Four Quantitative Metrics Describing Narrative Conflict

Stephen G. Ware, R. Michael Young, Brent Harrison, and David L. Roberts

Digital Games Research Center,
North Carolina State University, Raleigh, NC, USA
sgware@ncsu.edu,young@csc.ncsu.edu
beharri5@ncsu.edu,robertsd@csc.ncsu.edu

Abstract. Conflict is an essential element of interesting stories. In previous work, we proposed a formal model of narrative conflict along with 4 quantitative dimensions which can be used to distinguish one conflict from another based on context: *balance*, *directness*, *intensity*, and *resolution*. This paper presents the results of an experiment designed to measure how well these metrics predict the responses of human readers when asked to measure these same values in a set of four stories. We conclude that our metrics are able to rank stories similarly to human readers for each of these four dimensions.

Keywords: conflict, narrative, metrics, planning

1 Introduction

Narratologists, screen writers, game designers, and other researchers in computer narrative agree that conflict is an essential element of interesting stories [15, 5, 11, 2, 1]. Conflict provides an impetus for the plot to begin [6], and it keeps the audience engaged as the story unfolds, even if they already know the ending [9]. Conflict also structures the discourse of a story into meaningful units that together make up a coherent whole [6, 1].

Our previous work [17, 18] defined a formal computational model of narrative conflict that was inspired by research in narratology, based on AI planning, and designed for story generation. In short, conflict occurs when a goal seeking agent's plan is thwarted by another agent, the environment, or its own plans to achieve other goals. This definition is intentionally broad to cover the entire spectrum of conflict.

In order to provide greater control over story content, we identified seven dimensions from various narratological sources that can be used to distinguish one conflict from another. The first three—*participants, subject,* and *duration*—have discrete values which can be directly observed in the structure of the planning model and have already been experimentally validated [19]. The other four—*balance*, *directness*, *intensity*, and *resolution*—are quantitative, continuous values which require more contextual information. No consensus exists on how to measure these dimensions.

We provide four simple formulas intended to measure each of these dimensions and describe an experiment to test whether the observations of human readers correspond to the values predicted by our formulas. This paper presents the findings of that experiment along with an analysis and discussion of the results. We conclude that our formulas for *balance*, *directness*, and *resolution* rank stories in the same order as human readers, and that our formula for *intensity*, while less accurate, still ranks stories similarly to human readers.

This work is an attempt to operationalize a few of the tacit story metrics used by human readers into formulas which can be used by machines to evaluate the content of stories. Even if the formulas do not operate like a human mind, they can enable more human-like story analysis by modeling specific features of narrative structure. Many narrative-oriented virtual environments like role playing games, training simulations, and intelligent tutoring systems need to adapt their content in response to user actions. By capturing a model of how humans evaluate stories, we can guide story generation systems to produce content that is better suited to meet the expectations of the audience by leveraging the benefits provided by well-structured conflicts.

2 Related Work

Much previous work exists on modeling human perception with quantitative metrics. Yannakakis [20] provides a survey of research that measures concepts like *fun* and *flow* in the context of video games. Less work has been done specifically in narrative. Peinado and Gervs [13] collected four metrics from human readers evaluating the quality of stories produced by their ProtoPropp system: *linguistic quality*, *coherence*, *interest*, and *originality.*

Our approach differs from these because we measure properties of stories apart from their effects on the reader. The dimensions of conflict answer *who? what? when?* and *how?*; they are designed so that readers can agree on their values even when they disagree on how fun or interesting a given conflict is.

At least three story generation systems have attempted to reason about conflict quantitatively. IDtension [16] assigns a "conflict value" to each action in a story for the degree to which a character is forced to act against its moral principles. MEXICA [14] estimates the tension a reader perceives in the story at each world state and crafts a pattern of rising and falling action. The AI Director of the zombie survival game series *Left 4 Dead* [4] moderates the intensity of its conflicts by controlling the number and frequency of enemies, distribution of power-ups, and geography of levels. It monitors metrics such as the player's health and accuracy to measure stress, and uses this information to create a series of peaks and valleys in the story's intensity.

Because conflict is such a diverse phenomenon, we have chosen to measure many individual dimensions rather than attempt to quantify conflict as a single value. This higher level of detail will allow story generating systems to produce content with more specific constraints. We also hope to provide a model which can generalize to many domains.

3 Dimensions of Conflict

Complete formal descriptions for each dimension are given by Ware and Young [18]. Some essential notation is reproduced here.

We assume that a conflict exists between character c_1, who intends to carry out a sequence of actions f_1, and character c_2, who intends to carry out a sequence of actions f_2. Some action in f_1 conflicts with an action in f_2—that is, some action in f_1 prevents c_2 from executing one of the actions in f_2. Let E be the set of actions which actually occur in the story. E may contain some actions from both f_1 and f_2, but cannot contain all the actions from both (because the two character plans are incompatible).

Dimensions are measured from some character's point of view. In general, a dimension is expressed as $name(c)$ where *name* is the name of the dimension and c is the character from whose point of view the dimension is being measured. We also employ two additional functions in the range $[0, 1]$:

- $\pi(f)$ measures how likely some sequence of actions f is to succeed.
- $\text{utility}(c, f)$ measures how satisfied character c is with the state of the world after the sequence of actions f occurs. $\text{utility}(c, \emptyset)$ is the character's utility before the conflict begins.

Examples from the *Star Wars* films are provided to illustrate each dimension.

3.1 Balance

Balance measures the relative likelihood of each side in the conflict to succeed, regardless of the actual outcome:

$$\text{balance}(c_1) = \frac{\pi(f_1)}{\pi(f_1) + \pi(f_2)}$$

The range of *balance* is $[0, 1]$. If c_1 is likely to prevail—that is, $\pi(f_1)$ is close to 1, then balance is high for c_1. If the opposing participant, c_2, is is more likely to prevail, then balance is low for c_1.

When Obi Wan Kenobi challenges Darth Vader to a duel in *Star Wars: A New Hope*, he knows that he cannot win. Vader's skill is at its peak while Kenobi's skill is waning with age. In this conflict, the balance for Kenobi is low while the balance for Vader is high.

3.2 Directness

Directness measures how close the participants are to one another at the moment of the conflict:

$$\text{directness}(c_1) = \frac{\sum_{i=1}^{n} \text{closeness}_i(c_1, c_2)}{n}$$

We chose to measure 3 types of *closeness* in this experiment: familial, emotional, and interpersonal. The range of *directness* and each *closeness* is $[0, 1]$.

During the climax of *Star Wars: Return of the Jedi*, Luke Skywalker and Darth Vader are face to face and emotionally close because of their family ties. Interpersonal closeness is non-zero when one agent participates in the conflict via other agents. There is interpersonal distance between the Emperor and Luke because the Emperor participants in the conflict via his subordinate, Vader.

3.3 Intensity

Intensity is the difference between how high a participant's utility will be if she prevails and how low it will be if she fails (which can be estimated as how bad things will be if her opponent succeeds):

$$\text{intensity}(c_1) = |\text{utility}(c_1, f_1) - \text{utility}(c_1, f_2)|$$

The range of *intensity* is $[0, 1]$. Two factors influence this formula: how much can be gained and how much can be lost. Situations which are high risk or high reward have medium intensity, while situations which are both high risk and high reward have high intensity. Like balance, intensity is measured regardless of the actual outcome of the story.

The Rebel Alliance's plan to destroy the Death Star in *A New Hope* is very intense. If they succeed they will cripple the Empire, but if they fail their rebellion will be crushed. This is a high risk, high reward conflict.

3.4 Resolution

Resolution measures the change in utility a participant experiences after a conflict ends. Recall that E is the events from f_1 and f_2 that actually occur:

$$\text{resolution}(c_1) = \text{utility}(c_1, E) - \text{utility}(c_1, \emptyset)$$

The range of resolution is $[-1, 1]$. Luke and the Rebel Alliance overcome the Empire at the end of *Return of the Jedi*. Their resolution is high, while the resolutions for Darth Vader and the Emperor are low.

4 Experiment Design

The task of predicting the exact value a reader will report for some dimension is difficult considering how sensitive these concepts are to subtleties of interpretation. Simply predicting high or low is easier, but would provide less support for the strength of our model. As a middle ground, we tested whether our formulas could rank four stories in the same order as human readers. If readers agree on an ordering, and if that ordering agrees with our predictions, we assume that our formulas can approximate these dimensions of conflict.

Each participant was shown the four stories given in Figure 1 (initially in a random order) and asked to sort them from lowest to highest for each dimension.

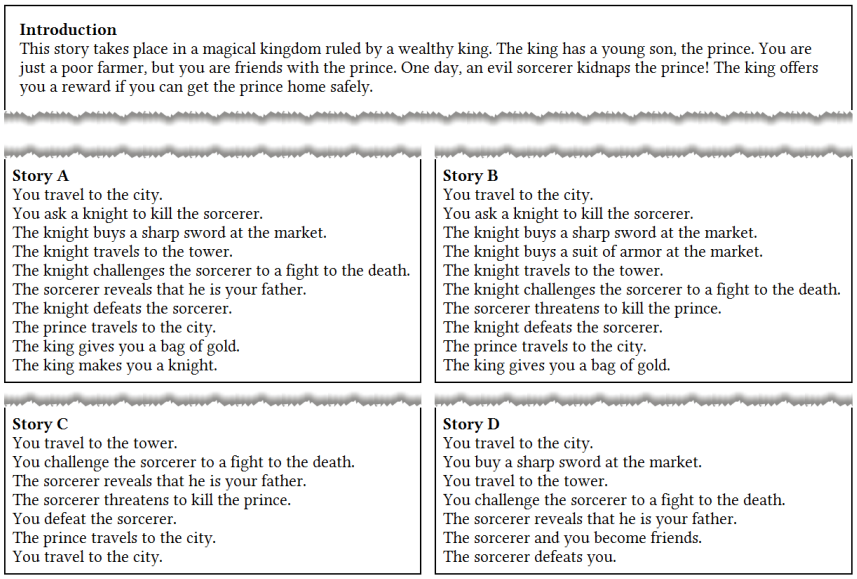

Fig. 1. The four stories used in the experiment. Each story has the same beginning, but a different middle and end. These stories can be generated by a narrative planner such as CPOCL [17] and translated into natural language using simple templates.

Liklihood You Will Win the Conflict
Rate the stories based on how likely you and your allies are to win out over the sorcerer. If you expect your team to win, rate the story high. If you expect your team to lose, rate it low. Do not consider whether or not you actually win. Only rate the stories based on what you expected to happen before someone gets defeated.

Fig. 2. Example dimension description given to participants for *balance*.

Dimensions were presented in a random order. All four stories had the same beginning, but different middles and ends. All stories were written in the second person such that the reader was the protagonist in conflict with an evil sorcerer. The text of the stories was composed of simple actions which can be formally expressed as STRIPS-style planning operators [7]. In other words, the stories were such that they could be produced by an automated planning system like the CPOCL algorithm [17].

The content of the stories was structured so that, given our orderings for each dimension, no two stories would appear at the same index for the same dimension (i.e. the story with highest intensity was not highest for any other dimension). Readers were not told of this constraint. To avoid confusion from vocabulary, the dimensions were not given names in the study. Participants were simply given a description of the concept and asked to sort the stories. An example description for balance is given in Figure 2.

4.1 Hypotheses

In this paper, we explore two hypotheses:

1. For each dimension, participants will rank stories similarly to one another.
2. For each dimension, participants will rank stories similarly to our metrics.

Our formulas predicted the following orderings:

Balance: $\{C\ D\ A\ B\}$ The protagonist (or the knight fighting for the protagonist) is more likely to succeed when wielding a sword, and even more so when wearing armor. The knight is more likely to win a fight than the protagonist (a poor farmer).

Directness: $\{B\ A\ C\ D\}$ Familial distance is low when the sorcerer is the protagonist's father, high otherwise. Emotional distance is low when the sorcerer and protagonist are friends, high otherwise. Interpersonal distance is low when the protagonist fights, high if he gets the knight to fight for him.

Intensity: $\{A\ B\ D\ C\}$ The protagonist's life is at stake when he fights the sorcerer himself. The prince's life is at stake when the sorcerer threatens to kill the prince. When neither life is at stake, intensity is low; when both are at stake, intensity is high. Participants were told to value their own lives higher than those of others, so D is more intense than B.

Resolution: $\{D\ C\ B\ A\}$ When the protagonist dies, resolution is lowest. Participants were asked to value riches over poverty, so some reward is better than nothing and 2 rewards is best of all.

This experiment does not require a commitment to specific formulas for $\pi(f)$ and utility(c, f) as long as those formulas produce the predicted orderings given above. For example, we assume that the knight is more likely to succeed when he has a sword and armor than when he has just a sword and no armor. It is not necessary to measure the exact difference in π between the two stories.

4.2 Notes on Analysis

The data collected from each participant was an ordering of four stories for each dimension. The task of choosing an ordering is similar to classification, but it is important to note that two orderings can still be substantially similar even if they are not exactly identical. Capturing this degree of similarity is important, which precludes certain standard statistical tests.

For example, Cohen's or Fleiss's κ coefficient is often used to measure inter-rater reliability, but κ assumes that the raters are choosing one of several discrete categories. The orderings {A B C D} and {A B D C} would be considered two different categories even though 5 of the 6 pairwise orderings are the same in both; in other words, A comes before B in both; A comes before C in both; etc. The various edit distance metrics, such as Hamming distance [10], suffer from similar problems. The Hamming distance between {A B C D} and {D A B C} is 4, the maximum possible.

Kendall's Tau Distance To account for similarity between responses, we used Kendall's τ distance [12] to compare orderings. τ counts the number of pairwise differences between two lists. Formally, let $\text{index}(x, S) = 1$ just when x is the first element in ordered set S, $\text{index}(x, S) = 2$ just when x is the second element in ordered set S, etc. Given two ordered sets M and N, an *inversion* is an ordered pair of elements (x, y) such that $\text{index}(x, M) < \text{index}(y, M)$ and $\text{index}(x, N) > \text{index}(y, N)$. This means that x is ordered before y in M, but x is ordered after y in N. The τ distance between two ordered sets can be expressed as $\tau(M, N)$ and is equal to the number of inversions that exist between M and N. Kendall's τ distance is symmetric, meaning $\tau(M, N) = \tau(N, M)$.

When comparing two orderings of length 4, the minimum τ distance is 0, when both orderings are the same. The maximum τ distance is 6, when one oridering is the reverse of the other. The τ distance between {A, B, C, D} and {D, C, B, A} is 6 because the pairs {A, B}, {A, C}, {A, D}, {B, C}, {B, D}, and {C, D} are inverted. If we fix M and choose N at random, assuming that all 24 permutations of the 4 stories are equally likely, then on average there will be a τ distance of 3 between M and N.

5 Results

30 people participated in the study—19 males and 11 females with 26 to 35 being the most common age group. Participants were recruited via e-mail and social networking websites. No compensation was offered.

5.1 Most Popular Orderings

In order to evaluate our formulas, we need to determine the most popular ordering for each dimension based on the data submitted by human readers. To account for similarity between answers, we chose the ordering with the lowest average τ distance from each participant's ordering.

For a given dimension of conflict, let $\{p_1, p_2, ..., p_n\}$ be the orderings chosen by the n participants for that dimension (here, $n = 30$). Let M be all 24 possible orderings of the 4 stories. For each of the 24 possible orderings, m, we calculated its average τ distance as:

$$\forall m \in M : \quad \tau_{\text{avg}}(m) = \frac{\sum_{i=1}^{n} \tau(m, p_i)}{n}$$

As an example, consider m ={A B C D}, the first of the 24 permutations in M. To calculate τ_{avg} for m for the dimension of *balance*, we calculate $\tau(\{\text{A B C D}\}, p_i)$ for all 30 orderings p_i that were reported by the participants for *balance*; then we average those 30 values. An ordering's τ_{avg} can be thought of as its average distance from each person's answer.

When an ordering's τ_{avg} is low, that ordering is more popular—it agrees more with the orderings reported by participants. If all 30 participants had reported the same ordering, that ordering's τ_{avg} would be 0 and the reverse ordering

would have the max τ_{avg} of 6. The most popular orderings for each dimension are given in the first row of Table 2, in Section 5.3, where we discuss how our formulas agree with readers.

5.2 Participant Agreement

Before demonstrating to what extent our formulas agree with human readers, we must first demonstrate that readers agree amongst themselves. In other words, we wish to know how strongly the participants agree that the most popular ordering is correct.

As discussed above, there is no clear way to calculate Fleiss's κ coefficient to measure inter-rater agreement for this data. However, it is possible to express agreement by comparing our data, shown in Figure 4, to distributions representing agreement and disagreement, shown in Figure 3:

- **Perfect Agreement:** If users agreed completely with one another, they would all report the exact same ordering for a dimension.
- **Relative Agreement:** Given the subjective nature of how people perceive stories, it may be impossible to achieve perfect agreement. It is more realistic to compare against a distribution which indicates high (but not perfect) agreement. One such distribution is given in Figure 3. This distribution assumes that $\frac{2}{3}$ of the participants will choose the most popular ordering, and then the function will decay exponentially by 3 from there.
- **Disagreement:** If there is complete disagreement, we would expect answers to appear as if they were given at random. This would result in a uniform distribution across the 24 possible permutations for the 4 stories. That uniform distribution, when plotted as τ distance from the most popular ordering, is a roughly normal distribution (as seen in Figure 3).

As a null hypothesis, we assume our observed distributions for each dimension will fit the *disagreement* distribution. To evaluate this, we used Fisher's

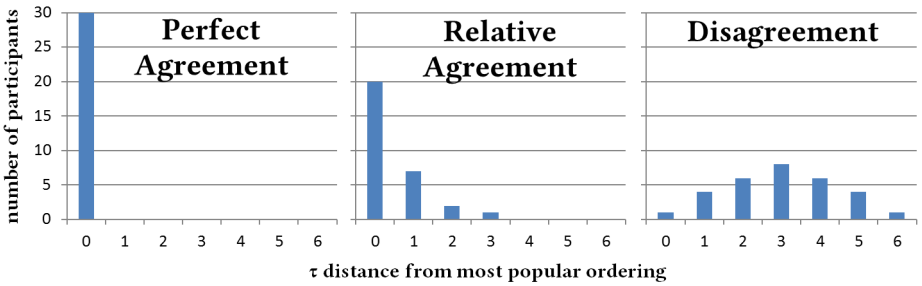

Fig. 3. The three distributions against which we compared our data. These histograms show how many participants (y axis) chose an ordering that was some τ distance (x axis) away from the most popular ordering for each dimension.

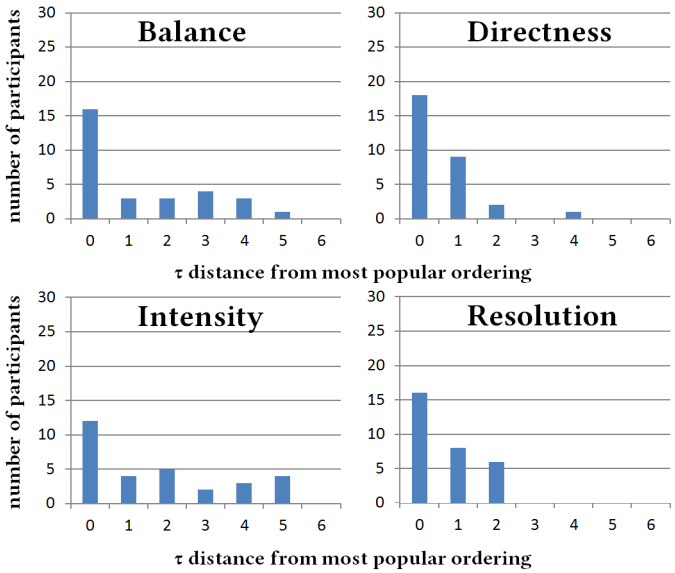

Fig. 4. The observed distributions for each dimension. These histograms show how many participants (y axis) chose an ordering that was some τ distance (x axis) away from the most popular ordering for each dimension.

exact test, which is similar to the χ^2 test but performs better for distributions with small expected values [8]. For all four dimensions, there was a statistically significant difference between our data and the *disagreement* distribution (for *balance* $p = 0.003$, for *directness* $p = 0.000$, for *intensity* $p = 0.028$, and for *resolution* $p = 0.000$). The null hypothesis is rejected—that is, participants do not disagree.

Now we can evaluate the alternative hypothesis—that users agree on the most popular ordering. For this, we employ a metric for measuring the similarity of two distributions called Bhattacharyya distance [3]. Bhattacharyya distance is 0 when two distributions are the same, and approaches 1 as the distributions become less similar. For each dimension, we want to know if the distribution defined by readers is most similar to the *agreement*, *relative agreement*, or *disagreement* distribution. Table 1 demonstrates that the dimensions of *directness* and *resolution* are more similar to the *perfect agreement* distribution than they are to the *disagreement* distribution; however, the dimensions of *balance* and *intensity* are more similar to *disagreement* than to *perfect agreement*. However, all four dimensions are most similar to the *relative agreement* distribution. These results support our hypothesis that users agree amongst themselves on a correct ordering for the four dimensions, especially for *directness* and *resolution*.

Table 1. The formula for Bhattacharyya distance, D_B, and the Bhattacharyya distances between the observed distributions for each dimension and the *Perfect Agreement* (Perfect), *Relative Agreement* (Agree), and *Disagreement* (Disagree) distributions. The lowest distance is highlighted in gray for each dimension.

Given two discrete probability distributions p and q over domain X,

$$\mathrm{D}_B = -\ln\left(\sum_{x\in X}\sqrt{p(x)q(x)}\right)$$

Dimension	Perfect	Agree	Disagree
Balance	0.314	0.108	0.240
Directness	0.255	0.037	0.619
Intensity	0.465	0.168	0.175
Resolution	0.314	0.040	0.650

Table 2. The top 6 orderings and the bottom ordering for each dimension based the on average τ distance. The orderings predicted by our formulas are in gray.

Balance		Directness		Intensity		Resolution	
Order	τ_{avg}	**Order**	τ_{avg}	**Order**	τ_{avg}	**Order**	τ_{avg}
CDAB	1.26667	BACD	0.56667	BACD	1.73333	DCBA	0.66667
CDBA	1.66667	BADC	0.96667	BADC	1.93333	DCAB	1.20000
DCAB	1.73333	ABCD	1.36667	ABCD	2.13333	CDBA	1.40000
CADB	2.00000	BCAD	1.36667	BCAD	2.26667	DBCA	1.40000
DCBA	2.13333	ABDC	1.76667	ABDC	2.33333	CDAB	1.93333
CBDA	2.26667	BDAC	1.90000	BDAC	2.33333	DACB	1.93333
...17...	...17...	...17...	...17...	...17...	...17...	...17...	...17...
BADC	4.73333	DCAB	5.43333	DCAB	4.26667	ABCD	5.33333

5.3 Accuracy of Our Formulas

For each dimension of conflict, Table 2 presents the 6 orderings with the lowest τ_{avg} (the top 6 best orderings for that dimension according to the participants). The orderings predicted by our formulas are highlighted in gray. For the dimensions of *balance*, *directness*, and *resolution*, the ordering predicted by our formula has the lowest τ_{avg}. For the dimension of *intensity*, the ordering predicated by our formula has the 5^{th} lowest τ_{avg}. These results support our hypothesis that participants will rank stories in the same order as our metrics. Our formula for *intensity* may need to be improved based on these results to better agree with human perceptions.

6 Discussion

These initial results are promising, especially for *balance*, *directness*, and *resolution*. Several factors may have contributed to the disagreement we observed.

Clarity of Descriptions Participants may have misunderstood the descriptions of one or more dimensions, which were intentionally brief and targeted at a high

school reading level. We attempted to address this by running a small pilot study before the experiment, which provided valuable feedback on how to clarify the definitions. *Intensity* was the most widely misunderstood dimension during the pilot. It is also possible that participants misunderstood the events of the story. At least one participant indicated a misunderstanding of the outcome of story D. To make the stories more G-rated, we used the text "X defeats Y," which does not make it explicit that Y is killed. Our predicted ordering for intensity is based on which characters' lives are at stake, so this may have caused confusion.

Dimension Synergy We assumed that each dimension could be measured independently of the others, but it is possible that participants perceived synergies between them. For example, if much was at stake (high *intensity*) but there was little chance that the sorcerer would prevail (low *balance*), participants might have given the story a low ranking for *intensity*. This may explain why story C is ordered before story D in the most popular ordering for intensity. We hope to investigate how dimensions influence one another in future work.

Knowledge of the Ending The two dimensions that showed the least participant agreement—*balance* and *intensity*—require the reader to measure them independently of the actual outcome of the story. If the protagonist appears likely to prevail, balance should be high regardless of whether or not he or she actually prevails. At least two participants reported difficulty ignoring their knowledge of the outcome. In future versions of this study, rather than ask participants to ignore the ending, we intend to leave the ending out. This may help to avoid the bias introduced by foreknowledge.

7 Conclusions and Future Work

Previous work focused on developing a formal model of conflict that encompasses the entire phenomenon [17]. This experiment was designed to validate four metrics for measuring specific dimensions of conflict which can be used to evaluate the content of individual stories. Based on our results, we draw three conclusions:

- The dimensions of *balance*, *directness*, *intensity*, and *resolution* are recognizable qualities of conflict.
- Readers demonstrate agreement on how to rank stories based on *balance*, *directness*, *intensity*, and *resolution*. We suspect that improvements to this experiment will yield higher agreement for *balance* and *intensity*.
- The orderings predicted by our formulas for *balance*, *directness*, *resolution*, and to a lesser extent *intensity*, corresponded with those chosen by readers.

The higher goal of this research is to identify what measurable qualities of stories readers perceive and how they evaluate different stories based on those criteria. We believe that this research represents progress toward that goal because it identifies quantitative metrics for evaluating conflict.

In the future, we hope to improve our formulas based on this data and guide the CPOCL algorithm's production of stories with constraints on the values of these dimensions. Constraints on each dimension will be based on observed patterns in various genres. For example, in most computer role playing games, the protagonist's conflicts with the antagonist become increasingly balanced and direct. Combined with the three discrete structural dimensions of conflict—*participants*, *subject*, and *duration*—we hope to gain considerable control over the content and quality of the stories we produce.

References

1. H. P. Abbott. *The Cambridge Introduction to Narrative.* Cambridge U. Pr., 2008.
2. H. Barber and D. Kudenko. Dynamic Generation of Dilemma-Based Interactive Narratives. In *AIIDE*, 2007.
3. A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, 35(99-109):4, 1943.
4. M. Booth. The ai systems of left 4 dead. In *Keynote, AIIDE*, 2009.
5. C. Crawford. *Chris Crawford on Game Design.* New Riders, 2003.
6. L. Egri. *The Art of Dramatic Writing.* Wildside, 1988.
7. R. Fikes and N. J. Nilsson. STRIPS: A New Approach to the Application of Theorem Proving to Problem Solving. *AI*, 2(3/4):189–208, 1971.
8. J. L. Fleiss, B. Levin, and M. C. Paik. *Statistical Methods for Rates and Proportions.* John Wiley & Sons, 3 edition, 2003.
9. R. J. Gerrig. *Experiencing Narrative Worlds: On the Psychological Activities of Reading.* Yale U. Pr., 1993.
10. R. W. Hamming. Error detecting and error correcting codes. *Bell System Technical Journal*, 29(2):147–160, 1950.
11. D. Herman, M. Jahn, and M. L. Ryan. *Routledge Encyclopedia of Narrative Theory.* Routledge, 2005.
12. M. Kendall. *Rank Correlation Methods.* Griffin, 1948.
13. F. Peinado and P. Gervás. Evaluation of automatic generation of basic stories. *New Generation Computing*, 24(3):289–302, 2006.
14. R. Pérez y Pérez and M. Sharples. MEXICA: A Computer Model of a Cognitive Account of Creative Writing. *Journal of Experimental & Theoretical AI*, 13(2):119–139, 2001.
15. M. L. Ryan. *Possible Worlds, Artificial Intelligence, and Narrative Theory.* Indiana U. Pr., 1991.
16. N. Szilas. IDtension: A Narrative Engine for Interactive Drama. In *TIDSE*, 2003.
17. S. G. Ware and R. M. Young. CPOCL: A Narrative Planner Supporting Conflict. In *AIIDE*, 2011.
18. S. G. Ware and R. M. Young. Toward a Computational Model of Narrative Conflict. Technical Report DGRC-2011-01, DGRC, North Carolina State University, Raleigh, NC, USA, 2011. http://dgrc.ncsu.edu/pubs/dgrc-2011-01.pdf.
19. S. G. Ware and R. M. Young. Validating a Plan-Based Model of Narrative Conflict. In *FDG*, 2012.
20. G. N. Yannakakis. How to model and augment player satisfaction: A review. In *1st Workshop on Child, Computer, and Interaction*, 2008.