Applying Narrative Theory to Aid Unexpectedness in a Story Generation System

Todd Pickering
School of Computing
University of Kent
Canterbury, Kent, UK
tmp9@kent.ac.uk

Anna Jordanous
School of Computing
University of Kent
Medway, Kent, UK
a.k.jordanous@kent.ac.uk

Abstract

Predictability is the polar opposite of originality, and as such it is a notable obstacle that should be overcome in the pursuit of computational creativity. Accurately modelling a human's understanding of predictability would be a monumental task, requiring a contextually rich network of social interaction, literature, news, and media. However, by artificially instilling a computer with some basic ideas about what is predictable in a given scenario, it can begin to gain an understanding of how to subvert expectation.

This project attempts to implement such a process into a specially designed story generation system known as *Chronicle*, inspired by Vladímir Propp's *Morphology of the Folk Tale*. *Chronicle* aims to fine-tune narrative direction and progression in a system modelled on predictability.

Decisions made during the story generation process are based on probabilities defined by the expectations of the typical reader, and are amassed to formulate an overall predictability rating. The decision making process is manipulated by the system in order to pursue a customisable predictability target.

Chronicle was demonstrably accurate at evaluating its output in some cases, and less accurate in other cases. Further refinement is required to increase its efficacy, but it presents a promising step towards negotiating predictability in computational creativity.

Introduction

Computers are capable of solving mathematical problems due to their comprehensive knowledge of the existing laws of mathematics. If we extrapolate this, we can surmise that computers may also be capable of solving other types of problems, so long as they possess the necessary knowledge, context, and understanding to do so.

However, this presents a significant challenge when negotiating computational creativity in written fiction, since there are many facets of creativity and authorship that cannot be so rigorously defined.

One such facet is the way in which a narrative is constructed: how one event follows another, and how this impacts future events. It demonstrates an understanding of cause and effect, contextual awareness, and a capability to make decisions, all of which are fundamental steps towards truly creative computing.

Objectives

The project objectives are as follows:

- Create a functional story generation system capable of producing a variety of outputs.
- Implement narrative theory during development to aid the generation process.
- Ensure that the system is able to affect narrative direction and progression based on its understanding of predictability.

The Importance of Unpredictability

Computers struggle to display any degree of spontaneity or unexpectedness due to the foundation of rules and constraints upon which they are built. This is advantageous for mathematical and scientific pursuits, where consistency and reliable behaviour are paramount. However, it becomes an issue wherever computational creativity is concerned, due to the disparity between predictability and originality. This is doubly the case when fiction is involved; a story with a predictable outcome can lead to an unsatisfying experience for the reader.

It is this intrinsic predictability that must be overcome in order to generate truly creative output. The first logical step is to make the computer aware of when it is and is not being predictable.

Existing Work

Computational creativity in written fiction has been explored and investigated in numerous ways since the early 1970s. However, it is the more recent developments that demand a greater focus, since they tend to address specific concepts and issues in greater and more relevant detail.

Unexpectedness

Kazjon Grace and Mary Lou Maher established that "unexpectedness is [...] a vital component of computational creativity evaluation" (Grace and Maher, 2014). They proposed a series of five properties with which to standardise expectations - "predicted", "prediction", "scope",

"condition", and "congruence" - and six dichotomies which "[categorise] creativity-relevant expectations based on these properties".

This model notes the salient differences between variegated degrees of expectation, and in doing so highlights some of the trickier aspects that should be negotiated. For instance, "prediction" requires an accurate numeric representation of the potential variance in values for a given expectation. Using an example from Grace and Maher's paper, this is relatively simple to apply when defining an expectation of an object's height as somewhere "between two and five metres". But when considering the expectation of progression of story events, it becomes rather more difficult to numerically assimilate a reader's opinion in quite the same way. Unfortunately, we often have to compromise and approximate subjective opinions in this manner (as is the case with *Chronicle*).

Due to the difficulty in translating these existing concepts, it would be beneficial to develop a new system with which to establish and monitor unexpectedness.

A computer must have a degree of awareness about what constitutes predictability before it can begin to overcome it. We can understand 'predictability' as a measure of how likely something is to happen. In that respect, it is closely linked to probability, which is an ideal concept to utilise considering its capacity to formalise chance and likelihood into hard data.

Every time a decision needs to be made during the story generation process, each of the possible options can be assigned a probability based on the typical reader's expectation. When a decision is made, the system can record the probability of the chosen option, and begin to gain an understanding of the overall predictability of the story.

Narrative Theory

Narrative theorists typically analyse fictional constructs for literary purposes, yet the act of breaking down a structure into its constituent parts to gain a better understanding of its composition is an intrinsically mechanical process. As such, the work of narrative theorists has an enormous potential relevance to computational creativity.

Vladímir Propp analysed one hundred Russian folk tales, and in doing so identified a series of recurring elements (Propp, 1968). He established seven character archetypes (or *actants*), and a series of thirty-one story events (or *functions*). Propp determined that each of the stories he analysed was comprised of a combination of these actants and functions. Specific details about these actants and functions varied from tale to tale, but the underlying structure remained largely the same.

For instance, in one crucial story function, a nefarious character commits an act of villainy. The details of this function could be fantastical, with a wicked witch or an evil dragon casting a spell or abducting a person. Instead, the details may be grounded in reality, with an odious stepmother or a belligerent neighbouring tsar ordering a murder. The specific details are not especially important; what matters is that an evil act occurs in order to upset the equilibrium and motivate the hero's quest.

An argument could be made that Propp's theory is a reductionist view of narrative structure, and therefore too restrictive for a story generation system; it is certainly true that not all fiction is comprised of these actants and functions. But logistically speaking, the computer has to be provided with a structure of some description, and from an engineering perspective, Propp's theory lays an ideal foundation for formalisation into software with a reasonable degree of freedom in its potential output.

Pablo Gérvas sought to employ Propp's *Morphology* to create new narrative structures (Gérvas, 2013). Gérvas accomplished this by incorporating Propp's functions and actants in fabula and plot driver generators.

However, Gérvas' research is not without its issues. He suggests that "some deviation is allowed [...] by shifting certain character functions to other positions in the sequence", while Propp's theory states that "the sequence of functions is always identical" (Propp, 1968). As such, Gérvas' use of plot driver generators to alternate the order of functions appears to deviate from Propp's literature.

Gérvas revisited Propp with another research project focusing on plot structure (Gérvas, 2016), which relies heavily upon his aforementioned 2013 project. A notable addition relates to what Gérvas names "long-range dependencies", which are the links established between corresponding story functions, such as a character being kidnapped, and the same character being rescued later in the story. Gérvas notes that these "long-range dependencies" have had a "very significant impact on the quality of the resulting plots".

Narrative theory can undoubtedly be used to aid the structure of story generation systems, but its efficacy can be affected by the degree of faithfulness to the original literature, and an evaluative process should be employed since the implementation of theory is not infallible.

Vladímir Propp's *Morphology of the Folk Tale* (Propp, 1968) presents itself as an ideal framework, since each of its constituent parts can be assimilated into computerised functions. This would not necessarily result in as rigid a structure as one might first anticipate; while "the sequence of functions is always identical", Propp stated that "by no means do all tales give evidence of all functions" (Propp, 1968). In other words, despite the fact that story functions are never rearranged, there is the potential for a number of them to be omitted: certain functions will only occur if their corresponding precursor functions have already occurred. For instance, if the hero does not enter into a chase, then they do not need to be rescued. This permits a relative degree of freedom within the confines of a defined structure, which is an ideal starting point for a computationally creative project.

Propp's *Morphology* has previously been utilised by Pablo Gérvas, whose work builds upon Propp's theory in an effort to generate entirely new plot structures, and also employ self-evaluative rating systems (Gérvas (2013), and Gérvas (2016)). *Chronicle* differs in that it focuses on variance of content and the pursuit of unexpectedness within an existing morphology, and the ratings it assigns relate to user expectation rather than a score based on conformance with a structure.

Chronicle also utilises an additional level of detail in each of the story functions, resulting in generated outputs constituted of fully-fledged sentences and paragraphs (with substitutions for randomised lexis, where appropriate), and seeks to incorporate subtle elements of humour. Considering *Chronicle's* output resembles a typical story (rather than a series of short sentences devoid of detail), it encourages a more genuine response from human volunteers, thereby overcoming Gérvas' issue with evaluators having to interpret "abstract representations". This is also a problem experienced by Rafael Pérez y Pérez, despite the nuanced plot structures his system generates (Pérez y Pérez, 2015). A story without detail is like a skeleton with no muscle, or the foundations of a house with no walls; we can recognise that it is a solid framework, but it lacks significant value when deprived of its defining details.

Output Evaluation

It is relatively simple to analyse the output of a computer that deals with hard data, because we know that 1 is always bigger than 0, smaller than 2, and not equal to 574. It is comparatively much more difficult to analyse creative output, particularly in a manner which enables meaningful comparison between multiple outputs. This is because the quality of a creative material is largely subjective. Once it has been established that a creative output satisfies the rules of grammar and incorporates a sufficient vocabulary, it largely becomes a matter of opinion as to how good (or otherwise) the text may be.

Additionally, there are multiple interpretations of what constitutes a 'good' text. For instance, a pithy crime thriller may have the capacity to excite a reader, but could fall short in the realms of literary excellence. Similarly, a great work from the literary canon may be considered the pinnacle of narrative innovation, but could send a reader to sleep should they attempt to negotiate it.

As such, a standardised evaluative process should be employed in order to overcome these hurdles, either eliminating or sufficiently accounting for the subjectivity of human evaluators.

Human Evaluation Anna Jordanous outlined the *SPECS* methodology, which is a three step process (Jordanous, 2012). The evaluator must declare their definition of creativity (in regard to the system they are evaluating), identify the standards for which they will be testing, and then test the system using those standards. It prioritises targeted feedback on specific aspects rather than an arbitrary numeric "creativity score".

The use of targeted feedback is certainly appropriate for attempting to quantify subjectivity, but the resulting lack of hard data increases the difficulty of accurate evaluation and comparison. Jordanous recognises this, and discusses the complications of attempting to reach a consensus while handling differing opinions.

The best that can be done when employing human evaluation is to follow a strict set of principles as objectively and realistically as possible, but even then, issues are likely to be encountered.

Self-Evaluation Rafael Pérez y Pérez outlined the "three layers" evaluation model as a potential evaluative methodology (Pérez y Pérez, 2014). This differs from Jordanous' approach in that it eliminates human opinion from the evaluation process. The first "layer" ensures that the plot is suitable and valuable enough to be evaluated. The second "layer" evaluates the "core characteristics" of the plot, in this case "climax", "closure", and "novelty". The third "layer" is responsible for "enhancers and debasers" which modify the overall "score" either positively (for original value) or negatively (for repetition).

This evaluative model is able to translate a creative output into hard data, which is arguably much more desirable than comparing subjective human feedback since it allows for clearer comparison and analysis. However, any plots or stories requiring evaluation need to have followed a particular format for the model to be applied correctly.

We may also wish to consider the implications of a computer evaluating its own output: if the output necessitates evaluation in the first place, it is indicative of the fact that we doubt its creative capabilities; as such, is it right that we rely on the same system to evaluate the output? Conceptually, this may sound like a valid concern, but in practice (due to the algorithmic nature of the evaluative process) this does not appear to be an issue.

Pérez y Pérez utilises a self-evaluative process referred to as an "engagement-reflection" (E-R) model (Pérez y Pérez, 2015). "Engagement" constitutes the generation of sequences and events, while "reflection" is responsible for evaluation and modification of the generated material. The story writing process is passed back and forth between two "agents" (each with a differing set of characteristics) which engage the E-R cycle at every step.

Similarly to Pérez y Pérez's three-layered approach (2014), this process encourages evaluation on multiple levels for each step of the story generation process, creating an autonomous feedback loop. The addition of a second agent simulates a collaborative environment in which multiple 'writers' contribute to the story, which more closely portrays the human creative experience (whether the second 'writer' is actually another person, or merely representative of the first writer's awareness of context). Due to the usage of numeric values for "emotional links" and the visual representation of agent "contextual knowledge structures", the process provides an amount of hard data that can be utilised for objective comparison.

Self-evaluation is not always this effective in its implementation. Gérvas' 2013 project results in the output of an "abstract representation", which is "plagued with difficulty" as far as human evaluators are concerned, and likely to lead to "difficulty of interpreting the representation". This means that for Gérvas, a system based evaluation is the only option.

In principle, this level of self-evaluation may sound like a desirable autonomous feature. However, Gérvas mentions these "quantitative procedures" with little detail. With no third parties able to interpret or quantify the output, and when the evaluation system runs "at a corresponding abstract level", the reliance of this self-evaluation is

questionable. His more recent research (Gérvas, 2016) suffers from the same evaluation issue: it operates at an "abstract level", and cannot be evaluated (or assessed for accuracy) by human volunteers.

Self-evaluation can be an incredibly powerful tool, allowing multiple stories to be autonomously generated and evaluated in the time it would take a human evaluator to even begin to read a single story. However, it must be possible to test the quality and accuracy of the self-evaluation process in order to quantify its efficacy and reliability, and this likely requires one or many human evaluators. As such, a desirable approach would be to build a system that can be evaluated by both human and computer, and does not rely too heavily on one or the other.

Implementation

Chronicle was developed in Java. It establishes a number of features whose functions need to be closely studied in order to achieve a greater understanding of the system as a whole.

Propp Story Functions

A method was implemented for each of Vladímir Propp's thirty-one story functions. The characters that appear within these functions each correspond to one of Propp's seven archetypes, such as "hero", "villain", and "princess" (Propp, 1968). Method calls are structured in such a way that no function will ever occur out of sequence, thereby maintaining narrative consistency. The manner in which these methods are negotiated is determined by probability and the system's decision making process.

Similarly to Pablo Gérvas's approach, the generation process ends "when the end of the sequence is reached" (Gérvas, 2016). Depending on the events of the story, this can potentially result in a premature ending in which the hero does not succeed in completing their quest.

Decision Making & Probability

Decision Making Decisions made during the story generation process are determined by probabilities based on what the average reader is likely to expect. For example, when the hero enters conflict with the villain, the typical reader will expect the hero to succeed. Thus, the probability of the hero succeeding is set at 80%, and the corresponding probability of the hero failing is set at 20%.

The system utilises roulette wheel selection in its decision making process; two probabilities are requested in the form of integers that sum to a total of 100, and boundaries are established based upon these integers. Using the above example, the system would establish boundaries of 1-20 for the hero failing, and 21-100 for the hero succeeding. A number between 1 and 100 would be randomly generated, and the system would return the corresponding decision based on which set of boundaries the value fell between.

Successive story functions are also determined with this process. Using the above example, if the hero failed the encounter, the narrative would end with their death. This would result in the omission of any remaining story functions. If it was determined that the hero succeeded in the encounter, a second decision would be made to determine whether the villain was killed outright, or merely injured and therefore able to flee. If the villain was killed, then the hero's journey home would be uninterrupted. If the villain had fled, then they would be able to pursue the hero on his or her journey home, leading to an additional encounter in which the hero would be endangered a second time.

Predictability Rating Every time a decision is made, the probability of that option's occurrence is recorded cumulatively, along with the total number of decisions that have been made within the current story. An overall predictability rating can then be calculated for each story by dividing the cumulative probability by the total number of decisions. A higher predictability rating indicates a high level of predictability, whereas a lower rating indicates unpredictability.

Predictability Target Every story has a user-customisable predictability target that it will attempt to meet with its overall predictability rating. In order to do this, the system utilises a probability modifier.

Probability Modifier The system can invert the decision making probabilities by subtracting them from a probability modifier and returning the absolute value. For example, with a default probability modifier value of 0, and a decision with 80% probability:

```
0 - 80 = 80
```

...hence, the probability is unchanged. When the odds are tipped, the probability modifier is adjusted to 100, and the calculation is as follows:

```
100 - 80 = 20
```

...hence, inverted. This approach works in any scenario in which there are two decisions whose probabilities sum to a total of 100.

The system will re-evaluate its predictability rating every time a decision is made, and will either maintain or invert the probabilities in order to pursue its target. With this functionality in place, the system is able to dynamically monitor the progress of the story, and consciously make changes during its generation.

It is important to note that while the probabilities are inverted, it is still the original (non-inverted) probability that is added to the cumulative total. This is because we want the rating to continue to reflect the typical user's expectations.

Word & Name Randomisation

Randomisation is typically a function that should be avoided in computational creativity, since it replaces autonomy with luck or random chance. However, its usage is justifiable in instances where the effects of the randomisation do not have a significant impact on the events of the story. Word randomisation is akin to a human writer selecting a different word from a thesaurus, and as such it can be justified as a reasonable introduction of variety into an otherwise repetitive story segment.

The system contains a bank of names (separated by gender) and words (separated into categories such as

locations, objects, colours, positive adjectives, negative adjectives, etc). In predetermined places, the system can be asked to pick a word from one of these categories. This means that while the structure of a particular segment may be quite similar from story to story, there is potential for lexical variety.

Article Selection: 'A' vs 'An'

Due to the irregularities and inconsistencies of the English language, it is difficult to predict whether a word chosen at random should be prefaced with 'a' or 'an'. We can have instances of "an honest man", or "a honeybee", and "a university" or "an unidentified object". The system negotiates this issue by comparing the first few characters of a given word to a number of predefined cases in sequential order until a match is found, and succeeds in assigning the correct article to a magnitude of different words.

Gender Pronoun Selection: 'Him' vs 'Her'

Each of the character archetypes can be portrayed by both males and females. In order to fully support this and maintain consistency, the system dynamically selects the correct pronouns based on the associated character's gender. For example, with a male character, one passage might read: "The man opened his eyes". With a female character, the same passage would read: "The woman opened her eyes".

Sample Output

A number of stories generated by *Chronicle* (as well as the application itself) can be found online at the following location: *https://github.com/toddpickering/chronicle*

Below is an excerpt from a predictable story, at the moment the hero learns of the villain's wrongdoing:

"Arthur hurriedly dialled his voicemail, and listened to the first message. It was Jenny. Her daughter had been kidnapped. Arthur took one last look at the desert, then turned and headed back towards the forest as quickly as he could. Upon arrival, he asked Jenny what had happened."

The same event in the narrative can read quite differently when the system pursues an unpredictable story:

"Daisy hurriedly dialled her voicemail, and listened to the first message. It was Evie. Her son had been kidnapped. Daisy didn't much care for Evie or her son, so she decided not to help, and spent the day at the canyon."

The system has a number of opportunities to end the narrative (in an expected manner or otherwise). Below is an example of a particularly unexpected story in which the main character chooses not to leave their home, resulting in a story only a single paragraph long:

"There was a man named Oliver. He spent most of his time at the swamp. Nearby, there lived a man named Steve. Oliver was feeling especially unadventurous, and decided not to leave the swamp. Steve thought this incredibly boring."

An excerpt from the ending of a much longer tale can be found below. At this point in the narrative, the false hero has been ridiculed after attempting to take credit for the hero's actions, and is now seeking revenge.

"Darren was ridiculed for his foolish claims. He was furious with Karl, deciding it was all his fault. Darren wanted revenge. Darren climbed to the top of the tree beside Karl's house, intending to attack him when he walked by. Unfortunately for Darren, he was at the wrong house. After several hours, Darren began to tire. He lost his grip and fell from the tree, breaking his neck."

User Test

Human-based evaluation can present issues of subjectivity. However, concepts such as creativity, originality, and unexpectedness are largely subjective themselves, and therefore can be difficult to evaluate from an analytical standpoint. These concepts play a crucial part in the understanding of how successful (or otherwise) *Chronicle* is. As such, it was necessary to acquire responses from third-party human participants in order to evaluate the supposed effectiveness of the software.

User Test Process

A survey was created and distributed online, containing a download link for the software and instructions on how to use it. The user was prompted to generate a story and read it in its entirety, and then assign the story two ratings.

The first rating concerned how entertaining the user found the story. The recording of this rating allowed comparisons to be made between user enjoyment, the user's perceived predictability of the story, and the system's perceived predictability of the story. 'Entertaining' is an open-ended, subjective word, and it was deliberately chosen for this reason. Since enjoyment is a subjective concept, it should be evaluated as such. Different people garner enjoyment and entertainment through different means. What is important is whether or not the user appreciates the creative output; their exact reasoning for this or the manner in which they do this is of less importance.

The second rating concerned how predictable the user found the events of the story. It was important to establish the scope of this question in an effort to avoid any existing impressions or biases that may be carried over from previous story generations. Ideally, we desired a higher predictability rating for stories that the software deemed to be predictable, and we desired a lower predictability rating for stories that the software deemed to be unpredictable. If this trend was followed, it would suggest that the software is accurate at identifying a user's perceived predictability of a story.

This process was repeated until each user had read and evaluated six stories. Unbeknownst to the user, the first three stories had a predictability target of 100% (very predictable), while the latter three stories had a predictability target of 0% (not at all predictable). The goal was to have the users' ratings match up with these targets.

It was important that the stories aiming to be unpredictable were generated last; generating these stories first could lead to inaccurate results, since a new user might evaluate a story as being unpredictable simply because they are unfamiliar with the software and have not yet picked up on its patterns, rather than the story being genuinely unpredictable as a result of the software's deliberate intervention. However, there is a possibility of introducing ordering bias by following this strategy.

The first three (predictable) stories gave the user the opportunity to get acquainted with the software, during which they would likely have begun to pick up on some of its patterns. Then, once it came to generating the latter three (unpredictable) stories, the user would have been better equipped to evaluate the overall predictability.

After completing the testing process, the user was prompted to submit statistics generated by the system during their session, and was given the chance to submit written feedback about their experience with the software.

There are numerous reasons why users were asked to generate six stories. Firstly, it had to be an even number, so that there could be an even distribution of predictability targets (e.g. three stories with a 100% target, and three stories with a 0% target). Secondly, a smaller number of stories (such as two or four) would not have given the user adequate experience with the software. With so few stories generated, it would have been difficult for the user to identify any recurring patterns, potentially leading them to assign an inaccurately generous predictability rating. Thirdly, it was necessary to have a relatively small workload for each user to complete in an effort to retain user focus, and to encourage more participants to respond to the survey. In a project such as this (where subjectivity and opinion play significant roles) it is far more useful to have a wider range of users evaluating a smaller number of stories than it is to have a very small number of users evaluating many stories.

Results & Analysis

Twelve surveys were completed, each containing evaluations of six stories. This resulted in a total of 72 evaluated stories. User predictability and entertainment scores were rated on a scale of 1-5, then translated to a scale of 0-100 for clearer comparison with system ratings.

Each user generated their own unique set of stories so that their evaluation could include their direct interaction with the software. However, it is possible that this approach introduced noise into the evaluation.

The relatively small sample size and potential for noise is indicative of the fact that these are merely preliminary results, and as such correlation measures have not been included. More conclusive results and comprehensive statistical analysis could be pursued as further research.

Predictability: User Rating vs System Rating

Human and system predictability ratings were closely related for some users, but for other users the system's ratings appeared to be less accurate. The averaged ratings for each of the evaluated stories are shown in figure 1.

The data in figure 1 demonstrates an increase in user perceived predictability across the first three stories, which reflects the expectation that users would begin to notice patterns in the generated stories as they became more familiar with the software. Nevertheless, the system was successful in generating unpredictable stories, evidenced by the decrease in user perceived predictability for stories 4-5. The increase in predictability for story 6 suggests that by this point in the test, the users may have become accustomed to the system's typical output.

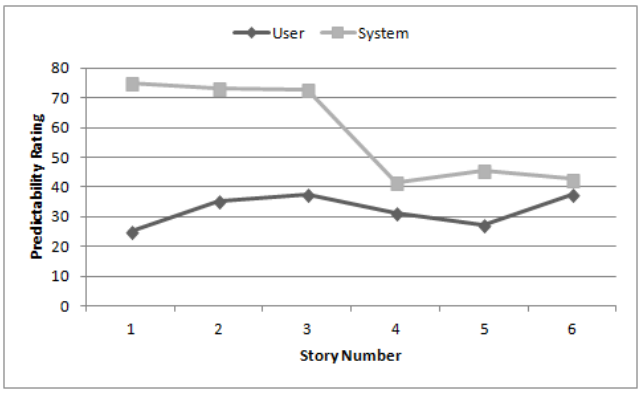

Figure 1: Averaged system and user predictability ratings for all 12 users. The lighter points indicate the ratings assigned by the system. The darker points indicate the ratings perceived by the user. A higher rating suggests a more predictable story, while a lower rating suggests a less predictable story. The system ratings have a mean of 58.5 and a standard deviation of 16.7. The user ratings have a mean of 32.3 and a standard deviation of 5.4.

Predictability: System Rating vs System Target

Figure 2 shows that while the system's predictability ratings never meet its targets precisely, it is capable of pursuing its assigned predictability target.

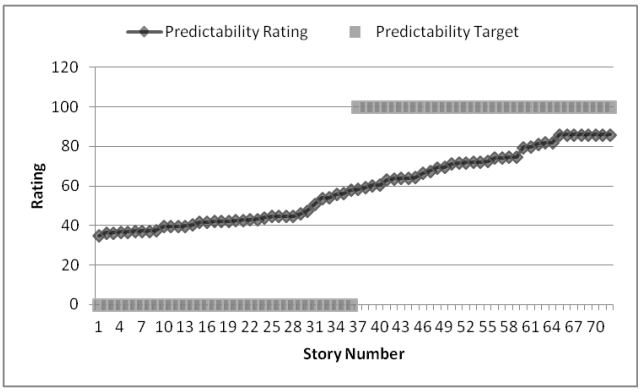

Figure 2: The system's predictability rating compared to the system's predictability target for all 72 evaluated stories, sorted by target (ascending) and rating (ascending).

User Entertainment vs User Predictability

Figure 3 demonstrates that a user's enjoyment of a story is inversely proportionate to their perceived predictability of said story: the more unpredictable a story, the more the user will enjoy it. This suggests that unpredictability is a desirable facet of a story generation system, aligning with Grace and Maher's assertion that "unexpectedness is [...] a vital component of computational creativity evaluation" (Grace and Maher, 2014), and validating *Chronicle*'s pursuit of unpredictability.

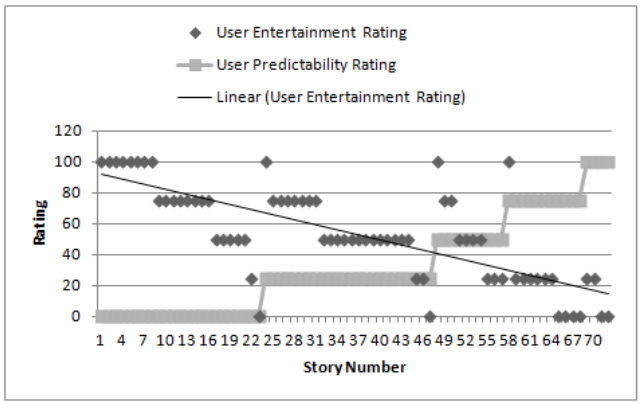

Figure 3: A comparison of user entertainment and predictability ratings for all 72 evaluated stories, sorted by predictability (ascending) and entertainment (descending).

User Entertainment vs System Predictability

Figure 4 demonstrates a similar (though admittedly much weaker) correlation between user entertainment and system predictability rating. This suggests that with a great deal more refinement, the system's predictability rating could potentially be used to predict a user's entertainment rating. However, in its current state, the system is quite a way from reaching such a goal.

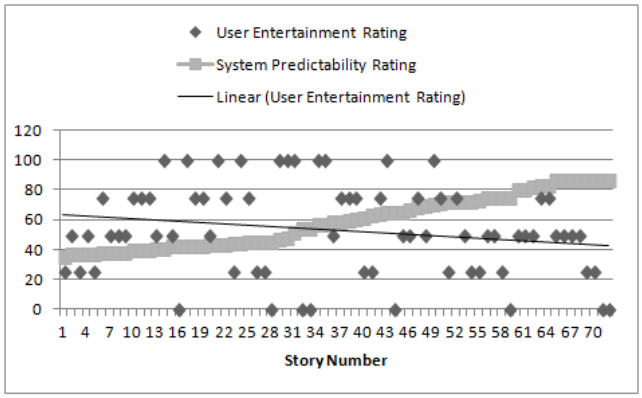

Figure 4: The users' entertainment ratings compared to the system's predictability ratings for all 72 evaluated stories, sorted by predictability rating (ascending) and entertainment rating (descending).

User Feedback

One user remarked that the endings which stood out the most and were the most enjoyable were those that were unexpected. Another user stated that it was the unexpected elements which made the greatest contribution to their enjoyment. A different user noted that on several occasions they were anticipating a certain end to the story, only for the system to develop an unexpected sub-plot.

Some users noted that it was difficult to keep track of the names of all of the story characters, and others made similar remarks about location names. This can be attributed to the fact that generic names are used, and is likely exacerbated when reading six stories consecutively. Others noted that word selection was occasionally jarring or inaccurate, which can be attributed to the categorisation of vocabulary (despite already being separated by word type and positivity/negativity); just because a word is a positive adjective, it does not mean it can be used in all contexts. Another user stated that they felt a greater degree of variety could have been added to story events.

Evaluation

With a sample of only 12 users, there is not enough data to show statistically significant evidence. However, interesting results are suggested by the results and feedback obtained thus far.

Opinion and accuracy of results differed from user to user; such an outcome was inevitable in a project that, from a reductionist perspective, attempts to turn subjective opinion into hard data. Additionally, it is worth noting that a certain amount of unpredictability is always going to be sacrificed when following a predetermined structure, even with added variation.

The system appears to be fully competent in pursuing its assigned predictability target. Based on the test results, the predictability rating system shows promise, but lacks accuracy in relation to user ratings.

When reading the source code, it is simple to keep track of characters and locations, since they are all clearly defined by their field names. From the user's perspective, when randomisation is introduced, it becomes understandably more difficult. Memorable names (potentially based on status or character type) and locations would help to alleviate this issue. While it is worth noting that only a small number of users commented on this aspect, it is a clear example of the importance of considering the user's perspective at all times during software development.

The users' experience with the software may have been adversely affected by response bias, potentially introduced by the mention of 'predictability' in the survey questions. This is undesirable, but subtly attaining this rating without directly naming the concept would have been very difficult. Allowances could at least be made for this issue in the test mode of the software, in which statistics relating to predictability remained hidden from the user until completion of the process.

Many of the story functions would benefit from a reassessment of their level of detail; some contain too much, while others do not contain enough. This creates an imbalance in certain stories, especially if several functions of a similar level of detail are selected, potentially leading to a story whose detail is overwhelmingly prevalent, or decidedly minimal. However, a noticeable improvement (in regard to detail) has been made on existing projects such as Pérez y Pérez's *MEXICA* system (2015), in which stories are constituted of simplified one sentence segments which arguably struggle to encourage user engagement.

Some story functions could benefit with a more liberal usage of word randomisation to increase variety; the rigidity

of a predetermined structure needs to be offset by sufficient variation in content. However, refinement of the word randomisation process is necessary to ensure contextual consistency. This could be achieved through a more nuanced categorisation system for each of the vocabulary files, with each clearly defined by purpose and tone.

Chronicle succeeded in meeting each of the project objectives, although there is certainly room for improvement in order to improve both the system's accuracy and the efficacy of its features.

Further Research

Predictability Rating Refinement

Chronicle has demonstrated the potential of the predictability rating system, but also its inaccuracies. Refinement of this system could be approached by standardising probabilities based on actual user expectations in specific scenarios, as opposed to estimated expectations. This could be accomplished with a large scale user survey concerning the events of story scenarios, but would require a sizeable number of participants.

Alternative Narrative Theorists

Propp's *Morphology* has proven to be an effective foundation for many story generation systems. Future researchers might consider assimilating the work of different narrative theorists for the purposes of plot development or character design.

- **Joseph Campbell** outlined the 'hero's journey', in which seventeen 'stages' of a narrative adventure are divided between three 'acts': 'departure', 'initiation', and 'return' (Campbell, 1949). Similarly to Propp's *Morphology*, these 'stages' would be apt for assimilation into computerised functions.
- **Tzvetan Todorov** theorised that every story follows a structure involving the upset and re-establishment of the equilibrium (Todorov, 1971). This theory is comparable to Propp's *Morphology* in its usefulness: it defines a loose structure for a system to follow, but allows an even greater degree of freedom in regards to the events of each of these constituent parts.
- **Roland Barthes** proposed a theory that narratives are understood on the basis of five 'codes', each with a different function (Barthes, 1970). One example is 'enigma codes', which relates to a reader's necessity to unfurl mysteries or uncertainties as a plot progresses. These 'codes' could be deconstructed to formulate story functions.
- **Richard Bartle** established a relatively modern character theory model, ascribing four archetypes to humans playing video games in virtual worlds based on their intended goals (Bartle, 1996). This provides an ideal basis for systems that are driven by character actions and motivations, rather than predetermined structures.

Conclusion

Refinements can be made in a number of places, but *Chronicle* was successful in achieving each of the project objectives. There is a good amount of variance in the system's output. Propp's *Morphology* was successfully incorporated during the development process and utilised during story generation. The system is capable of modifying a story's narrative direction mid-generation, and does so based on its understanding of predictability. While the system's ratings are less accurate in some places, they show promise in others, demonstrating at least a partial understanding of predictability and unexpectedness.

Ultimately, *Chronicle* constitutes a valid contribution to the field, and presents a solid foundation upon which further research can be undertaken.

Acknowledgements

This work was undertaken by Todd Pickering in fulfilment of an MSc Computer Science project and dissertation, and was supervised by Anna Jordanous, whose support and guidance proved invaluable throughout.

References

Barthes, R. 1970. *S/Z.* Paris: Editions du Seuil.

Bartle, R. 1996. Hearts, clubs, diamonds, spades: Players who suit muds. *Journal of MUD research* 1(1):19.

Campbell, J. 1949. *The Hero with a Thousand Faces*. New York: MJF Books.

Gérvas, P. 2013. Propp's morphology of the folk tale as a grammar for generation. In Finlayson, M. A.; Fisseni, B.; Löwe, B.; and Meister, J. C., eds., *2013 Workshop on Computational Models of Narrative*, volume 32 of *OpenAccess Series in Informatics (OASIcs)*, 106–122. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Gérvas, P. 2016. Computational drafting of plot structures for russian folk tales. *Cognitive Computation* 8(2):187–203.

Grace, K., and Maher, M. L. 2014. What to expect when you're expecting: The role of unexpectedness in computationally evaluating creativity. In Colton, S.; Ventura, D.; Lavravc, N.; and Cook, M., eds., *Proceedings of the Fifth International Conference on Computational Creativity*, 120–128. Ljubljana, Slovenia: Elsevier Procedia.

Jordanous, A. 2012. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation* 4(3):246–279.

Pérez y Pérez, R. 2014. The three layers evaluation model for computer-generated plots. In Colton, S.; Ventura, D.; Lavravc, N.; and Cook, M., eds., *Proceedings of the Fifth International Conference on Computational Creativity*, 220–229. Ljubljana, Slovenia: Elsevier Procedia.

Pérez y Pérez, R. 2015. A computer-based model for collaborative narrative generation. *Cognitive Systems Research* 3637:30–48.

Propp, V. 1968. *Excerpts from Morphology of the Folk Tale (American Folklore Society Bibliographical and Special Series)*. Indiana University Research Center in Anthropology, Folklore, and Linguistics, 2nd edition.

Todorov, T. 1971. The 2 principles of narrative. *Diacritics* 1(1):37–44.