Objective Metrics for Interactive Narrative

SZILAS, Nicolas, ILEA, Ioana

Abstract

This paper describes, implements and assesses a series of user-log indicators for automatic interactive narrative evaluation. The indicators include length and duration, diversity, renewal, choice range, choice frequency, and choice variety. Based on a laboratory experiment with players, a significant positive correlation has been observed between two indicators and some aspects of the interactive narrative experience measured by validated scales based on questionnaires.

Reference

SZILAS, Nicolas, ILEA, Ioana. Objective Metrics for Interactive Narrative. In: A. Mitchell, C. Fernándes-Vara & D. Thue. *Interactive Storytelling. 7th International Conference on Interactive Digital Storytelling, ICIDS 2014*. Berlin, Heidelberg : Springer, 2014. p. 91-102

DOI : 10.1007/978-3-319-12337-0_9

Objective Metrics for Interactive Narrative

Nicolas Szilas and Ioana Ilea

TECFA, FPSE, University of Geneva, CH 1211 Genève 4, Switzerland
{Nicolas.Szilas,Ioana.Ilea}@unige.ch

Abstract. This paper describes, implements and assesses a series of user-log indicators for automatic interactive narrative evaluation. The indicators include length and duration, diversity, renewal, choice range, choice frequency, and choice variety. Based on a laboratory experiment with players, a significant positive correlation has been observed between two indicators and some aspects of the interactive narrative experience measured by validated scales based on questionnaires.

Keywords: interactive storytelling, interactive narrative, interactive drama, evaluation, metrics, analytics.

1 From Theory-Driven to Player Data-Driven Interactive Narrative Design

Research in Interactive Narrative has only recently focused on the issue of evaluation. This is a central issue for the sake of scientific acceptance. It is also essential to guide the iterative design of a given system, to compare systems with each other and to "articulate overarching research directions for the field overall" [26].

Early systems were evaluated by merely showing one or more examples produced by the system [26], either manually chosen [3] or randomly selected [21]. This raises two issues:

- It is scientifically insufficient for it lacks enough examples to reach the threshold of statistical significance. Thus, neither proper evaluation criteria are set, nor independent observer is involved.
- The outputted story is evaluated, which is right for story generation systems but possibly inappropriate for systems targeting a moment-to-moment interaction, such as Defacto, IDtension, Façade, Thespian, U-DIRECTOR, Scenejo, etc. From a theoretical point of view, the interactive narrative experience is fundamentally different from the outputted story and the evaluation of the latter may not be equivalent to that of the former [1], [22].

The first above issue can be properly solved by adopting a more rigorous protocol for story evaluation. For example, Michael Young and his colleagues have a long standing experience in evaluating story generation algorithms by asking a pool of subjects to rate stories produced in various conditions (e.g. [4]). However, the second

A. Mitchell et al. (Eds.): ICIDS 2014, LNCS 8832, pp. 91–102, 2014.

issue mentioned above illustrates the necessity to include users (not readers) in the evaluation, because interactivity is the finality of most interactive narrative systems.

In the rest of the paper, we will only consider the case of an interactive experience that is in contrast to generated story.

Recent research has proposed a range of techniques to evaluate interactive narratives: post-experience questionnaires, in-game questionnaires, and interviews. These methods have their respective advantages and drawbacks. Post-experience questionnaires have a long tradition in user studies. They consist of building aggregated scales from several questions in order to describe one aspect of the experience (e.g. immersion, characters' believability, etc.). When validated, these scales can be used to compare different instances of a system and different systems, as for the IRIS scales [24]. In-game questionnaires help to better measure the user experience while paying, rather than user interpretation after playing [16], but it may disrupt the experience. All questionnaire-based methods suffer from the known limits as difference between the real experience and the recollected experience, social desirability bias, bias related to the context of questionnaire answering and the literacy level, etc. Less quantitative methods based on free or semi-structured interviews partially alleviate the problem. For example, Façade was evaluated via a protocol involving eleven users who played with the interactive narrative and who were then interviewed by researchers [18]. The transcripts were later analyzed by psychologists who extracted some emerging themes and concepts. This kind of approaches provides a better understanding of user perception.

All above-mentioned evaluation methods are costly, because they require organizing a playing session, setting questionnaires or interviews, processing the data. There exists a less costly approach that, surprisingly, has been rarely employed so far in the field of interactive narrative: computer traces (logs) extracted from playing sessions. Typically, as we deal with interactivity, it should be commonplace to measure the number of choices given to the player on average. However, this measure is only rarely adopted (but see [5], [20]).

In this paper, we claim that such objective measures, automatically extracted from real user experiencing the system, should be researched and implemented. Contrary to the above methods, such metrics may deliver quantitative feedback with limited requirements in terms of experiment setting and manual data analysis. In particular, if the interactive narrative is online, only a few dozens of users are required to play with a certain version of the system to get an immediate measure concerning this version. As we will show, what can be measured by using this kind of methods is limited. Therefore, this approach cannot replace the other evaluations. Instead, our claim is that a certain type of user feedback data is potentially there within the implemented systems, but it is largely unexploited today. Metrics for interactive narratives carry the goal of exploiting these data to provide to the system designers and the research community in general, more data to compare and improve systems.

This approach is related to the use of game analytics in research and industry [17]. Effectively, we are looking for metrics that help understanding the user experience, in what may be viewed as a specific type of game. The main two differences are: 1) no business goal is concerned but design and research goals only; 2) We are not concerned with the improvement of a specific interactive narrative, or even a specific engine; Instead, we aim at finding general metrics that should be useful for the vast majority of systems.

This system-independence restricts the range of analysis that may be carried out from computer traces, such as how many times this specific game element has been triggered, which character is most interacted with, etc. This kind of analysis is highly valuable and should be conducted in parallel with others. The system-independent approach brings another type of information that concerns the field as a whole: information across systems, that help to characterize the interactive narrative experience. In the long run, this may support a benchmarking approach that is currently lacking in the field.

Before getting straight to the heart of the matter, we would like to precise that the interactive metrics we propose is valuable but limited in scope. It only covers some features of the experience and may not apply to all systems. In particular, it requires access to a fully working system, preferentially accessible online, so that data can be collected in a central server. Our research lab hosts such an online interactive narrative [23].

In the rest of the paper, we are going to formalize a small number of indicators that are aggregated data representing certain features of the interactive narrative experience. Then, their implementation, as well as an attempt to validate these indicators on a sample of users will be described, followed by a conclusion.

2 Preliminary Indicators, Based on System-Independent Player Data

2.1 Raw Data Selection

Game analytics usually consist of two phases: The selection of raw data and the construction of meaningful indicators from these data [17]. Given our constraint of engine-independence, data that are external to the selection/calculus of actions by the engine will be considered. For that purpose, we will provide a loose formalization of the process of interactive narrative.

We will consider that an interactive narrative session contains a succession of narrative actions that include system narrative actions and player narrative actions; each of these actions is selected among a list of narrative actions called choices, including system choices and player choices. This statement calls for several comments:

- Narrative actions correspond to a certain level of granularity, above the physical actions such as moving in space. This distinction between narrative actions and physical actions is observed in many interactive narrative systems.
- This framework is discrete, in the sense that we do not consider the case in which the player chooses a continuous value in a given range. Furthermore, the framework is best suited to interactive narrative in which choices are explicitly given to the player, by a menu system for example [5], [9], [11], [19]. Nevertheless, the free input systems where the input is internally converted into one action among a list of possible actions also fit into the framework. It is for example the case of Façade, in which the player's input is converted into one of around 40 possible discourse acts [10].

- Despite the sequential nature of the above statement, it does not preclude parallelism and overlapping between actions. Also, there is no strict alternation between system and player actions.

To a session *s* is associated:

– a tuple with all performed narrative actions ai

– a tuple with all sets of choices C_i *(*$a_i \in C_i$*)*

with *i* being an index over the successive narrative actions, $i \in T(s)$. *T(s)* can be understood as the turns of session s, keeping in mind that there is no turn-by-turn requirement. A subset of these indices, called *P(s)*, contains all indices corresponding to the player's narrative actions.

In addition to these data, information related to the timing of these actions is also recorded. It consists of an additional tuple with a pair of dates: the starting date and ending date of the action *(start(ai),end(ai))*.

To summarize, raw data for a session consists of played actions – a_i – , choices – $\underline{C_i}$ –, turns – *T(s)* –, player turns – *P(s)* – and timing data – *(start(*a_i*),end(*a_i*))* –. As stated in introduction, these data are general and do not consider the meaning of each action.

2.2 Aggregated Indicators

The first type of indicators simply concerns the total length of a session, which is a useful hint about the size of an interactive narrative. Two kinds of lengths can be considered: the discrete length (in terms of number of actions) and the continuous length (in terms of time), which we denote respectively length (LEN) and duration (DUR). They are calculated as follows, in which card means the cardinality of a set (number of elements):

$$LEN(s) = card(T(s)) \qquad\qquad DUR(s) = end(a_{LEN(s)}) - start(a_0)$$

Another feature that can be easily measured is the number of different actions displayed. We therefore introduce the *diversity* that can be assessed in two ways: either from the point of view of a unique session, the intra-diversity, or from the point of view of a set of sessions, the global diversity.

Let us define S as a set of individual sessions. The intra-diversity is calculated as follows:

$$IDIV(s) = card(\bigcup_{i \in T(s)} a_i)$$

$$IDIV(S) = \frac{\sum_{s \in S} IDIV(s)}{card(S)}$$

The first line in the above formula consists in constructing a set of all actions that have been played in a session (symbol ∪ above being the union symbol) and counting the number of elements in this set. In other words, it involves counting the number of unique displayed actions in a session. The second line is the mean among a set of sessions.

The global diversity is calculated as follows:

$$GDIV(S) = card(\bigcup_{s \in S}(\bigcup_{i \in T(s)} a_i))$$

In this case, the counting of unique displayed actions is performed throughout all sessions. If the interactive narrative system is generative at the level of individual actions, that is if actions are not author-written but generated by combining small authored piece, the global diversity may be very high, typically several hundreds. Nevertheless, global diversity is not a measure of generativity, because an intensive authoring work may also lead to a high diversity. What is interesting to measure is the ratio between the global diversity and the intra-diversity. It represents the ability of the system to generate new content at each session and is denoted the renewal rate:

$$RENEW(S) = 1 - \frac{IDIV(S)}{GDIV(S)}$$

For example, a renewal rate of .6 would indicate that after one session, 60% of possible actions are still to be discovered by replaying. A renewal rate of zero corresponds to the case in which *IDIV=GDIV*, that when exactly the same set of actions are displayed in each session.

Artificial Intelligence-based researches in Interactive Narrative have often been explicitly driven by the goal of increasing player's freedom, in comparison with hypertexts and other branching-based approaches [8], [10], [14], [20]. This freedom, as one aspect of agency, has been advocated as a key promise of interactive narrative, in terms of enjoyment. Furthermore, in terms of motivation in a learning context, such freedom is also sought to increase player's motivation [8]. This freedom can be measured through range of choices and frequency of choices [7]. In order to effectively measure this range that is the number of choices offered to the player, we introduce an indicator called the *choice range*. A large choice range is often desired in interactive narrative research – otherwise usual branching techniques would be sufficient –, but this criteria must not be taken in isolation: a large choice range does not make sense if most choices lead to narrative incoherence. Furthermore, choice range is not the goal per se of interactive narrative research; rather, it is agency that is pursued [12], [25]. Nevertheless, choice range is a well-defined and measurable criterion that participates to more complex features such as motivation, control or agency. From the raw data exposed in the previous section, choice range (CR) is calculated as follows:

$$CR(s) = \frac{\sum_{i \in P(s)} card(C_i)}{card(P(s))}$$

Choice range is simply the mean of the number of choices ("*card(C_i)*"), over turns in which the user made a choice ("*P(s)*"). It may be defined at the level of a session, but also at the level of a set of sessions. This enables the researchers to evaluate an interactive narrative, providing that the set is representative of a certain population. This may also be used to compare different populations playing the same interactive narrative. The range of choices of a set of sessions is defined as follows:

$$CR(S) = \frac{\sum_{s \in S} CR(s)}{card(S)}$$

The next indicator also concerns the player's degree of freedom, but along the temporal dimension. It consists of the frequency of choices [7]. It can be determined either in discrete time, in terms of the proportion of actions that are played by the player or in real time, based on the elapsed time between two player actions. These indicators are calculated as follows:

$$DCFREQ(s) = \frac{card\big(P(s)\big)}{card\big(T(s)\big)} \qquad RTCFREQ(s) = \frac{card(P(s))}{DUR(s)}$$

The discrete choice frequency (DCFREQ) varies between 0 (no player action) and 1 (no system action). None of these extreme values make sense, but a discrete frequency of 0.5 means that there are as many player actions as system actions, while a value of 0.1 would indicate that for one action chosen by the player, nine system actions are displayed, which corresponds to a far more passive situation. The real time choice frequency (RTCFREQ) is measured in hertz (provided that dates are expressed in seconds), a value of 1 meaning one user action per second on average, which would constitute a very strong involvement of the player, comparable to an action game. Lower values are expected, and therefore, we will use centi-hertz (cHz) below.

Beyond the number of choices, what is also relevant is the content of the offered choices. More precisely, it would be relevant to distinguish an interactive narrative with the same choices repeated over all successive turns, from an interactive narrative when new choices appear while others disappear. We therefore propose an indicator called *variability* that measures, for each player turn, the difference between the choices offered since the beginning of the session and at the moment. For that purpose, we introduce $P_i(s)$, the set of indices of the player turns from the beginning of the session to the turn i, with $i \in P(s)$. The variability is calculated as follows:

$$CVAR(s,i) = 1 - \frac{card\big(\big(\bigcup_{j \in P_{i-1}(s)} C_j\big) \cap C_i\big)}{card(C_i)}$$

This formula simply calculates the intersection () symbol) between the current set of choices ("C_i") and all the choices proposed before. If the choices are entirely new, the intersection is empty, and $CVAR(s,i)=0$. The formula is extended to a whole session s and to a set of sessions S as follows:

$$CVAR(s) = \frac{\sum_{i \in P(s)} CVAR(s,i)}{card(P(s))} \qquad CVAR(S) = \frac{\sum_{s \in S} CVAR(s)}{card(S)}$$

We have defined a certain number of indicators based on raw generic data that can be extracted from an interactive narrative session. These indicators approximate intuitive properties that are usually sought by interactive narrative research such as generativity, replayability, degree of freedom, and variability. In the rest of the paper, these indicators will be implemented and tested.

3 Indicators' Evaluation

3.1 Goal

The goal of this experimental study is twofold. First, we want to check the practical usefulness of the indicators, in particular whether they differentiate effectively between interactive narratives. Second, we want to explore whether they relate to the subjective interactive narrative experience of users measured through a questionnaire. This questionnaire consists of a number of *scales* that are extracted from the measurement toolkit for interactive storytelling developed within the IRIS European project [6], [24]. The selected scales are, with the number of questions per scale in parentheses: Curiosity (3), Enjoyment (2), User satisfaction (2), Flow (5), Emotional state (positive (3) and negative (3)), Believability (3), Suspense (4), Role adoption (3), Aesthetic (3) and Usability (3). The English version of this questionnaire has been tested and validated on various games and interactive narratives. We translated it into French and used it in our research.

3.2 Method

The experiment consists of creating two interactive narratives, IN0 and IN1 that correspond to two variants of the same story-world. Both interactive narratives are based on the interactive drama *Nothing For Dinner*, a playable 3D interactive drama [5], [23]. *Nothing For Dinner* is generative in the sense that its individual actions and the order of these actions are not prewritten but generated from a calculation. It is build with *IDtension*, an Artificial Intelligence-based narrative engine [20]. The difference between the two variants lies in one parameter termed the *complexity*. This complexity defines the optimal number of narrative threads (or *nuclei*) that should be active in parallel. IN0 sets this number at two, which means whenever possible, two nuclei are active – this corresponds to the default value of the work. IN1 raises this value to ten, which is more than the total number of nuclei in this story-world. As a result, it is expected that IN1 will overload the user.

The narrative engine was modified to export raw data into a database. A separate Java program enriches the database with indicators. Subjects play with either one of the variants and fill the questionnaire composed of the IRIS scales (see above) and a few open questions. Given the modifications made on the second variant, significant differences are expected between the two groups, in terms of range of choice, and variability, while the global diversity and the duration should remain statistically identical. We expect a slight difference for the frequency of choice: because there is more to read in the second version, the user may take more time to play. Furthermore, the experiment will explore possible correlations between the questionnaire's scales and indicators. Note that this paper will not report the differences for the scales between the two groups, but focuses on the indicators.

40 subjects, mostly students, were recruited, 21 assigned to condition IN0 and 19 assigned to condition IN1. Experiments were conducted in groups of up to 8 subjects. After a short introduction to the experiment and the signing of consent forms, subjects played online with *Nothing For Dinner* until the end of the story or a maximum of 30 minutes. Then subjects were requested to fill-in a computer-based questionnaire that included the scales-related questions and open questions. They finally received 10 Swiss francs for their participation. A whole session lasts about 50 minutes.

3.3 Results and Interpretation of Indicator Measurements

During the experiment, some technical problems were encountered, which required us to restart the interactive narrative. Therefore, for some subjects, more than one single session was recorded. In these cases, only sessions with a whole duration above a given threshold (8000 sec.) were selected. In case two sessions remained for the same subject, the first one was selected. As a result, two subjects were discarded from group IN0, leading to 19 subjects for each condition. Also, this led us to not compare the indicators related to durations, because the stop criterion was not reliable.

The comparison of indicators between the two versions is displayed in Table 1. It includes four session indicators, for which an independent samples *t*-test was performed to calculate the statistical significance of the difference and two system indicators not aggregated from session indicators.

Table 1. Comparison of indicators result on the two conditions (N=19 for each condition). For the four session indicators, mean and standard deviation (in parenthesis) are displayed.

	Intra-diversity	Choice range	Choice frequency (cHz)	Choice variability	Global diversity	Renewal
IN0	65.05 (20.7)	**23.3** (5.37)	2.54 (.92)	**.208** (.083)	382	0.829
IN1	65.16 (16.6)	**34.9** (7.94)	3.00 (1.08)	**.139** (.051)	337	0.801
Sig.	>0.05	<.001	>0.05	.004	-	-

In both versions, users could act and see 65 different actions per session, on average. There is a slight difference between the two versions regarding the global diversity, which gives a renewal of 0.829 or 0.801. Therefore, in both versions, after one session, more than 80% of the content (in terms of distinct actions) has not yet been experienced. Choice range (CR) is 23 for IN0 and 35 for IN1, the difference being significant. This means that for the normal version, the user had 23 choices on average, while in the second version, 35 choices. Real-time choice frequency (RTCFREQ) varies from 2.54 cHz to 3 cHz in the two versions, the difference is not significant. This corresponds to a slow pace game, with 36 seconds between two player actions on average for both groups. Regarding choice variability (CVAR), IN0 has a variability of 0.21, while IN1 has a variability of 0.14, which means that on average, 21% of choices were new for IN0, while only 14% for IN1. The difference is statistically significant.

According to our hypothesis, choice range and choice variability changed between the two conditions. Because IN1 did not restrict the number of nuclei, more choices were offered on average for the user. However, because in both versions the narrative material is identical (it is the same story-world), the additional choices in IN1 were generated from the same data, hence a lower variability for IN1. Therefore, the greater choice range is compensated by a lesser choice variability. In other words, IN1 has more choices, but they are more repetitive. It is difficult to interpret the difference in Global diversity (GDIV) because the story-world is the same in the two conditions. With more sessions, the global diversity should increase and reach a plateau. Its value, in the order of the hundreds, illustrates the generative nature of the narrative engine, because these actions were produced with 144 elementary narrative elements. Finally, contrary to our hypothesis, we have not observed any significant difference in the choice frequency (RTCFREQ).

3.4 Results and Interpretation of Questionnaires-Indicators Correlations

In the correlation analysis, the two groups were merged into one single group of 38 subjects, in order to maximize the chances to obtain some correlations. Results of questionnaires comprised 13 scores regarding the IRIS scales – aggregated themselves from questionnaires answers – and 4 scores for indicators (intra-diversity, choice range, choice frequency, choice variability). All the data were normalized between 0 and 1. Intra-diversity and choice range were divided by the maximal value of the data in our sample.

Spearman correlations were performed between all these variables, and in particular between the scales and the indicators, which generates 48 correlation coefficients. Three correlations were found significant (significance lower than 0.05):

- The choice variability (CVAR) is correlated with the Flow scale, with a coefficient of .322 (sig. = .049). The flow scale's internal consistency is acceptable (Cronbach's alpha of 0.654, N=38).
- The choice variability (CVAR) is correlated with the Positive emotional state scale, with a coefficient of .442 (sig. = .005). This scale measures the positive affective response of the user via questions concerning specific emotions such as “Now, after the experience, I feel enthusiastic” [6]. The internal consistency is good (Cronbach's alpha of 0.802, N=38).
- The choice range (CR) is negatively correlated with the Positive emotional state scale, with a coefficient of -0.489 (sig. = .002).

Therefore, the fact that choices offered to the user change during the narrative experience is positively correlated with both the flow and positive emotional state, while the choice range is negatively correlated with positive emotional state. Note that Flow is also correlated with the Positive emotional state (.535). Note also that the scores for Flow and Positive emotional state are, for the first group 2.81 and 3.23 respectively, in a 5-point Likert scale (1–5), which correspond to rather high values (see [15] for comparisons with other systems tested with the same scales).

It is not straightforward to interpret the results. Regarding the flow, one interpretation could be that repetitive choices disrupt immersion and engagement. This is in line with some answers we had to the open question “What did you not like in the game?”: “The repetitive aspect of some situations”; “the repetition of some possible actions”. Another interpretation is that low variability, with many choices persisting between turns, gives the impression that user's actions are useless, because it does not change (enough) the choice list. This is in line with the fact that, at the level of individual questions, the highest correlation was between the variability and the answer to the question (before translation): “I had a good idea while I was performing about how well I was doing” (Spearman Correlation of 0.498).

Regarding the emotional state, it seems that choice variability, in a narrative context, tends to provide stronger positive emotions to the user. The positive effect of variability on user's emotion is interesting because, as game designer, one may consider that repetition is not bad, as it is frequent in many games. This result suggests that in a narrative context, users do not expect to have recurring choice but rather new choices appearing during the experience.

The negative correlation between choice range and the positive emotional state is puzzling, because it would tend to favor less choice range. Two remarks regarding the context of these measures are worth mentioning. First, we are dealing with a rather high choice range (35 for IN1), meaning that there may be a kind of saturation. The impact of choice range is certainly not linear. Second, our user interface for choosing an action consists of a flat list of actions for each Non Player Character. When too many choices are provided, a dozen of actions may be displayed to the user, creating a negative experience. This illustrates the facts that our first results must be interpreted in the context of the specific interactive narrative system used for the experiment.

Finally, correlations that were not observed may be interesting. For example, choice range, which seems a strong indicator in interactive narrative, did not correlate with the agency scale. This can be explained by the fact that in the proposed protocol, the difference in choice range is created by superimposing more content at the same time. To obtain an effect on agency, the additional choices should have been added around the same situation, while guaranteeing the impact of these new choices on the story.

4 Conclusion

We have proposed a series of indicators based on logs to assess various properties of an interactive narrative. The indicators have been created in a system-independent manner: they could be used for a large variety of systems. These indicators include length and duration, diversity, renewal, choice range, choice frequency, and choice variability. The indicators have been implemented on a fully implemented interactive drama, *Nothing For Dinner*, and have been tested with 38 subjects (plus 2 subjects later discarded) dispatched in two groups that corresponded to two variants of the interactive drama. Choice range and choice variability were different in the two variants, showing the interest of theses indicators to qualify an interactive narrative. Correlations were found between indicators and validated scales for assessing the interactive narrative experience. For example, choice variability was positively correlated with Flow (feeling immersed and engaged) and Positive emotional state (feeling excited). Therefore, choice variability appeared to be a relevant indicator of some experiential qualities of interactive narratives, which was not anticipated. This result illustrates that objective indicators constitute a promising complementary tool to evaluate interactive narrative works and systems.

To extend the approach, two research directions are envisioned. First, correlations between objective and subjective metrics need to be sought out beyond the scope of a unique system. The correlations that we found need to be reproduced for other interactive narratives, possibly with a different narrative engine and a different visualization engine. Second, more indicators could be built and tested. For example, we have not explored indicators that may use the *distance* between two sessions. This distance could either be calculated specifically by the interactive narrative engine or be calculated on the base of the presence/absence of actions and their ordering so as to stay in a system-independent approach (e.g. by adopting the *diff* function in Unix, or the Levenshtein distance as in [13]). With a distance metrics, dispersion values can be estimated, which provides a better measurement of story diversity. Clustering could also provide interesting insights on the variability of the gameplay. Indeed, indicators can be particularly useful for system designers and authors to obtain automatic feedback from users as initiated in [2].

References

1. Aylett, R., Louchart, S.: Being There: Participants and Spectators in Interactive Narrative. In: Cavazza, M., Donikian, S. (eds.) ICVS-VirtStory 2007. LNCS, vol. 4871, pp. 117–128. Springer, Heidelberg (2007)
2. Ben, S., McCoy, J., Treanor, M., Reed, A., Wardrip-Fruin, N., Mateas, M.: Story Sampling: A New Approach to Evaluating and Authoring Interactive Narrative. In: Foundations of Digital Games 2014, FDG 2014 (2014)
3. Cavazza, M., Charles, F., Mead, S.J.: Characters in Search of an author: AI-based Virtual Storytelling. In: Balet, O., Subsol, G., Torguet, P. (eds.) ICVS 2001. LNCS, vol. 2197, pp. 145–154. Springer, Heidelberg (2001)
4. Cheong, Y.-G., Young, R.M.: Narrative Generation for Suspense: Modeling and Evaluation. In: Spierling, U., Szilas, N. (eds.) ICIDS 2008. LNCS, vol. 5334, pp. 144–155. Springer, Heidelberg (2008)
5. Habonneau, N., Richle, U., Szilas, N., Dumas, J.E.: 3D Simulated Interactive Drama for Teenagers Coping with a Traumatic Brain Injury in a Parent. In: Oyarzun, D., Peinado, F., Young, R.M., Elizalde, A., Méndez, G. (eds.) ICIDS 2012. LNCS, vol. 7648, pp. 174–182. Springer, Heidelberg (2012)
6. Klimmt, C., Roth, C., Vermeulen, I., Vorderer, P.: The Empirical Assessment of The User Experience In Interactive Storytelling: Construct Validation of Candidate Evaluation Measures - IRIS FP7 D7.2, http://ec.europa.eu/information_society/apps/projects/logos/4/231824/080/deliverables/001_IRISNoEWP7DeliverableD72.pdf
7. Laurel, B.: Computers as Theatre. Addison-Wesley Professional, New York (1993)
8. Lester, J.C., Rowe, J.P., Mott, B.W.: Narrative-Centered Learning Environments: A Story-Centric Approach to Educational Games. In: Emerging Technologies for the Classroom, pp. 223–237. Springer, Heidelberg (2013)

9. Marsella, S.C., Johnson, W.L., LaBore, C.: Interactive pedagogical drama. In: Proceedings of the fourth International Conference on Autonomous agents - AGENTS 2000, pp. 301–308. ACM Press, New-York (2000)
10. Mateas, M., Stern, A.: Integrating Plot, Character and Natural Language Processing in the Interactive Drama Façade. In: Göbel, S., et al. (eds.) Proceedings of the Technologies for Interactive Digital Storytelling and Entertainment (TIDSE) Conference, pp. 139–151. Fraunhofer IRB, Darmstadt (2003)
11. Miller, L.C., Marsella, S., Dey, T., Appleby, P.R., Christensen, J.L., Klatt, J., Read, S.J.: Socially Optimized Learning in Virtual Environments (SOLVE). In: André, E. (ed.) ICIDS 2011. LNCS, vol. 7069, pp. 182–192. Springer, Heidelberg (2011)
12. Murray, J.H.: Hamlet on the Holodeck: The Future of Narrative in Cyberspace. Free Press, New York (1997)
13. Porteous, J., Charles, F., Cavazza, M.: NetworkING: Using Character Relationships for Interactive Narrative Generation. In: Gini, et al. (eds.) Proceedings of the 2013 International Conference on Autonomous Agents and Multi-agent Systems, pp. 595–602. IFAAMAS, Richland (2013)
14. Riedl, M.O., Young, R.M.: From Linear Story Generation to Branching Story Graphs. IEEE Comput. Graph. Appl. 26(3), 23–31 (2006)
15. Roth, C., Klimmt, C., Vermeulen, I.E., Vorderer, P.: The experience of interactive storytelling: Comparing "Fahrenheit" with "Façade". In: Anacleto, J.C., Fels, S., Graham, N., Kapralos, B., Saif El-Nasr, M., Stanley, K. (eds.) ICEC 2011. LNCS, vol. 6972, pp. 13–21. Springer, Heidelberg (2011)
16. Schoenau-Fog, H.: Hooked! – Evaluating Engagement as Continuation Desire in Interactive Narratives. In: André, E., et al. (eds.) ICIDS 2011. LNCS, vol. 7069, pp. 219–230. Springer, Heidelberg (2011)
17. Seif El-Nasr, M., Drachen, A., Canossa, A.: Game Analytics - Maximizing the Value of Player Data. Springer, Heidelberg (2013)
18. Seif El-Nasr, M., Milam, D., Maygoli, T.: Experiencing interactive narrative: A qualitative analysis of Façade. Entertain. Comput. 4(1), 39–52 (2013)
19. Short, E.: Versu, http://emshort.wordpress.com/2013/02/14/introducing-versu/
20. Szilas, N.: A Computational Model of an Intelligent Narrator for Interactive Narratives. Appl. Artif. Intell. 21(8), 753–801 (2007)
21. Szilas, N.: IDtension: a narrative engine for Interactive Drama. In: Göbel, S., et al. (eds.) Proceedings of the Technologies for Interactive Digital Storytelling and Entertainment (TIDSE) Conference, pp. 187–203. Fraunhofer IRB, Darmstadt (2003)
22. Szilas, N.: Requirements for Computational Models of Interactive Narrative. In: Finlayson, M. (ed.) Computational Models of Narrative, papers from the 2010 AAAI Fall Symposium, pp. 62–68. AAAI Press, Menlo Park (2010)
23. Szilas, N., Dumas, J., Richle, U., Habonneau, N., Boggini, T.: Nothing For Dinner, http://tecfalabs.unige.ch/tbisim/portal/
24. Vermeulen, I., Roth, C., Vorderer, P., Klimmt, C.: Measuring user responses to interactive stories: Towards a standardized assessment tool. In: Aylett, R., Lim, M.Y., Louchart, S., Petta, P., Riedl, M. (eds.) ICIDS 2010. LNCS, vol. 6432, pp. 38–43. Springer, Heidelberg (2010)
25. Wardrip-Fruin, N., Mateas, M., Dow, S., Sali, S.: Agency reconsidered. In: Proceedings of DiGRA 2009 (2009)
26. Zhu, J.: Designing an Interdisciplinary User Evaluation for the Riu Computational Narrative System. In: Oyarzun, D., Peinado, F., Young, R.M., Elizalde, A., Méndez, G. (eds.) ICIDS 2012. LNCS, vol. 7648, pp. 126–131. Springer, Heidelberg (2012)