

How to evaluate a computer narrator that evaluates itself

Rafael Pérez y Pérez^{1,2}

¹ Universidad Autónoma Metropolitana, Cuajimalpa, México

² Center for Digital Narrative, University of Bergen, Norway

Abstract

MEXICA is a narrative generator capable of evaluating the novelty, interestingness, and coherence of its own narratives. Recently, I developed a methodology to assess whether MEXICA effectively evaluates story coherence. This involved creating a questionnaire for a group of people to provide feedback on MEXICA's performance. I also conducted several meetings with computer science experts to gather their insights on the questionnaire. These discussions were invaluable, helping me refine both the questionnaire and how to apply it. This article details the questionnaire, key expert feedback, and the enhanced methodology for assessing MEXICA's mechanism to evaluate story coherence. This experience may benefit others in the community.

Keywords

Story generation, automatic assessment, story coherence

1. Introduction

MEXICA is a computer model of creative writing [1]. This is a long-term project whose main goal is to provide a framework for the reflection on ideas about how the creative process works and, in this way, contribute to its understanding. The system represents a plot as a sequence of actions described using simple texts, such as "I went-garden." Once the story is complete, MEXICA employs templates to produce more readable outputs, like "I went to the garden every evening after dark." In Section 2, box A there is an example of a story generated by MEXICA. In the past years, I have been developing mechanisms that allow my system to evaluate the novelty, interestingness and coherence of its own outputs (besides MINSTREL [2] that evaluates novelty, I am not aware of any other narrative generator that includes a similar self-evaluation process). Because of the difficulty of this endeavor, I divided the project into two parts. First, I developed a model to evaluate the novelty and interestingness of a tale [3]. Next, I started to work on the analysis of coherence. This paper focuses in this second aspect. The analysis of narrative coherence is complex. There are several characteristics to consider. For example, the adequate structure of the narrative, the congruence between the behavior of the characters and their goals, their personalities, their roles in the story, the social norms represented in the story, the cultural context (sometimes, what it makes sense in a specific culture might not make sense in others), among many others. That is why most narrative generators employ predefined story-structures, usually represented as goal-schemas or rules, to ensure the generation of logical plots (see [4] and [5]). MEXICA avoids the use of those types of structures and therefore it requires a mechanism to assess its stories. In this project, a story is defined as a sequence of actions where conflicts are introduced, developed, and finally sorted out. Thus, a story is coherent when all the events in the plot contribute to progress the conflict in a story or to preserve the coherence of the tale. This is known as the Model of Thematic Consistency (MTC) [6]. The model works as follows. To progress a conflict or to preserve the coherence of the narrative involves establishing what we refer to as functional "bindings" between two or more actions. Two actions are bounded when the consequences of the first action create the story-context necessary to perform the second action. We claim that in a narrative with thematic consistency, all actions must be

¹International Workshop on Computational Models of Narrative, 2025, Geneva

✉ rperez@cua.uam.mx

ORCID ID 0000-0001-8985-5897



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

bounded to at least one other of the actions in the tale. In other words, unbounded actions do not fulfil any purpose in the tale. Those, the main goal of the model is to identify those bindings. The Model contemplates four types of bindings: those related to common-sense knowledge, those that establish emotional relationships between friends and family members, those that trigger conflicts between characters, and those that cause two characters to be in the same place so they can interact with each other. Once bindings are identified, actions that share similar features are grouped. Groups represent sequences of actions that describe a particular situation. The MTC establishes that all groups in a tale must be connected or “chained” to another group. The current version of the model does not consider actions that might have other functionalities, e.g., building a specific atmosphere for the reader. I implemented all these ideas in MEXICA, so the system now has the capacity to evaluate the outputs it generates. Based on this evaluation, the program can, between other things, improve its stories by eliminating those actions that are not useful to progress the plot, and by reorganizing the text into sections that help the story to flow more smoothly. So far, the prototype has produced interesting results. So, the next step is to assess it. A recurrent topic in conferences in computational creativity is how to evaluate a creative agent. And most of the time, the discussions end in disagreements. Nobody seems to have a clear idea about how to solve this problem. In the literature, we can find some works aimed at evaluating models of the creative process (e.g. [7], [8]) and works aimed at evaluating computer generated narratives (e.g. [9], [10], [11], [12], [13]). Given the complexity of the task and the significant variations among narrative generators, achieving a consensus on the solution to this challenge remains elusive. So, I did what everybody does. I designed a questionnaire to ask human judges to assess my system. However, by contrast with previous similar experiences, I made two changes. Rather than focusing in the output (e.g. asking a group of subjects to evaluate an improved story), the questionnaire aims to get feedback from human judges about the decisions that the system takes to eliminate actions during the evaluation of the thematic consistency. The goal is to analyze if the way that MEXICA reasons, makes sense to these judges. And, inspired by some psychological methodologies, I decided to ask a group of experts their opinion about the questionnaire and the way I was planning to use it. Thus, I arranged meetings with Mike Sharples, Emeritus Professor in Educational Technology from the Open University in the UK, Professor Pablo Gervás, from the Universidad Complutense de Madrid in Spain, and Professor Dan Ventura, from the Brigham Young University in USA. Because these discussions allowed improving the methodology and highlighting interesting research aspects not considered before, sharing this experience might be useful for the community. Therefore, this paper has two primary objectives: to present the methodology I intend to use to evaluate MEXICA's ability to assess thematic consistency in stories, and to share some valuable insights gained from the group of experts.

2. The questionnaire

Thanks for participating in this study! Below you will find three texts, each inside a box labelled A, B, and C. These are the contents of each box:

- Box A contains a short story.
- Box B shows suggestions to eliminate those actions in the original story (in box A) that do not contribute to the development of a logical plot.
- Box C shows how text in box A looks like after the corrections suggested in box B are made, and how this story can be divided in sections.

A)

Original
short-story

B)

Suggestions to
eliminate some actions
to improve the original
short-story

C)

How the original story looks
like after the corrections have
been made and the story is divided
in sections

On the next page, you will find five statements. For each statement, you must choose one of the following options, depending on how much you agree or disagree with the statement: 1) Strongly disagree, 2) Disagree, 3) I am not sure, 4) Agree, 5) Strongly agree. Finally, if you wish you can write some comments about this study.

Instructions.

1. Read the story in box A.
2. Analyse the suggestions provided in box B to improve the story in A.
3. Read the improved version of the story in box C.
4. Indicate if you agree or disagree with each of the 5 statements using the scales above.
Please, indicate the reasons for your answer. If you wish, write some comments about this study.
5. Repeat the same process for a second story.

A)

B)

Original story Although it seemed like any other day at the hospital, the doctor had a feeling that something different would happen that day. I went to the garden every evening after dark. While I was walking, I had a terrible accident that produce severe injuries. Suddenly, the doctor bumped into me. He was surprised! The doctor realized that while I was walking, I had a terrible accident that produce severe injuries. The doctor made a potion that I drank and started to recover.	The following actions ought to be eliminated because they do not have a function in the narrative: -> ALTHOUGH IT SEEMED LIKE ANY OTHER DAY AT THE HOSPITAL, THE DOCTOR HAD A FEELING THAT SOMETHING DIFFERENT WOULD HAPPEN THAT DAY. -> I WENT TO THE GARDEN EVERY EVENING AFTER DARK. -> I WANTED TO GO TO THE FAMILY HOUSE. As a result of these changes, the following actions no longer have a function in the narrative and also must be eliminated: -> SUDDENLY, THE DOCTOR BUMPED INTO ME. HE WAS SURPRISED! Although it seemed like any other day at the hospital, the doctor had a feeling that something different would happen that day. I went to the garden every evening after dark. While I was walking, I had a terrible accident that produce severe injuries. Suddenly, the doctor bumped into me. He was surprised! The doctor realized that while I was walking, I had a terrible accident that produce severe injuries. The doctor made a potion that I drank and
--	--

I wanted to go to the family house. Paul took from the doctor some of his belongings. I realized that Paul took from the doctor some of his belongings. My irritation pushed me to look for Paul and to demand an explanation. As a result of my accusations, Paul was sent to jail.	started to recover. I wanted to go to the family house. Paul took from the doctor some of his belongings. I realized that Paul took from the doctor some of his belongings. My irritation pushed me to look for Paul and to demand an explanation. As a result of my accusations, Paul was sent to jail.
---	---

C)

The corrected story can be divided in 3 sections. This is the suggested final version: I While I was walking, I had a terrible accident that produce severe injuries. The doctor realized that while I was walking, I had a terrible accident that produce severe injuries. The doctor made a potion that I drank and started to recover. Paul took from the doctor some of his belongings. I realized that Paul took from the doctor some of his belongings. II My irritation pushed me to look for Paul and to demand an explanation. III As a result of my accusations, Paul was sent to jail. The end.
--

1. The story in Box A includes some actions that are not consistent with the rest of the text.

1) Strongly disagree	2) Disagree	3) I am not sure	4) Agree	5) Strongly agree
----------------------	-------------	------------------	----------	-------------------

Please, clarify the reason for your decision.

2. Box B suggests deleting some actions from the story in Box A. The actions that were crossed out are indeed inadequate for the proper development of the plot.

1) Strongly disagree	2) Disagree	3) I am not sure	4) Agree	5) Strongly agree
----------------------	-------------	------------------	----------	-------------------

Please, clarify the reason for your decision.

3. The coherence of the story in Box A improves when the recommendations proposed in Box B are carried out.

1) Strongly disagree	2) Disagree	3) I am not sure	4) Agree	5) Strongly agree
----------------------	-------------	------------------	----------	-------------------

Please, clarify the reason for your decision.

4. All the actions in the narrative in Box C follow a logical progression.

1) Strongly disagree	2) Disagree	3) I am not sure	4) Agree	5) Strongly agree
----------------------	-------------	------------------	----------	-------------------

Please, clarify the reason for your decision.

5. The story in box C has been divided into sections. To organize the narrative into sections helps the story to flow more smoothly.

1) Strongly disagree	2) Disagree	3) I am not sure	4) Agree	5) Strongly agree
----------------------	-------------	------------------	----------	-------------------

Please, clarify the reason for your decision.

6. Would you like to provide any comments about this study?

3. Meetings with the experts

I met Prof. Mike Sharples and his colleague Minji Xu in London at the beginning of November 2024. During our discussions about the main characteristics of the MTC and the questionnaire (see section 3), Mike and Minji spontaneously decided to cross out actions they deemed inadequate to maintain the thematic consistency of the story presented in section 2. Mike's results mostly differed from those produced by the program, as he presumed facts not explicitly mentioned in the narrative. However, he acknowledged that the actions excluded by MEXICA could also be valid. In contrast, Minji's removed actions closely mirrored those eliminated by the system, with only one exception. After explaining the model's logic, she expressed that the system had made a better decision. Consequently, Mike and I decided to divide the evaluation process into two phases. During Phase 1, participants are asked to read a story generated by MEXICA and cross out actions that do not contribute to the plot's development (i.e., those that do not meet the criteria of thematic consistency). They must also explain their reasoning. The results produced by humans are then compared with those generated by the system. The goal is to determine whether some people follow a process similar to the MTC and to gain insights into how people interpret the same plot. During Phase 2, a group of human judges will answer three questionnaires, referred to as Q1, Q2, and Q3. Each questionnaire will contain the same story in Box A, generated by MEXICA. In Q1, MEXICA will decide which actions to cross out for the story in Box B. In Q2, a human judge will make these decisions. In Q3, the actions in Box B will be randomly crossed out. The objective is to determine whether human judges agree with MEXICA's decisions and to compare the results across the three questionnaires.

In November 2024, I had the opportunity to meet Prof. Pablo Gervás in Madrid, shortly after seeing Mike Sharples. Pablo invited me to present the MTC to his research group at the Faculty of Computer Science, Complutense University of Madrid. The meeting was attended by five professors. During the presentation, I shared the narrative generated by MEXICA, as detailed in section 2, and asked the attendees to identify and cross out actions that did not contribute to the plot's progression. I then explained the key features of the MTC and the evaluation mechanism I planned to implement, i.e., phases 1 and 2. We compared the actions crossed out by MEXICA with those crossed out by the attendees. Interestingly, none of the human responses matched each other or MEXICA's exactly, though some were similar. This led to a lengthy and insightful discussion, revealing that each participant used different valid criteria for eliminating story elements. Here are a couple of examples

Two of the attendees decided to cross out the first sentence of the story, "Although it seemed like any other day at the hospital, the doctor had a feeling that something different would happen that day." Similarly, two other participants decided to cross out the second sentence, "I went to the garden every evening after dark." Only one of them deleted both (MEXICA also proposes to delete both). Based on the comments made by my colleagues during the discussion, I suspected that there

were two reasons for these results. For some of my colleagues, the opening sentence of a story is crucial as it sets the tone and shapes expectations for the rest of the tale, making it indispensable. Similarly, the description of a dark garden in the second action was viewed by some subjects as creating a narrative atmosphere that evoked a series of unconscious associations, influencing the perception of the rest of the story. Therefore, this description served a purpose and could not be omitted. However, for the rest of the attendees and MEXICA, these first two events were seen as unrelated to the subsequent plot developments.

Another significant point of discussion was the sequence: “While I was walking, I had a terrible accident that resulted in severe injuries. The doctor realised that while I was walking, I had a terrible accident that resulted in severe injuries.” Three participants felt that the information was repetitive. However, for MEXICA, it was not, as the first event describes the accident itself, while the second describes the moment the doctor became aware of the accident. Two participants appeared to agree with MEXICA.

As each participant explained their reasons for crossing out or keeping a particular event, the others concurred with the reasoning provided. This analysis highlighted that the interpretation of an action's usefulness to maintain the thematic consistency in a narrative can vary from person to person. The participants agreed that there are multiple ways to modify the story to improve it, and that MEXICA's approach was adequate. They also found the questionnaire to be an appropriate method for evaluating the Model of Thematic Consistency. Finally, several suggestions for improving MEXICA emerged. For instance, instead of deleting actions that do not contribute to the plot's development, the system could modify the story to utilize them. Additionally, the possibility of incorporating the MTC, or at least part of it, during the process of narrative generation was discussed.

I met Prof. Dan Ventura in Mexico City at the end of October 2024 and communicated with him via email towards the end of November 2024. Our discussions mainly centered around the questionnaire. His primary criticism was that, although the questionnaire could be a useful tool, it is challenging to formulate the right questions, find the best wording, and account for varying interpretations. Factors such as knowledge of narrative theories can influence responses. He also mentioned that the texts used to describe events might shape perceptions of how appropriate the actions are for the narrative. These are common problems when using questionnaires. Dan also made two recommendations: first, to get similar feedback, have participants identify unnecessary story elements (i.e., create box B in the questionnaire), and second, have them revise the story (i.e., create box C in the questionnaire). Creating box B corresponds to what I previously described as phase 1, while creating box C would require incorporating a new phase where participants must organize a given tale into sections.

4. Discussion

Based on my interviews with the group of experts, I plan to organize the assessment of the Model of Thematic Consistency as an online three-part process, with each part answered by different participants.

Part 1 (creating box B). A narrative generated by MEXICA will be presented to a group of human judges. They will be asked to cross out actions that they believe do not contribute to the thematic consistency of the story and to explain their reasoning. Then, they will view the same story, but with the actions crossed out by MEXICA. The participants will be able to read MEXICA's explanations for each decision to eliminate or not an action (subjects will be able to read sentences of the type “Is X a good reason to remove action Y?”). Finally, they will be asked to indicate whether they agree or disagree with the system's decisions and whether they would like to make any changes to the actions they crossed out.

Part 2 (box C). A narrative generated by MEXICA will be shown to a group of human judges, with all actions that do not contribute to the thematic consistency already removed. Participants will be asked to organize this story into sections to improve its flow, and then to explain their reasoning. There will also be the option of choosing not to divide the narrative into sections. Next, the

participants will see the same story, but this time organized by MEXICA. They will indicate whether they agree or disagree with the system's decision, and whether they would like to make any changes to the sections they proposed earlier.

Part 3 (the questionnaires). A group of human judges will be presented with three different questionnaires, referred to as Q1, Q2, and Q3, in random order. In all three questionnaires, the story in Box A will be generated by MEXICA. In Q1, MEXICA will determine which actions to cross out for the story in Box B and how to organize the story into sections in Box C. In Q2, a human judge will make these decisions. In Q3, the actions in Box B will be randomly crossed out, and the story will be randomly organized into sections in Box C.

Since the evaluation of the Model will be conducted online, I will design the interface required for each phase. I anticipate that in parts 1 and 2, a small group of human evaluations will align with those made by MEXICA, reflecting the diversity observed in human responses so far. Additionally, I expect that the majority of human judges will agree with MEXICA's decisions, both when crossing out actions and when organizing the tale into sections. In part 3, I predict that most human judges will assign high scores to questions in Q2, medium scores to questions in Q1, and low scores to questions in Q3.

One key aspect to examine is the variety of reasons individuals provide for crossing out or retaining an action. The study proposed in this paper aims to uncover insights into what influences participants to make decisions in one direction or another. Based on discussions with the experts, a significant factor appears to be how each person pictures the narrated events. For instance, in the sequence *While I was walking, I had a terrible accident that caused severe injuries. Suddenly, the doctor bumped into me. He was surprised!*, a reader imagining a scene with visible bloody injuries would naturally infer that the doctor immediately perceives the seriousness of the situation upon encountering the protagonist. Conversely, if the reader envisions an accident resulting in internal injuries, it might become less clear whether the doctor would recognize that an accident occurred before he arrived. This distinction shapes whether it is necessary for the narrative to explicitly state that the doctor realized about the accident. Thus, some possible reasons why individuals might eliminate events include perceiving redundancy in the information provided by an action, its irrelevance to the story's plot, or its inability to contribute to the development or resolution of conflicts within the narrative, among other considerations.

Final thoughts. Gaining firsthand insights from a group of computer scientists about my plans to evaluate the Model of Thematic Consistency was an inspiring experience. Including specialists from other fields such as psychology, creative writing, and narratology in these discussions will undoubtedly enrich them. The advice these experts offer is invaluable. The discussions on whether certain actions should be crossed out highlighted the diversity of possible interpretations of a story. From a computational perspective, these discussions reveal that multiple models can be developed for evaluating thematic consistency. This leads to new research questions: what modifications does the current Model of Thematic Consistency need to generate results similar to those produced by different human groups? What situations are impossible to represent with our current models? The limitations of our systems are evident, and this type of research would profoundly improve them. Recently, large language models (LLMs) have begun incorporating processes that seem to enable reflection on the narratives they generate, in a manner reminiscent of MTC. Companies like Anthropic have disclosed the use of hybrid systems to accomplish this, such as their Claude 3.7 Sonnet model. However, detailed information about the inner workings of these systems remains scarce.

While trying to figure out the best way to evaluate my model, I discovered new research paths. The type of evaluation proposed in this article can generate valuable information about the mechanisms humans use to interpret and judge a narrative, which might be useful in fields like psychology and narratology. We should encourage the creation of specialized forums to discuss how to creatively evaluate the processes and outputs of our computer models of narrative generation. Interesting developments will undoubtedly arise from those meetings. What are your suggestions to enrich my model to evaluate thematic consistency?

Acknowledgements

This research was partially financed by the National Council of Humanities, Science and Technology in México, project CF-2023-I-312, and by the Research Council of Norway through its Centers of Excellence Scheme, Project No. 332643.

Declaration on Generative AI

During the preparation of this work, the author used Google Translate for automatic text translation (few sentences) and Microsoft Copilot to paraphrase and reword. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

- [1] R. Pérez y Pérez, M. Sharples, MEXICA: a computer model of a cognitive account of creative writing, *Journal of Experimental and Theoretical Artificial Intelligence*, 13 (2001) 119-139. doi: 10.1080/09528130118867.
- [2] S. R. Turner, MINSTREL: A Computer Model of Creativity and Storytelling, PhD thesis, University of California, 1993.
- [3] R. Pérez y Pérez, The Three Layers Evaluation Model for Computer-Generated Plots, in: *Proceedings of the fifth International Conference on Computational Creativity*, Ljubljana, Slovenia, 2014, pp. 220-229.
- [4] M. Sharples, R. Pérez y Pérez, *Story Machines. How Computers Have Become Creative Writers*, Routledge, London, 2022. doi: 10.4324/9781003161431.
- [5] R. Pérez y Pérez, M. Sharples, *An Introduction to Narrative Generators: How Computers Create Works of Fiction*, Oxford University Press, Oxford, 2023. doi: 10.1093/oso/9780198876601.001.0001.
- [6] R. Pérez y Pérez, Computer models as tools for the reflection on ideas: a model for evaluating the thematic consistency of a narrative, in: *Proceedings of the 15th International Conference on Computational Creativity*, University of Jönköping, Sweden, 2024, pp. 75-82.
- [7] G. Ritchie, Some empirical criteria for attributing creativity to a computer program, *Minds and Machines* 17 (2007) 76–99.
- [8] A. Jordanous, A Standardised Procedure for Evaluating Creative Systems: Computational Creativity Evaluation Based on What it is to be Creative, *Cognitive Computation*, 4 (2012) 246-279.
- [9] Pereira, F. C.; Mendes, M.; Gervás, P., and Cardoso, Experiments with assessment of creative systems: An application of Ritchie's criteria, In *Proceedings of the 19th international joint conference on artificial intelligence*, 2005.
- [10] M. Riedl, M. Young, An Objective Character Believability Evaluation Procedure for Multi-agent Story Generation Systems, *Lecture Notes in Computer Science*, vol 3661 (2005), https://doi.org/10.1007/11550617_24
- [11] J. Rowe, S. McQuiggan, J. Sabourin, D. Marcey, J. Lester, STORYEVAL: An empirical evaluation framework for narrative generation, *AAAI Spring Symposium*, 2009, 103-110.
- [12] F. Peinado, V. Francisco, R. Hervás, P. Gervás, Assessing the Novelty of Computer-Generated Narratives Using Empirical Metrics, *Minds and Machines*, 20 (2010) 565-588.
- [13] C. Purdy, X. Wang, L. He, M. Riedl, Predicting generated story quality with quantitative measures, In *Proceedings of the Fourteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment Article 14*, (2018) 95–101.