl, 'Story generation with crowdsourced plot graphs.', in *AAAI*, (2013).

[33] K. F. MacDorman, T. Minato, M. Shimada, S. Itakura, S. Cowley, and H. Ishiguro, 'Assessing human likeness by eye contact in an android testbed', in *Proceedings of the XXVII annual meeting of the cognitive science society*, pp. 21–23, (2005).

[34] L. J Martin, P. Ammanabrolu, W. Hancock, S. Singh, B. Harrison, and M. O. Riedl, 'Event representations for automated story generation with deep neural nets', *arXiv preprint arXiv:1706.01331*, (2017).

[35] D. McNeill, 'So you think gestures are nonverbal?', *Psychological review*, **92**(3), 350, (1985).

[36] D. McNeill, *Hand and mind. What the hands reveal about thought*, Chicago: University of Chicago Press, 1992.

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

[38] I. Meir and O. Tkachman, *Iconicity*, Oxford University Press, 2014.

[39] I. Mittelberg, 'Methodology for multimodality', *MARQUEZ, MG; MITTELBERG, I. Methods in cognitive linguistics. Amsterdam: John Benjamins Publishing Company*, 225–248, (2007).

[40] M. Mori, 'The uncanny valley', *Energy*, **7**(4), 33–35, (1970).

[41] W. Nöth, *Handbook of semiotics*, Indiana University Press, 1995.

[42] R.E. Núñez and E. Sweetser, 'With the future behind them: Convergent evidence from aymara language and gesture in the crosslinguistic comparison of spatial construals of time', *Cognitive science*, **30**(3), 401–450, (2006).

[43] N. Parde, A. Hair, M. Papakostas, K. Tsiakas, M. Dagioglou, V. Karkaletsis, and R. D. Nielsen, 'Grounding the meaning of words through vision and interactive gameplay.', in *IJCAI*, pp. 1895–1901, (2015).

[44] C. Pelachaud, R. Gelin, J.C. Martin, and Q. A. Le, 'Expressive gestures displayed by a humanoid robot during a storytelling application', *New Frontiers in Human-Robot Interaction (AISB), Leicester, GB*, (2010).

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

[46] K. Pichotta and R. J. Mooney, 'Learning statistical scripts with lstm recurrent neural networks.', in *AAAI*, pp. 2800–2806, (2016).

[47] J. R. Searle, 'Minds, brains, and programs', *Behavioral and brain sciences*, **3**(3), 417–424, (1980).

[48] T. Veale, *Coming good and breaking bad: Generating transformative character arcs for use in compelling stories*, Proceedings of ICCC-2014, the 5th International Conference on Computational Creativity, Ljubljana, June 2014, 2014.

[49] T. Veale, 'Game of tropes: Exploring the placebo effect in computational creativity.', in *ICCC*, pp. 78–85, (2015).

[50] T. Veale, 'A rap on the knuckles and a twist in the tale', *AAAI spring symposium series*, (2016).

[51] T. Veale, 'Déjà vu all over again', in *ICCC*, pp. 245–252, (2017).

[52] T. Veale and A. Valitutti, 'A world with or without you', in *Proceedings of AAAI-2014 Fall Symposium Series on Modeling Changing Perspectives: Re-conceptualizing Sensorimotor Experiences. Arlington, VA*, (2014).

[53] P. Wicke. Ideograms as semantic primes: Emoji in computational linguistic creativity. Thesis DOI: 10.13140/RG.2.2.21344.89609 (2017).

[54] A. Wierzbicka, *Semantic primitives*, (Frankfurt/M.)Athenäum-Verl., 1972.

[55] G. Wilcock and K. Jokinen, 'Wikitalk human-robot interactions', in *Proceedings of the 15th ACM on International conference on multimodal interaction*, pp. 73–74. ACM, (2013).