A model for evaluating interestingness in a computer-generated plot

Rafael Pérez y Pérez, Otoniel Ortiz
Departamento de Tecnologías de la Información
Universidad Autónoma Metropolitana, Cuajimalpa
Av. Constituyentes 1054, México D. F.
{rperez/oortiz}@correo.cua.uam.mx

Abstract

This paper describes a computer model for evaluating the interestingness of a computer-generated plot. In this work we describe a set of features that represent some of the core characteristics of interestingness. Then, we describe in detail our computer model and explain how we implemented our first prototype. We assess four computer-generated narratives using our system and present the results. For comparison reasons, we asked a group of subjects to emit an opinion about the interestingness of the same four stories. The outcome suggests that we are in the right direction, although much more work is required.

Introduction

Evaluation is a core aspect of the creative process and if we are interested in building creative systems we need to develop mechanisms that allow them to evaluate their own outputs. The purpose of this project is to contribute in that direction.

This paper describes a model for evaluating the interestingness of a computer generated plot. It is part of our research project in computer models of narrative generation. Some time ago we developed a computer model of narrative generation (Pérez y Pérez and Sharples 2001; Pérez y Pérez 2007). Our model distinguished three core characteristics: coherence, novelty and interestingness. To test our model we built an agent that generated plots. Now, we are interested in developing a model to evaluate the coherence, novelty and interestingness of a computer-generated narrative. So, our storyteller agent will be able to evaluate its own outputs. In this way, we expect to understand better how the evaluation process works and, as a consequence, how the creative process works. Due to space limitations this document only discusses the central features of our model for the evaluation of interestingness (the reader can find some published work describing the main characteristics of our model for evaluation of novelty in Pérez y Pérez et al. 2011).

We are aware that human evaluation of interestingness is a very complex task and we are far from understanding how it works. Nevertheless, we believe that computer models, like the one we describe in this text, can provide some light in this challenging aspect of human creativity.

Related Work

There have been several discussions about how to assess computational creativity. For example, Ritchie (2007) suggests criteria for evaluating the products of a creative process (the process is not taken into consideration); in general terms such criteria evaluate how typical and how valuable the product is. Colton (2008) considers that skill, imagination and appreciation are characteristics that a computer model needs to be perceived to have. Jordanous (2012) suggests to have a set of human experts that evaluate characteristics like Spontaneity and Subconscious Processing, Value, Intention and Emotional Involvement, and so on, in a computer generated product. All these are interesting ideas, although some are too general and difficult to implement (e.g. see Pereira et al. 2005). Some work has been done in evaluation of plot generation:

> A computer model might be considered as representing a creative process if it generates knowledge that does not explicitly exist in the original knowledge-base of the system and which is relevant to (i.e. is an important element of) the produced output. Note that this definition involves inspection of both the output of the program and its initial data structures... we refer to this type of creativity as computerised creativity (c-creativity) (Pérez y Pérez and Sharples 2004).

Peinado et al. (2010) also have worked in evaluation of stories, although they work was oriented to asses novelty. An area that some readers might consider related to this work is interactive drama and drama managers. A good example of this type of systems is the work by Weyhrauch (1997). However, rather than evaluating the plot and the creative process, Drama managers focus in evaluating the user's experience while playing the game. Some other systems might employ different techniques, e.g. case base systems (Sharma et al. 2010), but the goal is the same: to provide a pleasant experience to the user.

Description of the Model

This work describes a model to evaluate the interestingness of a computer generated plot. Such a plot is known as the new story or the new narrative. For the purpose of this project, we consider a narrative interesting when it is recounted in a correct manner and when it generates new knowledge. A story is recounted in a correct manner when it follows the classical Aristotelian structure of a story: introduction, development, climax and resolution (or setup, conflict and resolution). Some previous work has shown the relation between the Aristotelian structure and the evaluation of interestingness in computer generated plots (Pérez y Pérez and Sharples 2001). We are particularly interested in evaluating the opening and the closure of a story. We consider that a story has a correct opening when at the beginning there are no active dramatic tensions in the tale and then the tension starts to grow. We consider that a story has a correct closure if all the dramatic tensions in the story are solved when the last action is performed. An important characteristic of the recountal of a story is the introduction of unexpected obstacles. In this work an obstacle is unexpected when the story seems to finish (final part of the resolution section) and then new problems arise.

Following Pérez y Pérez and Sharples (2004) we believe that the generation of new knowledge contributes to consider a narrative interesting. Some studies in motivation, curiosity and learning seem to support this claim (e.g. see Deckers 2005). In the same way, writers have pointed out how good narratives are a source of new knowledge (e.g. see Lodge 1996). In this work a new story generates new knowledge when:

- It generates knowledge structures that did not exist previously in the knowledge base of the system and that can be employed to build novel narratives.
- It generates a knowledge widening, i.e. when existing knowledge structures incorporate unknown information obtained from the new story. This information can be employed to build novel narratives.

Our computer model of evaluation is based on expectations. So, the assessment of the new knowledge structures and the knowledge widening is performed by analysing how much the new story modifies the knowledge base; then, comparing if such modifications satisfied the given expectations. In the same way, the evaluations of unexpected obstacles and the correctness of the narrative's recountal are performed by analysing the structure of the new narrative; then, assessing if such a structure fulfils there expectations. Finally, all these partial results are considered to obtain a final evaluation of interestingness. The following lines elaborate these ideas.

Generating Original Structures

One of the key aspects of c-creativity is the generation of novel and relevant knowledge structures. That is, a storyteller must develop narratives that increment its knowledge base (in this work we focus on how the knowledge base of the evaluator is incremented). Thus, a storyteller must include mechanisms that allow: 1) incorporating within its knowledge base the new information generated by its outputs, i.e. it must include a feedback process; 2) comparing its knowledge base before and after feeding back a new tale (an interesting point for further discussions is to compare the processes that different systems might employ to perform these tasks).

In this way, the first part of the model focuses in determining the proportion of new structures. It requires a parameter known as the Minimal Value of New Structures (Min-NS); it represents the minimum amount of new structures expected to be created by the new story. In this way, the Proportion of New Structures (PNS) is defined by the ratio between the number of new structures (NNS) created by the new narrative and the Minimal Value of New Structures (Min-NS). If the number of new structures is bigger than its minimal value, the Proportion of New Structures is set to 1.

$$\text{PNS} = \begin{cases} \dfrac{\text{NNS}}{\text{Min-NS}} & \text{IF NNS} \leq \text{Min-NS} \\ 1 & \text{IF NNS} > \text{Min-NS} \end{cases}$$

Besides calculating the number of new structures, it is necessary to determine how novel they are, i.e. to verify if they are similar to the information that already exists in the knowledge base. With this purpose we define a parameter known as the Limit of Similitude (LS) that represents the maximum percentage of alikeness allowed between two knowledge structures.

So, all those new structures that are too alike to already existing structures must be eliminated. In other words, one must get rid of all new structures that are at least LS% equal to any existing structure. The number of surviving structures is known as the Original Value (O-Value) and they represent new structures that are not similar to any old structures. Like in the previous case, the model requires an expected Minimum Original Value (Min-OV) to calculate the Proportion of the Original Value (POV). And, like in the previous case, this proportion never can be bigger than 1.

$$\text{POV} = \begin{cases} \dfrac{\text{O-Value}}{\text{Min-OV}} & \text{IF O-Value} \leq \text{Min-OV} \\ 1 & \text{IF O-Value} > \text{Min-OV} \end{cases}$$

So, POV represents in which percentage the new narrative satisfies the expected number of original new structures.

The Novelty of the Knowledge Structures (NKS) is defined as the ratio between the O-Value and the number of new structures (NNS).

$$\text{NKS} = \frac{\text{O-Value}}{\text{NNS}}$$

It represents which percentage of the new structures is original. In this way, if the O-Value is identical to the number of new structures the NKS is equal to 1 (100%). That means that all new structures satisfy the requirement of novelty.

A variant of the process of creation of knowledge structures is known as knowledge widening. It occurs when existing knowledge structures incorporate within its own structure unknown information obtained from the new story. This concept is inspired by Piaget's ideas about accommodation and assimilation (Piaget 1952). So, the model requires knowing the number of unknown information incorporated into the knowledge base; we refer to it as the number of new elements. So, in order to calculate the Proportion of Knowledge Widening (PKW) it is necessary to know the Number of New Elements (NNE) and an expected Minimum value of New Elements (Min-NE).

$$\text{PKW} = \begin{cases} \dfrac{\text{NNE}}{\text{Min-NE}} & \text{IF NNE} \leq \text{Min-NE} \\ 1 & \text{IF NNE} > \text{Min-NE} \end{cases}$$

Thus, PNS, POV, NKS and PKW provide information to evaluate how much new knowledge is generated.

Analysing the Story's Structure

We defined earlier that a story is recounted in a correct manner when it follows the classical Aristotelian structure: setup, conflict and resolution. The story's structure in this work is represented by the graphic of the curve of the dramatic tensions in the tale. Tensions represent conflicts between characters. When the number of conflicts grows the value of the tension rises; when the number of conflicts decreases the value of the tensions goes down; when the tension is equal to zero all conflicts have been solved. Thus, we analyse the characteristics of the graphic of tension to evaluate the presence of unexpected obstacles and how well recounted the story is. In this way, our evaluation model requires a mechanism to depict the dramatic tension in the tale.

There are four basic cases of graphics of tensions that we consider in this work: one complete curve (see figure 1-a); several complete curves (see figure 1-b); one incomplete curve (see figure 1-c); several incomplete curves (see figure 1-d). It is also possible to find combinations of these cases. A curve is defined as complete when its final amplitude is zero; that is, all tensions are resolved. By contrast, the final amplitude of an incomplete curve never gets the value of zero.

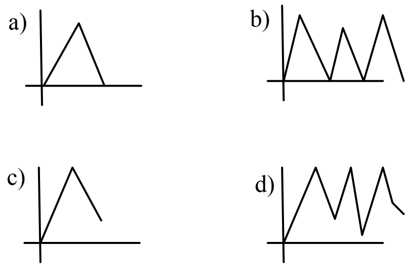

Figure 1. Examples of graphics of tensions.

The peak of a curve represents the climax of a narrative; if we have a sequence of curves we refer to the peak with the highest amplitude as the main climax. So, in a sequence, first the story reaches a situation with high levels of tensions, after that tensions start to loosen up and then they rise again; this cycle can be repeated. Each peak is a climax; each loosen up is a resolution of such a climax. We refer to the situation where a narrative has a resolution and then tensions start to rise again as *reintroducing-complications*.

We can find variations of the basic graphics of tensions we enumerated earlier. For example, the deepness of each valley in a sequence of incomplete curves might be different for each instance; in the same way, the amplitude of the peaks of sequences of complete or incomplete curves might change between them; and so on.

The difference between having a single curve and having a sequence of curves is that in the former there is only one high point in the story while in the latter we have two or more high points, i.e. new characters' obstacles are initiated reintroducing in this way complications.

The difference between a sequence of complete curves and a sequence of incomplete curves is that in the former all tensions are solved before new tensions arises; in the later new tensions emerge before the current ones are worked out. An incomplete curve is very similar to a complete curve if the fall of the tensions is close to 100% with respect to its peak, i.e. if the amplitude is close to the value of zero. On the other hand, if the fall of the tensions is close to 0% with respect to its peak, i.e. if the amplitude is close to the value of its peak, we practically do not have an incomplete curve. In this work we appreciate narratives that seem to end and then reintroduce new problems for the characters. In other words, we want narratives where all tensions are solved (complete curves) or are almost solved (incomplete curves with deep valleys) and then they rise again. This formula can be observed in several examples of narratives like films, television-series and novels (nevertheless, the model allows experimenting with different values of valley's profundity).

Thus, different graphics of tensions produce different characteristics in the narrative. We hypothesize that a story that includes more curves of tensions is more exciting than a story that includes fewer curves because the former reintroduces more complications. However, too many curves make the story inadequate. So, it is necessary to find a balance. In this way, our model requires to set a number that represents the perfect amount of complete curves that a story should comprise. We refer to this number as the Ideal Value of Complete Curves (Ideal-CC). So, because we can calculate the number of complete curves (Num-CC) in any new narrative and because we have defined an ideal number for them, it is possible to estimate how close the number of curves is to its ideal value. We refer to this number as the Proportion of Complete Curves (PCC):

$$\text{PCC} = \begin{cases} \dfrac{\text{Num-CC}}{\text{Ideal-CC}} & \text{IF NumCC} \leq \text{Ideal-CC} \\ 1 - \dfrac{\text{Num-CC}}{\text{Ideal-CC}} & \text{IF Ideal-CC} < \text{NumCC} \leq \text{Ideal-CC} \cdot 2 \\ 0 & \text{IF NumCC} > \text{Ideal-IC} \cdot 2 \end{cases}$$

It is important to explain how the NUM-CC is calculated. As it is going to be explained some lines ahead, a story must include at least one complete curve to be considered as properly recounted. But this curve itself does not reintroduce problems. The reintroduction of complications occurs when the current ones are sorted out and then new complications (i.e. new complete curves) emerge. In this way, NUM-CC only registers those complete curves that actually reintroduce new conflictive situations.

The process to calculate the incomplete curves is a little bit different. The goal is to calculate how close the set of incomplete curves are to its ideal value. Remember that too many curves or too few curves produce inadequate results. It is necessary to know the number of incomplete curves (Num-IC) and the Ideal Value of Incomplete Curves (Ideal-IC) to calculate the Proportion of Incomplete Curves (PIC):

$$\text{PIC} = \begin{cases} \dfrac{\text{Num-IC}}{\text{Ideal-IC}} & \text{IF NumIC} \leq \text{Ideal-IC} \\ 1 - \dfrac{\text{Num-IC}}{\text{Ideal-IC}} & \text{IF Ideal-IC} < \text{NumIC} \leq \text{Ideal-IC} \cdot 2 \\ 0 & \text{IF NumIC} > \text{Ideal-IC} \cdot 2 \end{cases}$$

Now, it is necessary to analyse each of the curves to see how close they are to its ideal value. One starts getting the amplitude of the first peak and the amplitude of the bottom part of its valley; the ratio between the valley and the peak indicates the percentage, with respect to its peak, that the valley needs to be expanded to reach zero. So, if the peak's amplitude is 10 and the valley's is 4, the valley needs to be expanded 40% to reach zero. The process is repeated for all incomplete curves. The summation of these results is known as the Summation of Incomplete Curves (SIC):

$$\text{SIC} = \begin{cases} \displaystyle\sum_{i=1}^{\text{Num-IC}} \frac{\text{Amplitude-Valley}_i}{\text{Amplitude-Peak}_i} & \text{If Num-IC} > \text{Ideal-IC} \\ & \text{If Num-IC} \leq \text{Ideal-IC} \\ \left[\displaystyle\sum_{i=1}^{\text{Num-IC}} \frac{\text{Amplitude-Valley}_i}{\text{Amplitude-Peak}_i}\right] + (\text{Ideal-IC} - \text{Num-IC}) & \end{cases}$$

Notice that, if the number of incomplete curves is minor to the ideal value of incomplete curves, the difference between them is added to the summation. So, the value of SIC represents how far the set of incomplete curves is from its ideal value. So, if SIC ≈ 0 the new narrative totally satisfies the requirement for reintroducing complications (all curves have deep valleys); if SIC ≈ Ideal-IC the valleys are so small that practically we do not have incomplete curves.

Now, given an Ideal Number of Incomplete Curves (Ideal-IC), it is possible to calculate in what percentage the amplitude of all incomplete curves is similar to its ideal value. We refer to this value as the Total Amplitude of Incomplete Curves (TAI), which is defined as follows:

$$\text{TAI} = \begin{cases} \dfrac{\text{Ideal-IC} - \text{SIC}}{\text{Ideal-IC}} & \text{IF SIC} \leq \text{Ideal-IC} \\ 0 & \text{IF SIC} > \text{Ideal-IC} \end{cases}$$

If SIC > Ideal-IC we have too many incomplete curves whose amplitudes do not provide useful information for the evaluation.

Regarding the recountal of a story, we consider that a narrative follows the classical Aristotelian structure when its graph of tension includes at least one complete curve, i.e. the tension at the beginning and at end of the story is zero, and at least once the value of the tensions between these two points is different to zero. So, in this project we analyse if the story under evaluation has an adequate opening and adequate closure in terms of tensions. A story has an adequate opening (A-Opening) when the tension in the story goes from zero at the beginning of the story to some value greater than zero at the first peak.

$$\text{A-Opening} = \frac{\text{Amplitude First Peak - Amplitude (t=1)}}{\text{Amplitude First Peak}}$$

In this way, because our goal is to have a continue tension growing from zero to the first peak, this formula indicates which percentage of this goal is achieved.

One common mistake, particularly between inexperienced writers, is to finish a story leaving loose ends. Thus, following Pérez y Pérez and Sharples, a story "should display an overall integrity and closure, for example with a problem posed in an early part of the text being resolved by the conclusion" (Pérez y Pérez and Sharples 2004). In this way, in order to have an Adequate Closure (A-Closure) all conflicts must be worked out at the end of the story. That is, the value of the tension in the last action must be equal to zero. So, it is necessary to perform a similar process to the one employed to calculate the incomplete curves: one needs to get the amplitude of the curve's main peak, the amplitude of the bottom part of the last valley, and then calculate in what percentage the tension goes down. If the final amplitude of the curve is zero, i.e. if it goes down 100%, the Adequate Closure is set to 1; if the curve goes down 30% the Adequate Closure is set to 0.3; and so on.

$$\text{A-Closure} = 1 - \frac{\text{Amplitude Last Valley}}{\text{Amplitude Main Peak}}$$

Calculation of Interestingness

Thus, our model employs the following characteristics:

- Proportion of new structures (PNS)
- Proportion of the Original Value (POV)
- Novelty of the Knowledge Structures (NKS)
- Proportion of Knowledge Widening (PKW)

- Adequate Opening (A-Opening)
- Adequate Closure (A-Closure)

- Proportion of Complete Curves (PCC)
- Proportion of Incomplete Curves (PIC)
- Total Amplitude of Incomplete Curves (TAI),

The first six characteristics (PNS, POV, NKS, PKW, A-Opening and A-Closure) are known as the core characteristics (CoreC); the last three are known as the complementary characteristics (ComplementaryC). This distinction emerges after talking to some experts in science of human communication that pointed out to us that a story can be interesting even if there are no reintroductions of complications (that is, even if there are no extra complete or incomplete curves). The experts agreed that the reintroduction of problematic situations might add interest to the story, but they are not essential to it. So, we decided that they would complement the evaluation of the core characteristics (a kind of extra points).

It is necessary to set a weight for each of the core characteristics. The sum of all weights must be equal to 1. Thus, the Evaluation of Interestingness (I) is equal to the summation of the value of each core characteristic (Core-C) multiplied by its weight (W):

$$I = \sum_{i=1}^{6} CoreC_i \cdot W_i$$

The Complement (Com) is equal to the summation of the value of each complementary characteristic multiplied by its complementary weigh (w). The sum of all complementary weights ranges from zero to 1.

$$Com = \sum_{i=1}^{3} ComplementaryCi \cdot wi$$

Thus, the total value of interest (TI) is giving by

$$TI = \begin{cases} 1 & \text{if } (I + Com) > 1 \\ I + Com & \text{if } (I + Com) \leq 1 \end{cases}$$

If we combined the values obtained from the correct recountal of a story and the reintroduction of complications, then we can calculate a parameter that we referred to as excitement (E):

$$E = \text{A-Closure} \cdot W + \text{A-Openning} \cdot W + PCC \cdot w + PIC \cdot w + TAI \cdot w$$

Thus, E assigns a value to the increments and decrements of tension during the story.

Implementation of the Prototype

We have implemented a prototype to test our model. Our prototype evaluates the interestingness of four stories generated by our storyteller. Details of our computer model for plot generation can be found in (Pérez y Pérez and Sharples 2001; Pérez y Pérez 2007). In this document we only mention two characteristics that are important to learn in order to understand how the prototype of the evaluator works:

1. Our plot generator employs a set of stories, known as the previous stories, to construct its knowledge base. Such narratives are provided by the user of the system. Any new story generated by the storyteller can be included as part of the previous stories.

2. As part of the process of developing a new story the storyteller keeps a record of the dramatic tension in the story. The following are examples of situations that trigger tensions: when the life of a character is at risk; when the health of a character is at risk; when a character is made a prisoner; and so on. Every tension has assigned a value. So, each time an action is performed the system calculates the value of all active tensions and records it. With this information the storyteller graphs the curve of tension of the story (see figure 3).

Now we explain some details of the implementation of the prototype for the evaluation of interestingness. The model includes several parameters that provide flexibility. The first step is to set those parameters. We start with the expected or ideal values: Minimal Value of New Structures (Min-NS), Minimum value of New Elements (Min-NE), Minimum Original Value (Min-OV), Ideal Value of Complete Curves (Ideal-CC) and Ideal Value of Incomplete Curves (Ideal-IC). To determine the value of these parameters we employ the previous stories as a reference. (The previous stories employed in this work were made long time before this project started. They represent well-formed and interesting narratives. So, they are a good source of information). The process works as follows. We select seven previous stories. With six of them we create the knowledge base; the 7th is considered a new story (as if it had been produced by our storyteller). Then, we analyse how many new structures, new elements, new original value structures, and new complete and incomplete curves are generated by the 7th previous story and record these results. We repeat the same process for each of the previous stories. Then, after eliminating the highest and lowest values, we calculate the means of each result obtained. Following this procedure we conclude that the parameters should be set as follows: Min-NS = 7; Min-NE = 4; Min-OV = 5; Ideal-CC = 1; Ideal-IC = 1. That is, in average each previous story generates seven new knowledge structures, four new elements, five original structures, one complete curve and one incomplete curve.

The next step is to set the weights. Based on empirical experience of experts in human communication, the weight of the generation of new knowledge is set to 50% and the weight of the correctness of the way the narrative is recounted is set to the other 50%.

The characteristics that define the generation of new knowledge are: Proportion of new structures (PNS), Proportion of the Original Value (POV), Novelty of the Knowledge Structures (NKS) and Proportion of Knowledge Widening (PKW). Table 1 shows their assigned weights. We considered Novel knowledge structures more important than Knowledge Widening structures. The correctness of the way the narrative is recounted is defined by the parameters A-Opening and A-Closure. Both are important and both received the same weight. Finally, the LS was set to 85%.

Regarding the complementary parameters and weights, they contribute with a maximum extra value of 10% distributed as follows: 5% for the complete curves and 5% for TAI. This decision is based on our own experience.

Core Characteristic	Weight
Proportion of new structures (PNS)	10
Proportion of the Original Value (POV)	10
Novelty of the Knowledge Structures (NKS)	15
Proportion of Knowledge Widening (PKW)	15
Adequate Opening (A-Opening)	25
Adequate Closure (A-Closure)	25

Complementary Characteristic	Weight
Proportion of Complete Curves (PCC)	5
Proportion of Incomplete Curves (PIC)	0
Total Amplitude of Incomplete Curves (TAI)	5

Table 1. Weights of the characteristics

Finally, if the value of the correct recountal of the story (A-Closure + A-Opening) does not reach at least 50% of its highest possible value, the story is considered as unsatisfactory. In this way we avoid evaluating stories that lack enough quality (the reader must remember that in this paper we do not evaluate coherence; that is a different part of the project. However, this constraint in the prototype helps to avoid processing pointless stories).

Testing the Model

To test our model our storyteller generated four narratives known as short-1, short-2, long-1 and long-2 (see Figure 2). Figure 3 shows their graphics of tension. The following lines describe the main characteristics of each narrative.

Short-1 lacks an introduction; it starts with a violent action. One gets the impression that everything occurs very fast. It is not clear what happens to the virgin once she escapes and has an accident. Also it is unclear the fate of the enemy.

Short-2 has a brief introduction and then the conflict starts to grow (the killing of the knight). The end is tragic and all tensions are sorted out.

Long-1 has a nice long introduction. The conflict between the princess and the lady grows nicely and slowly until it reaches a climax. However, at the end, we do not know the destiny of the characters. Who got the knight? So, the story has an inadequate conclusion.

SHORT 1 The enemy kidnapped the virgin The virgin laugh at the enemy The enemy attacked the virgin The virgin wounded the enemy The virgin ran away The virgin had an accident ***The End***	**SHORT 2** Jaguar knight was a citizen The artist prepared to sacrifice the jaguar knight The jaguar knight became free The jaguar knight fought the artist The artist killed the jaguar knight The artist committed suicide ***The End***
LONG 1 Jaguar knight was a citizen The princess was a citizen The princess was fond of jaguar knight The princess fell in love with jaguar knight The lady was in love with jaguar knight The princess got jealous of the lady The jaguar knight was in love with the princess The princess attacked the lady The lady wounded the princess The lady ran away The lady had an accident ***The End***	**LONG 2** Jaguar knight was a citizen The enemy was a citizen The enemy got intensely jealous of jaguar knight The enemy attacked jaguar knight The jaguar knight fought the enemy The enemy wounded jaguar knight The enemy ran away The enemy went to Texcoco lake The enemy did not cure jaguar knight The farmer prepared to sacrifice the enemy The enemy ran away The jaguar knight died because of its injuries ***The End***

Figure 2. Four computer-generated stories

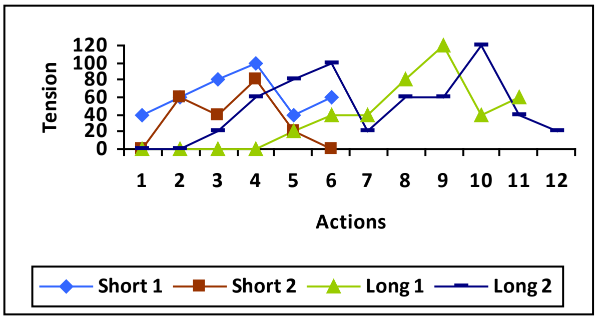

Figure 3. Graphics of Tensions for the four stories

Long-2 starts introducing the characters of the narrative. The tension grows fast until the story reaches a climax when the enemy wounded the knight. The tension decreases when the enemy decides to run off; however, it increases again when the enemy returns and the farmer attempts to kill him. Finally, he escapes again and the knight dies.

Based on our personal taste, our favourite narrative was short-2, then long-2, long-1 and finally short-1. We evaluated these four stories with our prototype. Table 2 shows the results; figure 4 shows the normalised values for the following features: generation of new knowledge, adequate closure, excitement and the total value of interestingness. Against our prediction, the system selected Long-2 as the most interesting story. There were two main reasons that explained why Long-2 beat short-2: 1) Long-2 generated more knowledge structures than Short-2; 2) Long-2's complements were slightly better evaluated than Short-2's. So, Short-2 obtained the second best result.

	Long-1	Long-2	Short-1	Short-2
PNS	10	8.57	4.29	4.29
POV	0	10	6	6
NKS	0	15	15	15
PKW	3.8	3.75	11.25	3.75
A-Op	25	25	15	25
A-Clo	13	20.83	10	25
I	**51.25**	**83.15**	**Unsatisfactory**	**79.04**
Com	3.4	3.15	3	1.65
TI	**54.6**	**86.30**	**Unsatisfactory**	**80.69**
E	41.4	48.98	28	51.65

Table 2. Numerical values of the evaluation.

In third place was Long-1; it did not produce any original structure and therefore its characteristic NKS got a value of zero. Also, its closure was poor. In last place was Short-1. The system evaluated Short-1 as an unsatisfactory story; i.e., it did not satisfy the minimum requirements of a correct recountal of a story (as we can see in table 2, the opening only got 15 points and the closing 10!). Nevertheless, we included the value of Short-1's closure and excitement in figure 4.

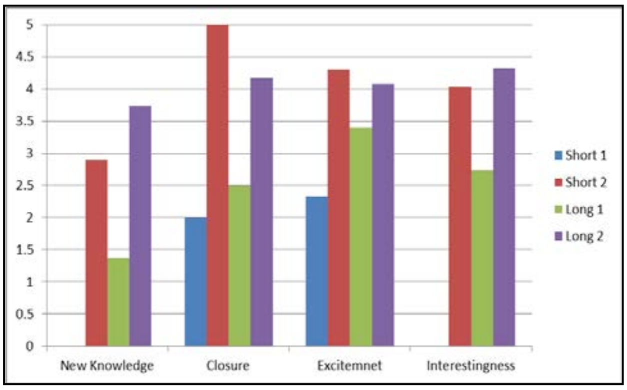

Figure 4. Graphics of the results of the evaluation.

We thought it could be interesting to compare the opinion of a group of subjects about the four stories under analysis to the results generated by our computer evaluator. Thus, we decided to make a survey by applying two questionnaires: 22 subjects answered questionnaire 1 and 22 subjects answered questionnaire 2; 25% were females and 75% were males; 13% had a PhD degree, 29% had a master degree, 27% had a bachelor's degree and 29% had other types of degree. We decided to group the narratives by their length. So, the first questionnaire included the two short narratives while the second questionnaire included the two long narratives. In both questionnaires we asked subjects to evaluate the adequateness of the closure and the interestingness of the stories. Subjects could rank each feature with a value ranging from 1 to 5, where 1 represented the lowest assessment and 5 the highest one. Figure 5 shows the results of the evaluation of interestingness. Short-2 was considered the most interesting narrative; Long-2 seemed to be in the second position followed close behind by Short-1 and Long-1. These last results were not conclusive. We were surprised that Short-1 was not clearly in the last position. We speculated that human capacity of filling gaps when reading a narrative might contribute to this result. Although our computer agent calculated a higher evaluation to Long-2 than to Short-2, both stories got a very similar score (the difference was less than 6%; c.f. with the score of Long-1). So, we felt that subjects' opinion about these two narratives was close to the results we obtained from our computer prototype. However, by contrast, our system clearly rejected Short-1 and left Long-1 in a clear third position while subjects' evaluation was unclear.

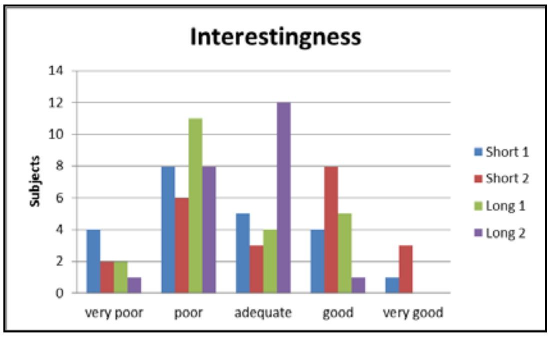

Figure 5. Subjects' evaluation of interestingness.

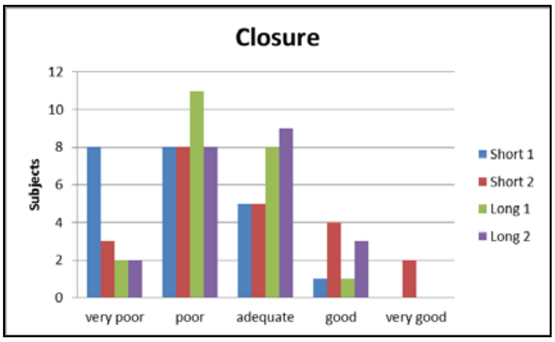

Figure 6. Subjects' evaluation of closure.

Figure 6 shows the results for the evaluation of closure. Subjects ranked Short-2 as the story with the best closure, followed by Long-2, Long-1 and Short-1. There was a total coincidence between the computer agent evaluation and the human evaluation.

Discussion and Conclusions

This paper reports a computer model for the evaluation of interestingness. It is part of a bigger project that attempts to evaluate the interestingness, coherence and novelty of computer generated narratives. The model presented in this paper emphasises two properties: generation of new knowledge and the correctness of the recountal of a story. Regarding the generation of new knowledge, we developed a process to calculate how much new information was produced by a computer generated story. In the same way, motivated by Piaget ideas about accommodation and assimilation, we defined two different types of knowledge structures: new knowledge and widening knowledge. We went further by identifying those new knowledge structures which were very different to the existing ones. Regarding the recountal of a story, we worked on previous research that had illustrated the relation between the dramatic tension of a story and its interestingness. In this work we expanded this idea by analysing the opening and closure of a story, and verifying if new obstacles were introduced along the plot. Thus, we have been able to create a model that allows a computerised agent to perform a detailed evaluation of the stories it produces.

The implementation of our prototype has allowed testing the ideas behind the model. We are satisfied with the results. But we are more excited about what we are expecting to achieve with this new characteristic. The capacity to evaluate its own outputs allows a storyteller to distinguish positive and negative qualities in a narrative and therefore to learn from its own creative work; it also incorporates the possibility of evaluating and learning from narratives generated by other systems. In our case, we expect that our storyteller agent will be able to determine autonomously which stories, either produced by itself, by other systems or by humans, should become part of its set of previous stories. That is our next goal.

We have compared the results produced by our automatic evaluator to the results obtained from a questionnaire answered by a group of 44 human evaluators. In general terms, the results obtained from both approaches were similar. This suggests that the subjects that answered the questionnaire might consider acceptable the outputs produced by our system. Nevertheless, it is intriguing why the story Short-1 got a relative high evaluation from the subjects. We need to analyse further this result and see if we require adjusting our model.

As it has been showed in this work, we consider the generation of new knowledge an important characteristic of computational creativity. So, it is not enough to evaluate the creative-product and/or the creative-process, as it has been suggested by some researchers. We believe that it is also necessary to considerate how much such products and/or processes modify the characteristics of the storyteller agent and the evaluator agent (that in our case is the same). So, any evaluation process must consider this aspect. This idea is inspired by the fact that, any creative act performed by humans will influence their future creative acts. We need to represent this feature in our computer models.

The qualities that make a story interesting, coherent and novel are complex and many times overlap each other. Our work seems to illustrate part of this overlapping complexity. For example, the generation of new structures might be employed to evaluate novelty; the adequate opening and closure might be employed to evaluate coherence; however, at the same time, they are essential elements to evaluate interestingness. This seems to confirm our idea that a general model of evaluation of narratives at least must contemplate coherence, novelty and interestingness. We are currently working on producing such a general model.

Hopefully this model will be useful not only for those working in plot generators but also to those researchers working in similar areas (e.g. interactive fiction). We are aware that many features not considered in this work might contribute to make a story interesting (e.g. suspense, intrigue). As mentioned earlier, human evaluation is very complex and we do not comprehend yet how it works. Nevertheless, we expect this research contributes to understand better the mechanisms behind it.

References

Colton, S. 2008. Creativity versus the perception of creativity in computational systems. Creative Intelligent Systems: Papers from the AAAI Spring Symposium. 14–20.

Deckers, L. 2005. Motivation Biological, Psychological, and Enviromental. Pearson.

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

Lodge, D. 1996. The practice of writing: essays, lectures, reviews and a diary. London: Secker & Warbug.

Pease, A.; Winterstein, D.; and Colton, S. 2001. Evaluating machine creativity. In Weber, R. and von Wangenheim, C. G., eds., *Case-based reasoning: Papers from the workshop programme at ICCBR 01Vancouver*. Canada 129–137.

Peinado, F.; Francisco, V.; Hervás R. and Gervás, P. 2010. Assessing the Novelty of Computer-Generated Narratives Using Empirical Metrics. *Minds and Machines*. 20(4):565-588.

Pereira, F. C.; Mendes, M.; Gervás, P., and Cardoso, A. 2005. Experiments with assessment of creative systems: An application of Ritchie's criteria. In Gervás, P. Veale, T. and Pease, A., eds., *Proceedings of the workshop on computational creativity*, 19th international joint conference on artificial intelligence.

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

Pérez y Pérez, R. and Sharples, M. 2004. Three Computer-Based Models of Storytelling: BRUTUS, MINSTREL and MEXICA. *Knowledge Based Systems Journal*. 17(1):15-2.

Pérez y Pérez, R. 2007. Employing Emotions to Drive Plot Generation in a Computer-Based Storyteller. Cognitive Systems Research 8(2): 89-109.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17:76–99.

Perez y Perez, R., Ortiz, O., Luna, W. A., Negrete, S., Peñaloza, E., Castellanos, V., and Ávila, R. 2011. A System for Evaluating Novelty in Computer Generated Narratives. In *Proceedings of the Second International Conference on Computational Creativity*, México City, México, pp. 63-68

Piaget, J. 1952. The Origins of Intelligence in Children.

London: Routledge and Kegan Paul, 1936 (French version published in 1936, translation by Margaret Cook published 1952).

Sharma, M., Ontañón, S., Mehta, M. and Ram, A. 2010. Drama Management and Player Modeling for Interactive Fiction Games. *Computational Intelligence Journal*, 26(2): 183-211.

Weyhrauch, P. 1997. Guiding Interactive Drama. PhD Dissertation, School of Computer Science, Carnegie Mellon University.