Live, Die, Evaluate, Repeat:
Do-Over Simulation in the Generation of Coherent Episodic Stories

Stefan Riegl
Institute of Cognitive Science
University of Osnabrück
D-49069 Osnabrüuck, Germany
sriegl@uni-osnabrueck.de

Tony Veale
School of Computer Science and Informatics
University College Dublin
Belfield D4, Ireland
tony.veale@ucd.ie

Abstract

Struggling writers are sometimes tempted to throw away their current effort and start over with a a blank page and a fresh approach. But the cost of yielding to this temptation is high, especially when one has already sunk a great deal of time and energy into a work that must be discarded. However, as computational creativity increases the speed and lowers the cost of narrative generation, the option of a fresh do-over becomes ever more attractive. So we consider here a simulation-based approach to the generation of episodic stories in which stories are generated, evaluated and frequently discarded in a rapid, coarse-grained cycle of engagement and reflection. The goal of simulation is to better exploit the situated possibilities for information transfer amongst the characters in a story, while the goal of *repeated* simulation is to find the story that achieves maximal coherence amongst its episodic parts.

Introduction

A compelling story is like the juiciest gossip, so it is likely that people have been sharing views about what constitutes a good story long before Aristotle ever wrote the *Poetics*. As with gossip, *how* we are told is as crucial as *what* we are told, and linguistic framing is as important as the events that make up the causal substance of the story. But in addition to stylistic subtleties, listeners also appreciate how the presentation of a story adapts to the constraints imposed by its medium. How did an ancient Greek bard arrange the narration of the *Iliad* to efficiently hold the audience's attention span? How did George Martin portion the sweeping *A Song of Ice and Fire* into books of several hundreds of pages each, or screenwriters parcel it into episodic films of an hour apiece, while maintaining continuity and coherence throughout?

Long stories are often divided into episodic chunks to facilitate distribution and consumption, but each episode must be relatively self-contained while coherently linking to what has gone before and what will come next. Episodes can benefit from a unifying theme, such as a common goal, antagonist or location, yet each must slot into the grander sweep of the narrative by echoing past events or foreshadowing future ones. The use of echoing and foreshadowing, as reflected in what characters say and do, creates coherence in what might otherwise seem a rambling sequence of disjoint events. So a character that commits an egregious act in one episode may be punished or blackmailed for it in the next, while others may alter their views on the basis of this knowledge. In this way, information-sharing across episodes lends purpose to action and unites episodes into a coherent whole.

When a story is coherent and its actions well-motivated, certain elements can be usefully left unsaid, as these gaps will be filled by an engaged listener (Abbott 2008). But which actions are the key-frames of a story and which can be interpolated between them? We argue that it is the events that promote future information transfer (gossip, blackmail, boasting, threatening, etc.) that are key to understanding the mindset of a story's characters (Owens, Bower, and Black 1979). The stories that specify actions to which other characters visibly relate are the narratives that listeners can relate to also. The goal of episodic story generation should thus be to maximize the opportunities for a narrative to create coherent relations between characters and with the audience.

We present here a simulation-based approach to story generation that is based principally upon the *Scéalextric* model of (Veale 2017), but which also integrates elements of the *engagement and reflection* cycle as identified by (Sharples 1999) and implemented by (Pérez and Sharples 2001) in the MEXICA system. *Scéalextric Simulator* [1] generates a series of short self-contained episodes that it links together to form a long over-arching story. Episodes can be chained in different ways, though popular convention suggests that a persistent main character is the best way to create a single narrative thread. Even when one character – a hero – persists, episodes must cross-relate in other ways too, so that actions are seen to have far-reaching consequences. *Scéalextric Simulator* uses repeated generation, simulation and evaluation to find the threaded sequence of local episodes that maximizes a global measure of information coherence.

Many different genres of stories exist in the wild, of which some impose less stylisitic constraints than others, like some poems that derive their charm from their lack of a specific form. For the stories presented here we follow the tradition of a structuralist view of stories, implying a form of narratological events that supports coherence in a story.

Related Work

Creative writing is a pasttime that has been practiced for millenia by novices and professionals alike. The sheer diversity

[1] http://afflatus.ucd.ie/simlextric/

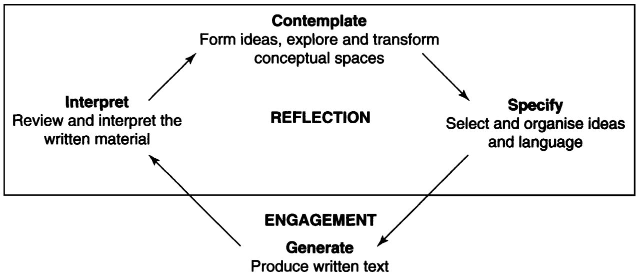

Contemplate
Form ideas, explore and transform conceptual spaces
Interpret
Review and interpret the written material
REFLECTION
Specify
Select and organise ideas and language
ENGAGEMENT
Generate
Produce written text

Figure 1: The writer's cycle of engagement and reflection, from (Sharples 1999).

of expression makes the process of creative writing particularly difficult to understand. The accumulated wisdom about the writer's craft, such as popular guides and maxims (e.g., "a story has a beginning, middle and an end"), often serve as metaphors or as simplified ideals, and rarely contribute to a deeper understanding of the act of writing itself.

Sharples (1990) has championed the view that writing is a process of problem-solving and creative design. He argues that writers do not start with a single intention but with a set of constraints that frames a space of possible solutions. Such constraints are limiting, but they usefully tell an author where focus can most effectively be placed. Even when making a humble shopping list, writing is more than the physical act of etching words on paper; it requires cognitive effort to evaluate an emerging text – what has been written and what remains to be said – and to satisfy the constraints of the given task. Sharples argues that these two processes occur sequentially in an alternating fashion (see Fig. 1). An engaged author continues to produce text until a cause, external or internal, necessitates a moment to stop and reflect. During reflection, the existing text is reviewed, new ideas are conceived, and a selected few are prepared for integration into the text, at which point engagement can continue.

The MEXICA model of (Pérez and Sharples 2001) gives an implemented, computational form to Sharples' model of engagement-reflection. MEXICA was designed to automate the generation of short stories and thereby facilitate the study of human creativity. It builds each new story from a stock of predefined story-structures, a set of adjustable parameters, and a list of explicit external and internal constraints. MEXICA operates with an inventory of primitive story actions, and a memory of previous stories composed only of those actions; those past stories illustrate how actions combine to form coherent plot sequences. In line with (Sharples 1999), MEXICA alternates between an engagement and a reflection phase: during engagement, new and appropriate actions are appended to the emerging story in a way that obeys rhetorical and thematic constraints. If the logical consequence of a new action contradicts the emerging story context, the story become incoherent and the system begins to reflect on why this is so. The emerging story is assessed in terms of novelty, interestingness and coherence, and appropriate actions are chosen for insertion into the story to restore coherence.

Coherence is a central concept in story generation, and one that operates over – and thus varies with – the stock of elements from which a story is to be constructed. For instance, early in the 20th century the Russian structuralist Vladimir Propp created a system to formalise Russian folk tales as sequences of character functions (Propp 2010). A text may contain one or more tales, of which each contains one or more *moves* that each permit a more-or-less self-contained narration. Each move comprises a sequence of functions, of which Propp defines 31 for Russian folk tales. Sequential moves can be interleaved in a variety of ways and exhibit dependencies of varying strength to each other. So when a later move introduces a new villain, the hero established in an earlier move will naturally reappear. Coherence in the Proppian scheme requires an author to ensure that moves are connected in ways that satisfy their associated character functions (Gervás 2013). The sequence of moves in a tale can be reordered provided the connections between functions remains satisfied.

A landmark in the history of computational storytelling is TALE-SPIN (Meehan 1977), a system that generates diverse stories from initial character descriptions or a desired moral. The movements of TALE-SPIN's characters through its story world are simulated so that they perform their actions rationally at each juncture, to realize their predefined goals. The sequential movements of a focal character are then arranged to provide the rendered story. TALE-SPIN's simulator uses knowledge of multiple domains to infer character beliefs, to resolve goals, and to translate goals into actions. TALE-SPIN is a spinner of short tales and no attempt was made to divide a narrative into a series of episodes or Proppian moves. In each state of the simulated world, the next is derived by allowing characters to perform their selected plans of action, and it is the rationality of these goal-based actions that imbues the story with a coherent shape.

The Knowledge-Intensive Interactive Digital Storytelling (KIIDS) model of (Gervás et al. 2005) is a framework for story generation that offers a set of ontologies modeling knowledge (and *meta*-knowledge) of the interactions in stories. Built on this is ProtoPropp, an automated story generator that gives Propp's analysis a computational form. ProtoPropp's stories typically feature multiple episodes that are created sequentially, where case-based reasoning selects the next episode based on constraints explicated by "the current state of the narration and using explicit knowledge about narrative and world simulation" (Peinado and Gervás 2006). When ProtoPropp's performance was evaluated in a user study, judges were asked to rate the coherence and novelty of ProtoPropp's stories amongst a set of randomly generated stories and human stories taken from a corpus. Overall, ProtoPropp's stories were deemed to be considerably closer to the human stories then to the random-generatd ones.

Generating and Integrating Episodes

A coherent episodic narrative is built from locally-coherent chapters (or *episodes*) that unite to maximize a global objective function. The *Scéalextric Simulator* presented here uses the *Scéalextric* system to generate the individual episodes which it then stitches together, simulates, and evaluates. As in *Scéalextric*, we assume that each episode is a tale of two characters (Veale 2017), but at least one of these can vary from episode to episode to yield a narrative with potentially many characters. *Scéalextric* is used to generate episodic plots with generic characters A and B, and *Scéalextric Simulator* then decides how these placeholders are to be filled.

Generating Episodes

(Veale 2017) modeled a two-character plot as a random-walk in a forest of causal links between action verbs. Each verb relates an A to a B, and sucessive verbs typically shift the focus from A to B and back to A. Bookend texts are defined for each verb so that a story can meaningfully begin with any action and terminate at any action. If we view the causal forest as a search-space, a coherent path can often be found between any two actions. Using the *Flux Capacitor* of (Veale 2014) to suggest actions at the beginning and at the end of a character's journey through a stereotypical category – e.g. going to medical school and losing one's medical license sit at opposite ends of one's journey through the Doctor category – *Scéalextric Simulator* can select meaningful start and end actions for the focal character in an episode, and use *Scéalextric* to fill in the rest of the plot. *Scéalextric* also provides ample templates for the rendering of plot actions in idiomatic forms, as well as causally-pairwise connections (such as *but, then, so*) to link plot actions.

There are as many episodes as there are character arcs, since each episode represents the journey of a character through a certain category, say from *pauper* to *millionaire* or from *believer* to *apostate*. To speed up story generation itself, we pre-generate a large inventory of these episodes – 12,000 in all – to be assembled into long-form narratives by the system. This inventory comprises alternate pathways through the causal forest for each of the character arcs defined in (Veale 2014). It allows *Scéalextric Simulator* to focus on the simulation of the plot in each episode rather than on its generation, and on the integration of multiple episodes into a coherent whole.

Linking Episodes

During simulation, the system maintains a placeholder view of characters. They are defined by their actions in the story, not by any prior knowledge. As characters interact in an episode, they gain information about each other. Interaction may involve issuing a threat, making a promise, or sharing a story about a past event. Over time, characters learn the things that make other characters proud or shameful, and can exploit this information to advance their own goals. Unlike TALE-SPIN and subsequent planning systems (e.g. (Riedl and Young 2010), these goals are not advanced incrementally; recall that the plot structure for each episode is generated prior to simulation. Rather, when an action is performed by a character as per the plot, and that action allows acquired information to be exploited, it is rendered as such, and the system records its contribution to global coherence.

Scéalextric Simulator provides such a view. For each of the 800+ A-B actions defined by *Scéalextric*, we associate at least one *continuation* action that an observer C may be likely to perform on A as a result. Thus, when episode n concludes with the event $A\ V_1\_act\_on\ B$ then a follow-up episode $n+1$ is chosen on the basis that its first action, $C\ V_1\_act\_on\ A$ is a valid continuation, inasmuch as $V_1\_act\_on$ suggests $V_2\_act\_on$ in the three-body causal model. There are episodes for any starting action, and each action is associated with multiple continuations. The selection of follow-on episodes is thus non-deterministic, and the simulator selects a follow-on randomly from the available choices.

Creating a World of Feelings and Memories

The *Scéalextric Simulator*'s knowledge-base describes characters, their actions, their locations and their beliefs. Each character resides at one location at a time, and can move between locations at episode boundaries. Each possesses a set of *affective* beliefs, each of which concerns a past action known to the character, as well as the intensity and type of the character's ”feeling” toward it. Intensity, an integer, ranges from 1 to 9, while type can be *proud*, *guilt*, *admiration* or *shock*. Agents feel pride or guilt for their own actions and admiration or shock for the actions of others. If an agent feels *pride* for its own action, an observer will feel *admiration* for that action, while if the agent feels *guilt* for its own actions, others will feel a degree of *shock*. Information is gained by characters either by observing their own actions or the actions of others. All 800+ of Scéalextric's actions have been assigned at least one pairing of type and intensity.

(Veale and Valitutti 2017) describe how a rich inventory of famous characters, called the NOC list, can be integrated into the story-generation process, so that their attributes and proclivities are reflected in the story's rendering. We do not exploit this depth of prior character detail here, and use the NOC list only to suggest the names of story characters (e.g., Bill and Hillary, Tom and Jerry, etc.). Character names are assigned to placeholders in the narrative (A, B, C, etc.) after the episodic plot structure has been created. At this point, each episode is associated with a different locale, drawn from a range of vivid options in the NOC list (e.g., a seedy nightclub, a ritzy hotel lobby).

The protagonist (denoted A) is the character that persists across all episodes. All other characters are antagonists (denoted B, C, etc.). For all antagonists a ”common past” is generated as a collection of shared beliefs. For a number of iterations of the simulator, a variable number of antagonists (but at least two) are assigned to a temporary, virtual location. Two characters from this virtual location are chosen to participate in a randomly-generated plot structure. Each action from the plot structure is simulated, which allows all the characters in the virtual location to observe and affectively react to the action, adding new beliefs to their memories.

Building on this foundation of shared memories, episodes are incrementally added to the emerging story. The first is chosen randomly, and subsequent episodes connect to the last via a causal continuation of its final action. Each new episode is set in a new locale and introduces a new antagonist. Glue text is inserted between episodes to explain the change of locale, and before the first episode and after the last episode to frame the story as a whole. In rare cases, a story cannot be progressed because no continuation can be found to launch a new episode. In these cases the simulation ends early, and the failed story is punitively scored.

Simulate, Reflect, Repeat.

In the MEXICA system of (Pérez and Sharples 2001) the unit of engagement is the action. New actions are added to an evolving tale in a process of engagement, prompting reflection on the consequences of each addition. The unit of engagement in the *Scéalextric Simulator* is not the action but the episode. New episodes are added to an evolving tale in a process of engagement, prompting reflection on their contribution to information transfer and global coherence.

A key issue concerns *when* reflection should take place. Should it occur incrementally, after the addition of each new episode, or upon the completion of a story? MEXICA employs incremental engagement, insofar as it pursues a greedy approach to generation: an evolving story is worked and reworked until it satisfies the desired constraints. Yet in a non-greedy approach that explores many alternate stories, it makes sense to reflect *after* each is completed and simulated. Scoring each story using an objective function that rewards global coherence, the highest-scoring story can be selected from a run of perhaps thousands of successive simulations.

Engagement in the *Scéalextric Simulator* governs the integration and simulation of new episodes into the story, while reflection governs the evaluation of each story once it has been completed. Engagement thus includes the transfer of information amongst characters, either by observing an action or reporting it to others. Reflection evaluates the impact of this transfer on the global coherence of the story.

Certain Scéalextric actions are defined as vectors of information transfer. Among others, these include *deceive*, *teach*, *confide_in* and *share_stories_with*. Deceitful communication requires the simulator to invent a false belief to communicate, while truthful communication (which is otherwise assumed) causes the most intense belief of the appropriate type to be transferred from sender to recipient (and all other observers). The action *confess_to* requires the type of the transferred belief to be one for which the speaker feels *guilt*. So when one character blackmails another, the extortion is assumed to relate to the most guilty feeling held by the victim. The coherence of a blackmail action is a function of the intensity of the guilty secret that one is blackmailed about, so a story in which A is blackmailed for killing B is preferred over one where A is blackmailed for merely insulting B.

This scoring of a story, to reward stories with well-motivated actions and punish those with weaker rationales, is a matter for the reflection phase. Whenever an action in one episode is motivated not just by the local plot, but by the actions of an earlier episode that are accessible due to information transfer, the overall story is rewarded accordingly. These connections across episodes also influence the rendering of the finished story, either through the insertion of mini-flashbacks, or by the rendering of direct speech that explicitly harks back to an earlier motivating action.

Once connections between episodes are established, and repeated simulation has identified a sequence of connected episodes to serve as a global plot, the plot is rendered as text. This rendering proceeds largely as in (Veale 2017) with some additions to create a more pleasant reading experience. Character names are rendered in different ways (first long, then short) or are replaced by gender-appropriate pronouns to avoid over-use of names. Adjectives that are suited to the respective action (e.g. *violent* for *attack*) are inserted into the idiomatic text to embellish the rendering of the protagonist and/or antagonist. Connecting text between episodes is inserted to note the movement of the persistent character between locales, and to establish the locale of the new episode.

An Objective Function for Global Coherence

The *Scéalextric Simulator* runs not one, but many simulations, one for each of its many successive attempts at generation. It scores each run according to an objective function that rewards global coherence, and renders the highest-scoring narrative into a polished idiomatic text. As an initial computational evaluation we quantify several features that influence the cross-episodic coherence of a story. The objective function $o(S)$ for the simulation S of a newly-generated narrative is defined as a summative score for the features in S divided by the square root of the number of actions in S:

$$o(S) = \frac{\sum_{f \in F(S)} s(f)}{\sqrt{\sum_{e \in S} |e|}}$$

Here $e \in S$ denotes an episode in the narrative under simulation in S and $|e|$ denotes the number of actions in the plot structure assigned to e by *Scéalextric*. $F(S)$ denotes the set of scorable features in S and $s(f)$ denotes the score for a single feature f, as defined in table 1. We divide by the square root of the total number of actions in a narrative to punish unnecessarily long stories whose actions do not earn their keep by contributing to the global coherence of the narrative. We divide by the square root because our empirical investigations show that the number of features does not tend to increase linearly with the length of a narrative.

Net Score	Feature Description
+10	information transfer (direct speech)
+6	– a character is lying
+n	– intensity n of the belief to be shared
+10	– the protagonist is talking
+20	– reference to past action
+50	– information shared for 2nd, 3rd, etc. time
+30	continuation of past action

Table 1: Relative contribution of features to story coherence.

The values in table 1 have been determined empirically over many trial simulations during feature development. The objective function thus rewards stories whose episodes are tightly cross-stitched by information transfers between characters. As noted earlier, these transfers are weighted by intensity and scored accordingly, since coherence is heightened when a narrative hinges on actions with pervasive influences that extend across episodes. So the actions that prove pivotal to a story, insofar as they motivate multiple future actions, are scored most generously of all. Actions that do not contribute to the score of a narrative are dead-weight, and repeated generation and simulation is a means of minimizing this narrative flab. This intuition is captured in the maxim

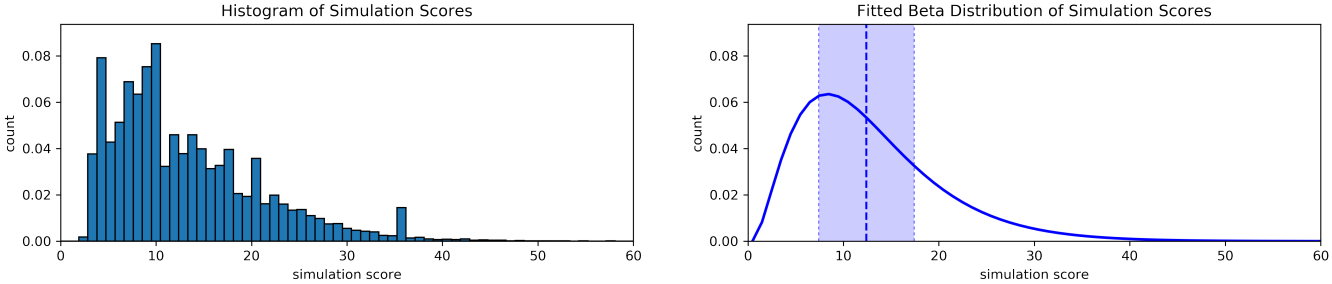


Figure 2: Distribution of global coherence score for 100,000 stories. A probability density function of a beta-distribution was fitted to the score distribution.

of "Chekov's pistol," which dictates that eye-catching flourishes (such as a gun prominently mounted on the wall) must earn their keep by meaningfully influencing the narrative.

Figure 2 presents the distribution of coherence scores produced by our objective function when we simulate 100,000 different three-episode narratives. The mean and standard deviation are 12.5 and 4.97 respectively, but we can greatly improve on the mean score with successive cycles of generation and simulation, retaining the highest-scoring narrative and discarding all others. To find a higher-scoring narrative, we simply run more cycles of generation and simulation. Figure 3 graphs the rate of increase in the score of the best narrative across repeated runs of generation and simulation. Each point in this graph represents the mean of the score for 100 three-episode stories. As also shown in Figure 3, a logarithmic function has been fitted to these mean values, demonstrating a logarithmic increase in mean score for repeated runs of the system. In addition to the mean for each iteration, Figure 4 also shows the standard deviation for each successive iteration across all stories.

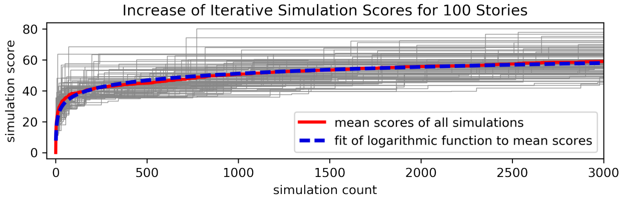


Figure 3: The changing scores of simulations for 100 generated stories. Mean values are per iteration over all stories and approximate a logarithmic function.

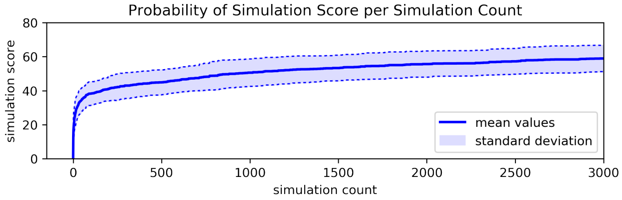


Figure 4: The changing scores of simulations for 100 generated stories, with mean values and standard deviation per iteration over all stories.

But how much is ever enough? When should the generator by satisfied with a particular narrative and a particular score? To achieve excellence we must quantify excellence, by e.g. imposing explicit minimum scores that must be exceeded by a qualifying narrative. For instance, as shown in Figure 3, 1000 cycles of generation, simulation and evaluation – which requires just seconds to execute – is typically sufficient to achieve a score that is multiple standard-deviations higher than the mean score achieved for any given narrative. Alternately, we can express this threshold as the number of standard deviations above the mean that is required for excellence. A *six sigma* threshold thus requires a successful narrative to score more than six standard deviations above the mean on the system's measure of global coherence. Or we can do without thresholds altogether, and simply use the available time to the fullest. In a *just-in-time* setting, the system generates new narratives for as long as it is permitted. When the system is interrupted, or its allotted time has run out, it simply returns the highest-scoring narrative it has thusfar generated.

Evaluation

We have seen that repeated do-overs tend to steadily increase the scores achieved by narrative generation on our objective function, which has been defined to codify our own notions of global narrative coherence. But does our notion of coherence comport with the intuitive view held by others, such as by the end consumers of the stories that are generated? What prospect is there of asking lay judges to rate the "coherence" of a long narrative in a way that facilitates meaningful empirical evaluation?

(Veale and Valitutti 2017) present the results of crowd-sourcing experiments in which anonymous judges were asked to rate stories generated by *Scéalextric* under two conditions and along six dimensions: *eventfulness, imagination, laughter, silliness, entertainment* and *vividness*. Stories from *Scéalextric Simulator* were evaluated according to these six dimensions using the same approach and the same testing conditions as (Veale and Valitutti 2017), again using the CrowdFlower platform with comparable demographics for the 50 test subjects a.k.a. judges. While the mean results (of 10 judgments per story per dimension) showed meaningful

The story of Maura

Maura needed a place to live and Saoirse had plenty of it.

Maura found Saoirse at an underground lair. Maura rented accommodation from her. She paid Saoirse what she owed. Saoirse could not achieve bossy Maura’s lofty goals. She refused to honour Saoirse’s commitments to her, so Maura ripped off rich Saoirse’s best ideas. Anguished Saoirse evicted Maura from Saoirse’s home.

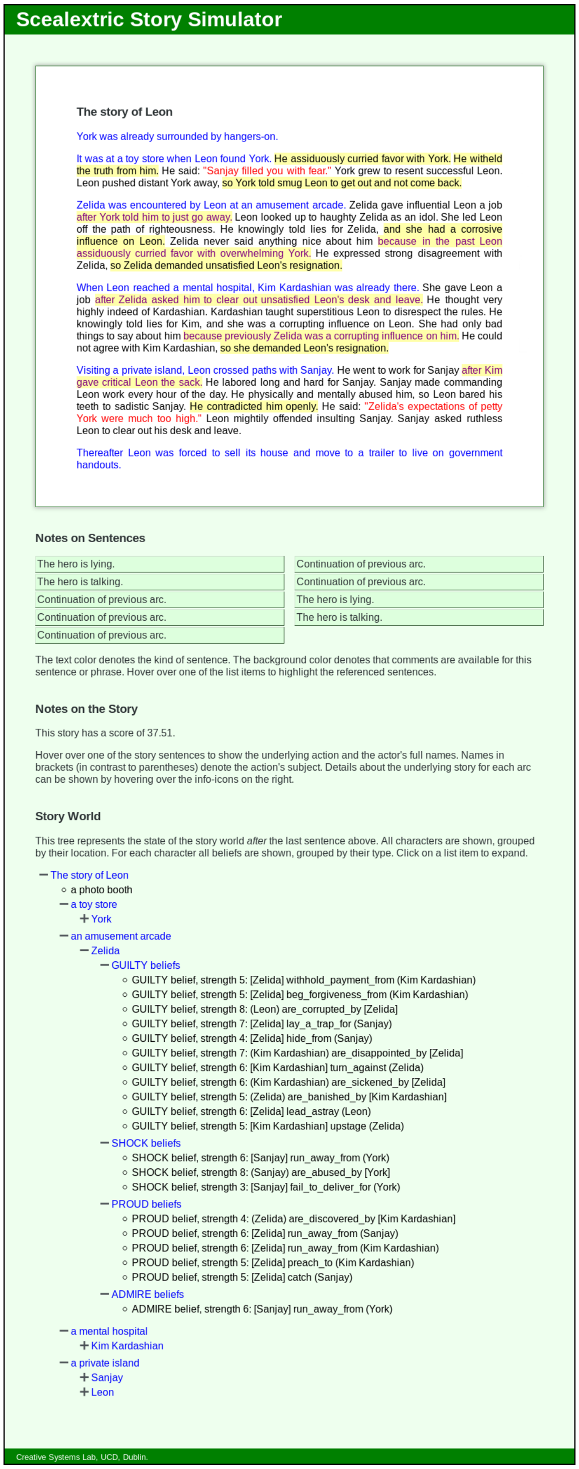
At a smoke-filled back room Maura met Oscar Wilde. Maura assiduously curried favor with dictatorial Oscar after cheated Saoirse evicted Maura from Saoirse’s home. Maura told eager Wilde a pack of lies. Maura said: ”Dolores wrote propaganda to promote your cause.” His attitude hardened toward Maura. He openly disrespected Maura because earlier she took everything that Saoirse had. Maura tried to tune out loudmouthed Oscar’s voice. Bossy Oscar Wilde wrote Maura off as a loser, so he coldly dismissed Maura and turned away.

It was at the red carpet when Maura found Rina. Maura started a new job for influential Rina after unsatisfied Oscar told Maura to get out and not come back. Rina took full advantage of her. She pulled the wool over Maura’s eyes. She said: ”Saoirse was a real suck-up to aristocratic Wilde.” Maura could not reach the bar set by bossy Rina. She was very disappointed in her, so ”Get out! You’re fired” said Rina.

It was at a recording studio when Maura found Dolores. Authoritarian Dolores recruited Maura into her ranks after Rina asked her to clear out her desk and leave. Maura took the spotlight from lackadaisical Dolores. Dolores withheld due payment from lazy Maura. Maura criticized sinful Dolores in public. She said: ”Saoirse showed no shame in sucking up to influential Wilde.” She broke with her and went her own way.

What do you think? Can Maura and Dolores ever mend their relationship?

Figure 5: An full example story generated by the Scealéxtric Simulator. The story has four episodes with a variable number of sentences, an opening and closing sentence and episode intros. Direct speech is used to present transfer of information and inserted clauses at the end of sentences refer to motivating actions from previous episodes.



Scealextric Story Simulator

The story of Leon

York was already surrounded by hangers-on.

It was at a toy store when Leon found York. He assiduously curried favor with York. He witheld the truth from him. He said: "Sanjay filled you with fear." York grew to resent successful Leon. Leon pushed distant York away, so York told smug Leon to get out and not come back.

Zelida was encountered by Leon at an amusement arcade. Zelida gave influential Leon a job after York told him to just go away. Leon looked up to haughty Zelida as an idol. She led Leon off the path of righteousness. He knowingly told lies for Zelida, and she had a corrosive influence on Leon. Zelida never said anything nice about him because in the past Leon assiduously curried favor with overwhelming York. He expressed strong disagreement with Zelida, so Zelida demanded unsatisfied Leon's resignation.

When Leon reached a mental hospital, Kim Kardashian was already there. She gave Leon a job after Zelida asked him to clear out unsatisfied Leon's desk and leave. He thought very highly indeed of Kardashian. Kardashian taught superstitious Leon to disrespect the rules. He knowingly told lies for Kim, and she was a corrupting influence on Leon. She had only bad things to say about him because previously Zelida was a corrupting influence on him. He could not agree with Kim Kardashian, so she demanded Leon's resignation.

Visiting a private island, Leon crossed paths with Sanjay. He went to work for Sanjay after Kim gave critical Leon the sack. He labored long and hard for Sanjay. Sanjay made commanding Leon work every hour of the day. He physically and mentally abused him, so Leon bared his teeth to sadistic Sanjay. He contradicted him openly. He said: "Zelida's expectations of petty York were much too high." Leon mightily offended insulting Sanjay. Sanjay asked ruthless Leon to clear out his desk and leave.

Thereafter Leon was forced to sell its house and move to a trailer to live on government handouts.

Notes on Sentences

The hero is lying.
The hero is talking.
Continuation of previous arc.
Continuation of previous arc.
Continuation of previous arc.
Continuation of previous arc.
Continuation of previous arc.
The hero is lying.
The hero is talking.

The text color denotes the kind of sentence. The background color denotes that comments are available for this sentence or phrase. Hover over one of the list items to highlight the referenced sentences.

Notes on the Story

This story has a score of 37.51.

Hover over one of the story sentences to show the underlying action and the actor's full names. Names in brackets (in contrast to parentheses) denote the action's subject. Details about the underlying story for each arc can be shown by hovering over the info-icons on the right.

Story World

This tree represents the state of the story world *after* the last sentence above. All characters are shown, grouped by their location. For each character all beliefs are shown, grouped by their type. Click on a list item to expand.

− The story of Leon
  ○ a photo booth
  − a toy store
    + York
  − an amusement arcade
    − Zelida
      − GUILTY beliefs
        ○ GUILTY belief, strength 5: [Zelida] withhold_payment_from (Kim Kardashian)
        ○ GUILTY belief, strength 5: [Zelida] beg_forgiveness_from (Kim Kardashian)
        ○ GUILTY belief, strength 8: (Leon) are_corrupted_by [Zelida]
        ○ GUILTY belief, strength 7: [Zelida] lay_a_trap_for (Sanjay)
        ○ GUILTY belief, strength 4: [Zelida] hide_from (Sanjay)
        ○ GUILTY belief, strength 7: (Kim Kardashian) are_disappointed_by [Zelida]
        ○ GUILTY belief, strength 6: [Kim Kardashian] turn_against (Zelida)
        ○ GUILTY belief, strength 6: (Kim Kardashian) are_sickened_by [Zelida]
        ○ GUILTY belief, strength 5: (Zelida) are_banished_by [Kim Kardashian]
        ○ GUILTY belief, strength 6: [Zelida] lead_astray (Leon)
        ○ GUILTY belief, strength 5: [Kim Kardashian] upstage (Zelida)
      − SHOCK beliefs
        ○ SHOCK belief, strength 6: [Sanjay] run_away_from (York)
        ○ SHOCK belief, strength 8: (Sanjay) are_abused_by [York]
        ○ SHOCK belief, strength 3: [Sanjay] fail_to_deliver_for (York)
      − PROUD beliefs
        ○ PROUD belief, strength 4: (Zelida) are_discovered_by [Kim Kardashian]
        ○ PROUD belief, strength 6: [Zelida] run_away_from (Sanjay)
        ○ PROUD belief, strength 6: [Zelida] run_away_from (Kim Kardashian)
        ○ PROUD belief, strength 5: [Zelida] preach_to (Kim Kardashian)
        ○ PROUD belief, strength 5: [Zelida] catch (Sanjay)
      − ADMIRE beliefs
        ○ ADMIRE belief, strength 6: [Sanjay] run_away_from (York)
  − a mental hospital
    + Kim Kardashian
  − a private island
    + Sanjay
    + Leon

Creative Systems Lab, UCD, Dublin.

Figure 6: An excerpt of a story in the web interface as of February 28, 2018. Introspection tools can be used to analyse the presented story, including tool tips and colour-coded annotation of the story text, interactive highlighting of sentences and clauses that increased the score and a hierarchical representation of locations, characters and beliefs. The web interface is accessible here: `http://afflatus.ucd.ie/simlextric/`

separation between the two conditions, it was also ruefully noted that many judges opt for the middle rating when presented with a Likert scale. Given that the stories evaluated in (Veale and Valitutti 2017) were short single-episode affairs, we can expect more judges to take the lazy middle option when presented with multi-episode narratives that are considerably longer and more taxing to analyse.

This was indeed the case when we replicated those earlier experiments on three-episode narratives generated by the *Scéalextric Simulator*. Although plots have now been given a greater sense of direction through the use of *Flux Capacitor* to specify the start and end actions of a causal path and thereby create a more meaningful arc for the protagonist, and episodic plots are integrated using 1000 cycles of simulation to select the highest-scoring narratives, the mean judgments of anonymous human raters tend to be lower than either condition in the original experiments. For the most part, the new results fall within a standard deviation of the old, with one notable exception: the *laughter* dimension.

The two conditions in (Veale and Valitutti 2017) relate to humorous intent. In the simple generic condition, straight narratives – which is to say, narratives that are not intended to be humorous – are generated using baseline *Scéalextric* capabilities. In this generic condition, A and B character placeholders are simply filled with random animals in the Aesop tradition (e.g. the monkey and the snake). In the NOC condition, A and B are chosen to be familiar characters that are, in a deliberate metaphorical sense, well-matched. These characters are further chosen so as to evoke a meaningful incongruity in the guise of postmodern irony. Thus, fictional characters are paired with real people, as in Lex Luthor and Donald Trump, or characters are paired with similar entities from different eras, as in Steve Jobs and Leonardo Da Vinci, or characters are paired on the basis of a shared screen portrayal, as in Frank Underwood and Keysar Söze. In this condition, actions are rendered into text using vivid details of the characters as provided by the NOC. The goal is to foster humorous incongruity in character and action while using the same basic plotting mechanism of the generic condition.

Narrative coherence is no laughing matter. All things being equal, we can expect a long narrative with low internal coherence to strike a reader as more laughable – perhaps more laughably bad – than one whith high internal coherence. Most contemporary theories of humour thus emphasize the role of incongruity in the construction of laughter-inducing texts (see e.g., (Suls 1972), (Raskin 1985), (Ritchie 1999), (Ritchie 2003)). An incongruity is any logical impasse or jarring misalignment of expectation and reality that stops readers in their tracks. Some incongruities are more subtle than others, while others are significantly more dramatic. A vexing question surrounds the actual role of incongruity in humour: is it a profound phenomenon that tickles the funny bone and triggers cognitive recovery mechanisms, or is it merely an epiphenomenon that accompanies, but does not explain, most instances of humour (Veale 2004)?

In any case, laugher is a visceral response to a situation. Unlike the other dimensions evaluated in (Veale and Valitutti 2017), which each need a brief explanation so that judges can understand what they are supposed to score, laughter needs no explanation. Narrative coherence, quite unlike laughter, is a rather abstract quality that is far from visceral, and requires an even more substantive explanation to judges than any dimension in the original experiments. However, insofar as low coherence creates the conditions for unintended incongruities to arise in a text, we can expect a straight narrative with low coherence to evoke more laughter on average than a straight narrative with high coherence.

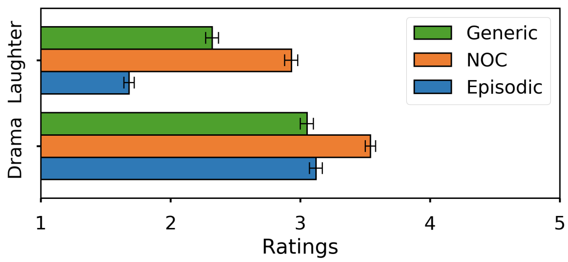


Figure 7: Comparison of mean ratings for the dimensions of drama and laughter for the NOC and Generic conditions from (Veale and Valitutti 2017) and the Episodic condition from this evaluation. Error bars denote standard errors.

The experiments reported in (Veale and Valitutti 2017) show that humour can be engineered in a text by fostering the kinds of incongruity – between real and fictional, or contemporary and historical, or between distinct fictional worlds – that encourage laughter. Those earlier experiments report a mean Laughter score for NOC stories of 2.93, significantly higher than the mean Laugher score for generic stories, 2.32. In effect, the narratives that were engineered to be creatively incongruous were deemed to be significantly more humorous than their straight counterparts.

Given that our multi-episode narratives are *not* engineered to be humorous, and are, moreover, engineered to be as internally consistent and free of incongruity as possible, we expect the mean laughter scores for these narratives to not only fall far short of those attained for NOC narratives, but to also fall significantly short of those attained for generic narratives. As shown in Figure 7 this is indeed the case. With a mean score of 1.68 for laughter, our episodic narratives have seemingly been drained of their inconsistencies by repeated engagement and reflection, so that they prompt much less unintended laughter than comparable stories that are not chosen for their coherence.

From Meet-Cute to Cliff-Edge and Beyond

The flow of information between the characters in a story – who knows *what*, and *how/why/when* do they know it? – is every bit as as important as the flow of information from author to audience. For how this flow is managed will dictate the coherence of the narrative and influence the feelings it engenders in a reader. For example, as shown in (Delatorre et al. 2017), a tightly-managed information flow can greatly enhance the enjoyable sense of suspense that authors hope to nurture in consumers of thrilling or mysterious stories. In this work we have focused more on the feelings of story characters than on story consumers, in the hope that the latter will feel the benefits of coherence that accrue from the careful tracking of the former. This is especially important for the generation of longer stories that incorporate multiple episodes, characters and locales.

We have shown that repeated simulation of what characters know, how and when they know it, and how they exploit this knowledge to advance the plot, underpins a measure of global coherence that can be steadily increased over repeated cycles of generation and iterative evaluation of episodes. While this approach departs substantially from how humans create stories, we believe it can nonetheless be considered a coarse-grained version of the engagement-reflection loop that is championed by (Sharples 1999) and implemented in the MEXICA system of (Pérez and Sharples 2001). It is an approach that makes a virtue of starting over, of failing fast and of failing better, because in conditions like ours it is more costly to fix a broken, highly-constrained episode than to make a fresh start from the last known good episode. Moreover, it facilitates a *just-in-time* view of the story-generation process that is ideally suited to the implementation of that process as a creative web service (see e.g., (Veale 2013) and (Concepción, Gervás, and Méndez 2017)).

The crowd-sourced evaluation has been conducted as a pilot for a more comprehensive study yet to come. Vexing challenges with the evaluation of long stories have forced us to look for validation of our objective function via a roundabout and creative interpretation of the experimental results. Understandably, anonymous raters who are paid small amounts per rating cannot be trusted to fully engage or to give a reliable picture of anything but the most visceral of phenomena. As we improve our objective function to capture additional aspects of global coherence, we will have to find other means of evaluating the resulting stories.

As shown in (Delatorre et al. 2017), a viable area of improvement concerns the fostering of suspense at the boundaries of adjacent episodes. In the "cliff-hanger" serials of old, in which long cinematic narratives were broken into a series of weekly instalments, each episode would conclude with a moment of high suspense by placing the protagonist in a position of impending doom. Subsequent episodes would quickly deflate the suspense, only to create a new predicament for the hero to endure. Long stories that manage the ebb and flow of suspense in this way should do more than hold the interest of an engaged reader: they may also go some way toward holding the sustained attention of an otherwise disinterested crowd-sourcing volunteer.

References

Abbott, H. P. 2008. *The Cambridge introduction to narrative*. Cambridge University Press.

Concepción, E.; Gervás, P.; and Méndez, G. 2017. An api-based approach to co-creation in automatic storytelling. In *6th International Workshop on Computational Creativity, Concept Invention, and General Intelligence. C3GI*.

Delatorre, P.; León, C.; Salguero, A.; Mateo-Gil, C.; and Gervás, P. 2017. Impact of interactivity on information management for suspense in storytelling. In *Proceedings of the 4th AISB Symposium on Computational Creativity*. AISB.

Gervás, P.; Díaz-Agudo, B.; Peinado, F.; and Hervás, R. 2005. Story plot generation based on cbr. *Knowledge-Based Systems* 18(4-5):235–242.

Gervás, P. 2013. 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.

Meehan, J. R. 1977. Tale-spin, an interactive program that writes stories. In *Ijcai*, volume 77, 91–98.

Owens, J.; Bower, G. H.; and Black, J. B. 1979. The "soap opera" effect in story recall. *Memory & Cognition* 7(3):185–191.

Peinado, F., and Gervás, P. 2006. Evaluation of automatic generation of basic stories. *New Generation Computing* 24(3):289–302.

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

Propp, V. 2010. *Morphology of the Folktale*, volume 9. University of Texas Press.

Raskin, V. 1985. *Semantic Mechanisms of Humor*. D. Reidel.

Riedl, M. O., and Young, R. M. 2010. Narrative planning: balancing plot and character. *Journal of Artificial Intelligence Research* 39(1):217–268.

Ritchie, G. 1999. Developing the incongruity-resolution theory. In *Proceedings of the AISB Symposium on Creative Language: Stories and Humour*. AISB.

Ritchie, G. 2003. *The Linguistic Analysis of Jokes*. Routledge Studies in Linguistics, 2. Routledge.

Sharples, M. 1999. *How we write: Writing as creative design*. Psychology Press.

Suls, J. 1972. A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis. In Goldstein, J. H., and McGhee, P. E., eds., *The Psychology of Humor*. New York, NY: Academic Press. 81–100.

Veale, T., and Valitutti, A. 2017. Tweet dreams are made of this: Appropriate incongruity in the dreamwork of language. *Lingua* 197:141–153.

Veale, T. 2004. Incongruity in humor: Root-cause or epiphenomenon? *HUMOR: The International Journal of Humor* 17(4):419–428.

Veale, T. 2013. A service-oriented architecture for computational creativity. *Journal of Computing Science and Engineering* 7(3):159–167.

Veale, T. 2014. Coming good and breaking bad: Generating transformative character arcs for use in compelling stories. In *Proceedings of ICCC-2014, the 5th International Conference on Computational Creativity, Ljubljana, June 2014*.

Veale, T. 2017. Déjà vu all over again: On the creative value of familiar elements in the telling of original tales. *In Proceedings of ICCC 2017, the 8th International Conference on Computational Creativity, Atlanta, Georgia, June 19-23*.