BRANCHES: A Dataset and Analysis of Narrative Graphs in Visual Novels

Giovanni Carbinatti[1], **Wandemberg Santana Pharaoh Gibaut**[1]

[1]Instituto de Pesquisas Eldorado
Av. Alan Turing, 275 – 13083-898 – Campinas – SP – Brasil

{giovanni.carbinatti, wandemberg.gibaut}@eldorado.org.br

Abstract. Introduction: *Interactive narratives in visual novels present complex branching structures that remain difficult to analyze systematically at scale. Despite the widespread use of the Ren'Py engine, there is a lack of standardized datasets enabling computational analysis of such structures.* ***Objective:*** *This work introduces BRANCHES, a dataset of narrative graphs extracted from visual novels, aiming to support large-scale analysis of interactive storytelling.* ***Methodology:*** *We collected 592 publicly available visual novel projects and implemented a static analysis pipeline to parse script files, extracting narrative nodes, transitions, and player choices. Each project is represented as a directed graph, and structural metrics such as branching factor and choice density are computed.* ***Results:*** *The dataset reveals a wide spectrum of narrative structures, with most projects exhibiting near-linear progression (median branching factor of 1.31) and a small subset showing highly complex branching (up to 20.32). The number of choices ranges from 0 to 3303, indicating substantial variability in interactivity. These findings show that visual novels span from simple linear narratives to highly interactive structures, demonstrating the potential of the dataset for large-scale analysis.*

Keywords *Interactive Narrative, Visual Novels, Narrative Graphs, Static Analysis, Dataset*

1. Introduction

Interactive narrative has become an increasingly important domain within digital games research [Riedl and Young 2010, Kybartas and Bidarra 2016], enabling players to influence story progression through meaningful choices and branching narrative structures. Among the various forms of interactive storytelling, visual novels (VNs) have emerged as one of the most prominent and widely adopted formats [Camingue et al. 2021]. Visual novels combine textual storytelling, character dialogue, player choices, and audiovisual elements to produce highly narrative-driven experiences.

From a computational perspective, VNs offer a particularly interesting form of structured narrative. Unlike traditional linear media such as novels or films, VN narratives are inherently branching systems composed of dialogue sequences, decision points, and multiple story routes [Camingue et al. 2021]. This structure makes them a valuable resource for research in areas such as interactive storytelling, narrative modeling, dialogue systems, and computational narratology.

Despite this potential, the availability of publicly accessible datasets for studying interactive narrative structures remains limited. Existing narrative datasets are typically

derived from books, movie scripts, or conversational corpora, which generally assume linear story progression. These resources do not capture the branching and decision-driven properties that characterize interactive narratives. As a result, researchers interested in modeling or analyzing branching storytelling systems lack standardized data resources for experimentation and evaluation.

Visual novels are a particularly suitable domain for constructing such datasets. Many VNs are developed using scripting-based engines such as Ren'Py [Ren'Py Team 2004], where narrative content is written in structured script files that define dialogue, player choices, and narrative flow. These scripts implicitly encode the narrative graph of the game, making them amenable to automatic parsing and structural analysis. An example of this representation is shown in Figure 1.

In this work, we introduce BRANCHES (Branching Narrative Corpus for Human-Interactive Entertainment Systems), an open dataset of visual novel scripts designed to support research in computational analysis of interactive narratives. The dataset is constructed from publicly available visual novel projects and includes structured representations of dialogue, player choices, and narrative graphs extracted from script files, facilitating tasks such as dialogue analysis, narrative graph modeling, and structural evaluation of branching stories.

Our contributions are threefold:

1. A curated dataset of narrative graphs and dialogue structures extracted from publicly available visual novel projects.
2. A parsing pipeline capable of extracting dialogue, choice structures, and narrative graphs from visual novel scripts.
3. A structured dataset representation that enables research tasks involving interactive narrative modeling and analysis.

By making this dataset publicly available at `https://huggingface.co/datasets/TensorGio/BRANCHES`, we aim to support future research on interactive storytelling, narrative structure analysis, and computational approaches to branching narratives in digital games.

2. Related Work

2.1. Interactive Narrative and Branching Storytelling

Interactive narrative has long been studied as a domain in which computational systems support storytelling experiences in which users influence the progression of the narrative. Early research explored formal models of branching narratives and player agency, emphasizing the tension between authorial control and player choice [Mateas and Stern 2003, Murray 2017]. Subsequent work expanded this perspective through the development of systems and frameworks for interactive storytelling, procedural narrative generation, and narrative planning [Riedl and Young 2010].

A central aspect of interactive narrative systems is the representation of the narrative structure. Branching story models are often represented as graphs where nodes correspond to narrative states and edges correspond to player choices or story transitions. Several works have investigated computational approaches for modeling and analyzing

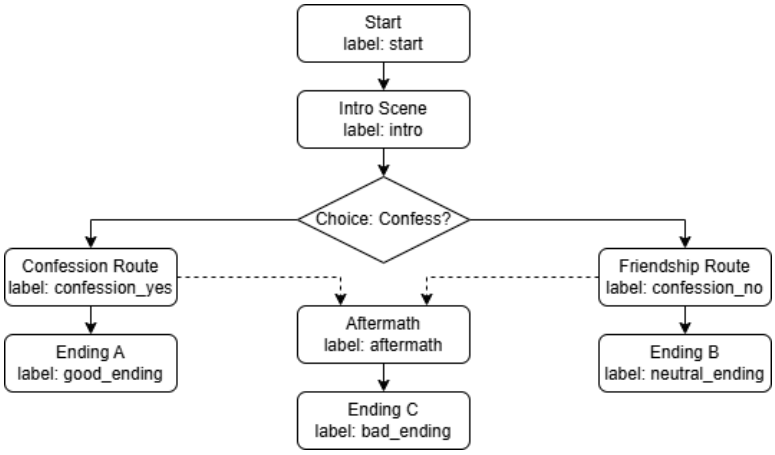

Figure 1. Example of a visual novel narrative represented as a directed graph. Nodes correspond to narrative labels, while edges represent transitions. Branching occurs at player choice points, leading to different routes.

narrative graphs [Young et al. 2013], including studies on narrative planning, drama management, and story generation systems.

Visual novels represent a particularly prominent form of branching interactive narrative in digital games. Their structure typically consists of dialogue sequences punctuated by decision points that lead to different story routes. Because VNs rely heavily on script-based authoring, they offer a natural source of structured narrative data suitable for computational analysis.

2.2. Narrative and Dialogue Datasets

Datasets play a central role in enabling research in natural language processing (NLP) and narrative modeling. A number of narrative datasets have been proposed in the literature, including corpora derived from movie scripts, books, and conversational transcripts. For example, the Cornell Movie-Dialogs Corpus provides a large collection of conversational exchanges extracted from movie scripts [Danescu-Niculescu-Mizil and Lee 2011], while the BookCorpus dataset contains narrative text from thousands of novels [Zhu et al. 2015].

More recent datasets have focused on structured story understanding and narrative reasoning tasks. The ROCStories dataset, for instance, provides short commonsense stories designed to support narrative completion tasks [Mostafazadeh et al. 2016].

Despite the availability of these resources, most existing narrative datasets assume linear story progression. As a result, they do not capture the branching structures and player-driven decision points characteristic of interactive narratives. This limitation restricts their applicability for research on interactive storytelling systems and computational models of player choice.

2.3. Static Analysis and Structural Extraction

The extraction of structured representations from source artifacts is a well-established problem in software engineering, particularly in the context of static analysis. Techniques for parsing code and constructing control flow graphs (CFGs) [Allen 1970] have been widely used to analyze program behavior and structure.

In the context of narrative systems, similar ideas can be applied to script-based formats, where control flow constructs define possible progression paths. However, relatively little work has explored the application of static analysis techniques to interactive narrative scripts, especially at scale.

Recent efforts in game data mining have focused on extracting assets and metadata from game repositories [Drachen et al. 2013], but the underlying narrative structure of script-based games has received comparatively little attention. Our work addresses this gap by applying static analysis to visual novel scripts developed with the Ren'Py engine, transforming them into typed directed graphs that capture control flow, player choices, and conditional branches, enabling systematic large-scale structural analysis of interactive narratives across hundreds of projects simultaneously.

3. Dataset Construction

3.1. Game Collection

To investigate structural properties of branching narratives in VNs, we constructed a dataset of games developed with Ren'Py, using two sources: GitHub and itch.io [GitHub, Inc. , itch.io].

3.1.1. GitHub

Repositories were automatically collected from GitHub using the platform's search API with four complementary queries targeting Ren'Py-specific patterns: `extension:rpy renpy`, `filename:script.rpy`, `"label start" extension:rpy`, and `"menu:" extension:rpy`. These queries were designed to maximize recall of narrative Ren'Py projects by targeting both file extension and structural script patterns characteristic of visual novel authoring. The results were restricted to public repositories, yielding 196 successfully cloned repositories.

For each cloned repository, the project directory was scanned recursively for `.rpy` script files. All 196 repositories contained at least one script file, consistent with the search filters applied during collection.

After parsing with the pipeline described in Section 3.2, 183 repositories produced valid narrative graphs and were included in the final dataset, representing a parse success rate of 93.4%. The remaining 13 repositories were excluded as they produced no extractable narrative structure, likely due to non-narrative Ren'Py usage such as engine configuration files, template projects, or repositories containing only compiled `.rpyc` files without readable source scripts.

3.1.2. Itch.io

In addition to GitHub repositories, we collected visual novel projects from itch.io, an open indie game distribution platform with a large catalog of Ren'Py-based titles. Unlike GitHub, which primarily hosts developer-shared and often incomplete projects, itch.io contains published games with associated player ratings, review counts, and curated metadata, providing a complementary source of finished narrative works.

Games were discovered through itch.io's browse pages using a set of ten tags selected to maximize recall of Ren'Py-based VNs. Two tags directly identify the target engine and format (`visual-novel`, `renpy`); five target the dominant genres of the Ren'Py ecosystem on itch.io (`otome`, `romance`, `yuri`, `yaoi`, `dating-sim`); two identify narrative-driven games regardless of genre (`story-rich`, `narrative`); and one captures the visual aesthetic strongly associated with VNs on the platform (`anime`). For each tag, we crawled ten pages of results, collecting game URLs and removing duplicates across tags. This ten pages limit was set to balance corpus size against collection cost, as itch.io's tag-based ranking returns the most relevant results in early pages. This process yielded 2,078 unique game pages.

For each discovered game, we retrieved the page HTML and extracted structured metadata, including title, author, description, player rating, review count, price, platform availability, and genre tags. Rating and review data were extracted from JSON-LD structured data embedded in each page, while platform information was obtained via the itch.io API.

Of the 2,078 discovered games, 813 were freely available for download, as determined by the presence of accessible download links on the game page. To further filter for Ren'Py-based titles, we computed a heuristic relevance score $s \in [0, 1]$ based on the presence of engine-specific terms in the page text, defined as:

$$s = \min\big(1,\, 0.6 \cdot \delta(\texttt{ren'py}) + 0.3 \cdot \delta(\texttt{renpy}) + 0.4 \cdot \delta(\texttt{.rpy}) + 0.2 \cdot \delta(\texttt{visual novel engine})\big) \tag{1}$$

where $\delta(t) = 1$ if term t is present in the page text and 0 otherwise. The weights reflect the specificity of each term: `ren'py`, written with its characteristic apostrophe, is the most unambiguous identifier of the engine, while `.rpy` and `visual novel engine` may appear in broader contexts. Games with $s \geq 0.4$ were selected as download candidates, meaning that at minimum the term `renpy` must appear in the page text alongside at least one corroborating signal. This process yielded 424 titles.

Game archives were retrieved via the itch.io API using authenticated requests, targeting Windows builds where available. Downloaded archives were extracted and scanned for Ren'Py-specific file types (`.rpy`, `.rpyc`, `.rpa`). Archives containing .rpa asset packages were processed with the unrpa tool prior to script extraction. Of the 424 downloaded games, 409 were successfully extracted and confirmed to contain Ren'Py script files. The remaining 15 failed during archive extraction due to unsupported formats or corrupted downloads. After parsing with the pipeline described in Section 3.2, all 409 itch.io games produced valid narrative graphs and were included in the final dataset, yielding a parse success rate of 100% in the extracted corpus.

We further analyze metadata tags associated with the collected projects, as shown in Figure 2, highlighting common thematic categories in the dataset.

3.2. Script Parsing

Ren'Py narratives are organized as sequences of labeled blocks, referred to as labels, which function as nodes in the narrative flow. Each label may contain dialogue lines

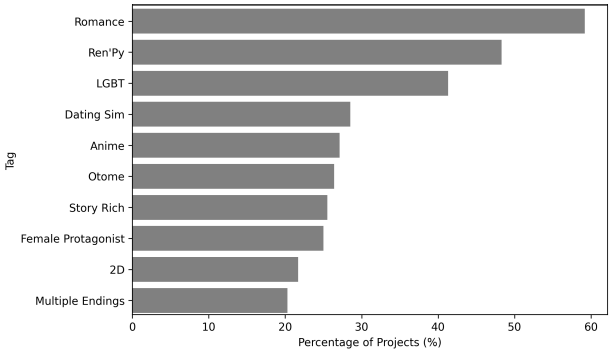

Figure 2. Most frequent tags associated with visual novel projects in the BRANCHES dataset. Romance and LGBT themes are particularly prevalent, reflecting common narrative focuses in the domain.

attributed to named characters, narration without an explicit speaker, player choice menus, and control flow transitions to other labels via `jump` or `call` statements.

To extract these structures, we implemented an indentation-aware static parser that processes all `.rpy` files in each project. Before the main parsing pass, a pre-pass collects character alias definitions of the form `define x = Character("Name")`, enabling dialogue lines to be attributed to full character names rather than script-level identifiers. Multi-line string literals are joined into single lines prior to parsing to ensure complete dialogue extraction. The parser identifies five primary elements:

- Dialogue lines, consisting of a resolved character name and text.
- Narration, represented as quoted text without an explicit speaker.
- Menu choices, which represent player decisions and induce branching edges in the narrative graph.
- Unconditional transitions, defined through `jump` and `call` statements linking narrative labels.
- Conditional transitions, defined through `if`, `elif`, and `else` blocks containing control flow statements. Since the outcome of conditional branches cannot be determined through static analysis alone, all possible targets are included as edges in the graph, tagged as conditional to distinguish them from guaranteed transitions.

Labels containing only a `return` statement are marked as explicit terminal nodes, representing narrative endings or subroutine exits. From these elements, a directed graph $G = (V, E)$ is constructed for each game, where nodes correspond to narrative labels and edges are typed as `jump`, `call`, `choice`, or `conditional`. This representation captures the possible progression paths of the narrative independently of runtime execution, while the edge type annotation preserves the structural distinction between player-driven branching and authorial control flow.

3.3. Dataset Representation

For each parsed game, the extracted data is stored in five structured files:

- `dialogue.csv` contains all dialogue and narration lines, with fields for line index, narrative label, resolved character name, and text content.
- `choices.csv` contains player choice options, with fields for the source label, choice text, and target label.

- `nodes.csv` lists all narrative labels with a boolean field indicating whether each node is a terminal node, i.e., a narrative ending or subroutine exit.
- `edges.csv` stores directed transitions between labels, with fields for source node, target node, edge type (`jump`, `call`, `choice`, or `conditional`), and the condition expression for conditional edges.
- `metadata.json` provides per-game summary statistics including number of labels, dialogue lines, choices, nodes, edges, conditional edges, terminal nodes and terminal node ratio.

In addition to per-game files, two dataset-level files are provided: `all_dialogue.csv` aggregates all dialogue lines across games, tagged with a game identifier, enabling cross-game textual analysis without requiring individual file access. `summary.csv` consolidates the metadata statistics for all games into a single table, facilitating large-scale comparative analysis and dataset-level metric computation.

This representation supports both textual analysis, through the dialogue and choice content, and structural analysis, through the typed directed graph and associated metrics.

3.4. Dataset Statistics

The final dataset comprises 592 visual novel projects (183 collected from GitHub and 409 from itch.io), each represented as a directed narrative graph. Table 1 summarizes key descriptive statistics across the dataset.

Table 1. Descriptive statistics of the BRANCHES dataset.

Metric	Mean	Std	Min	Median	Max
Labels	84.37	135.38	1	41	1230
Dialogue lines	4499.77	9661.60	0	1452.50	79195
Choices	38.61	172.61	0	4	3303
Branching factor	1.49	1.46	0	1.31	20.32
# Terminal Nodes	40.92	71.89	0	24	792
% Terminal Nodes	57.91	24.17	0	64.52	100
Choice density	0.28	0.45	0	0.10	3.80

Content Advisory

A subset of the dataset contains content intended for adult audiences. Among the itch.io subcorpus, 36 games (8.8%) were flagged as age-restricted based on the presence of mature content indicators on their game pages, including explicit content warnings and platform-level age gate markers. These games are retained in the dataset to preserve the structural diversity of the corpus, as adult-oriented VNs frequently exhibit complex branching structures and high choice density. However, researchers using the dataset for NLP tasks involving dialogue content should be aware of this characteristic and apply appropriate filtering using the `age_restricted` field provided. It should be noted that this classification was performed through static page analysis and may underestimate the true proportion of mature content, as some games may not explicitly advertise age restrictions on their page text. The GitHub subcorpus does not include

equivalent metadata, as repository pages do not provide standardized content rating systems; consequently, the presence of mature content in GitHub-sourced projects cannot be systematically determined from the collected data, and researchers should exercise independent judgment when working with the full dataset.

4. Narrative Structure Analysis

4.1. Graph Representation of Narratives

Each visual novel in the dataset is represented as a directed graph $G = (V, E)$, where V is the set of nodes corresponding to narrative labels and E is the set of directed edges representing transitions between labels, as described in Section 3.2. This representation abstracts away engine-specific execution details while preserving the underlying narrative topology, enabling each visual novel to be analyzed as a discrete graph and facilitating the application of graph-based metrics to study branching structure, narrative depth, and interactivity.

4.2. Structural Metrics

To characterize the narrative complexity of the collected works, we compute a set of graph-based metrics for each game.

4.2.1. Branching Factor

The branching factor measures the average number of outgoing edges per node, capturing how frequently the narrative diverges into different paths. It is defined as:

$$B = \frac{|E|}{|V|} \tag{2}$$

where $|V|$ is the number of nodes (narrative labels) and $|E|$ is the number of edges (transitions). Higher values indicate narratives with greater structural branching.

4.2.2. Terminal Nodes

Terminal nodes are defined as nodes with no outgoing edges, corresponding to narrative endings or dead-ends. We compute both their total count and their proportion relative to all nodes, revealing how frequently narratives conclude and how endings are distributed.

4.2.3. Choice Density

Choice density measures how frequently player decisions occur within a narrative, defined as:

$$D = \frac{C}{|V|} \tag{3}$$

where C is the number of choices and $|V|$ is the number of nodes. This metric distinguishes highly interactive narratives from predominantly linear ones.

4.3. Dataset-Level Analysis

We analyze the structural properties of the dataset using the metrics defined in Section 4.2. The results reveal distinct patterns of variability and skewness across visual novel narratives.

The distribution of branching factor indicates that most projects remain close to linear, with a median value of 1.31 and a third quartile of 1.51. This suggests that, although branching is present, many narratives favor limited divergence. However, extreme cases with branching factors above 20 demonstrate the existence of highly complex structures with dense connectivity. These patterns are illustrated in Figure 3, which shows a strong concentration of projects with low branching factors alongside a long-tailed distribution.

Narrative scale also varies substantially. The number of nodes ranges from 1 to 1230, with a median of 41, indicating that most projects are relatively small, while a minority of large-scale works significantly increases the overall average. A similar heavy-tailed pattern is observed for dialogue lines, which reach up to 79,195 lines.

The number of choices further reinforces this variability. While the median number of choices is 4, some projects include thousands of decision points (up to 3303), indicating that interactivity is unevenly distributed across the dataset.

Choice density provides a complementary perspective on interaction patterns. Although the mean value is 0.28, the presence of projects with density above 3.0 suggests that certain narratives concentrate multiple choices within localized regions of the graph, leading to highly interactive segments embedded within otherwise sparse structures.

Terminal nodes exhibit a median of 24 per project and a mean of 40.92, with a maximum of 792. The distribution is strongly right-skewed, indicating that most projects include a relatively small number of endings, while a minority accumulates a large number of terminal states. The ratio of terminal nodes to total nodes provides additional insight into narrative closure: higher ratios suggest fragmented structures with multiple local endings, whereas lower ratios indicate narratives that converge toward a smaller set of definitive conclusions.

4.4. Discussion

The structural analysis reveals that VNs occupy a spectrum between linear storytelling and highly interactive branching narratives. While the *Ren'Py* framework enables complex branching, the majority of projects exhibit relatively low branching factors and sparse choice distributions, suggesting that simpler narrative structures are more common in practice.

At the same time, the presence of extreme cases with very high branching factors and choice densities highlights the expressive potential of the medium. These projects demonstrate that VNs can support highly complex narrative designs, but at a higher authoring cost.

The proposed graph-based representation and associated metrics enable systematic comparison across games and support future research in automated narrative analysis, procedural story generation, and interactive storytelling evaluation.

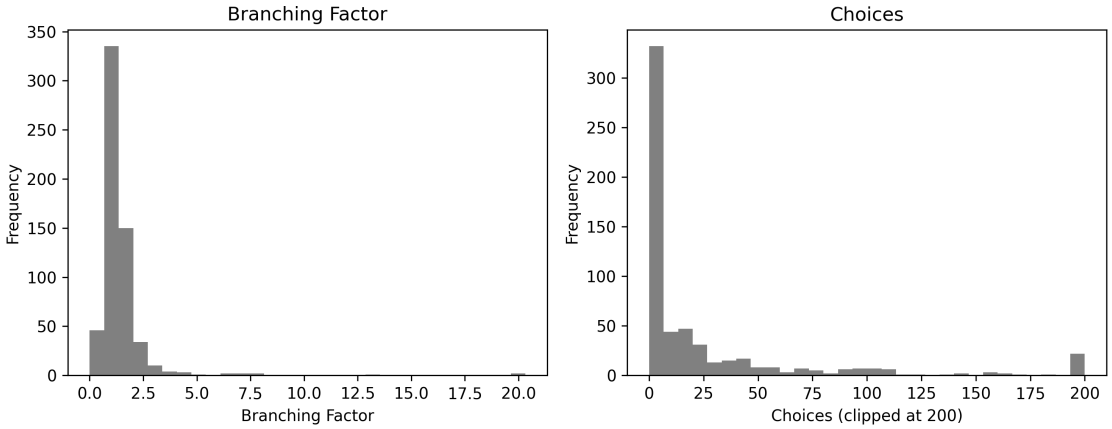

Figure 3. Distributions of branching factor (left) and number of choices (right) in the BRANCHES dataset. Most visual novels exhibit low branching factors and a limited number of choices, while a small subset of projects shows highly complex and interactive structures.

5. Conclusion

In this paper, we introduced BRANCHES, a dataset of narrative structures extracted from publicly accessible VNs developed with the Ren'Py framework. By automatically collecting repositories from GitHub and games from itch.io and parsing their script files, we constructed a structured representation of interactive narratives as directed graphs.

Our approach enables the transformation of visual novel scripts into analyzable artifacts, capturing dialogue, player choices, and control flow transitions. This representation allows both textual and structural analysis, bridging a gap between interactive storytelling and computational methods.

Through structural analysis, we demonstrated that VNs span a broad spectrum of narrative designs. The proposed metrics provide a foundation for future quantitative research on interactive storytelling.

5.1. Future Work

Several directions can extend the impact of this work. First, we plan to enrich the dataset with semantic annotations, including sentiment, emotional tone, and thematic labels. This would enable deeper analysis of how narrative structure interacts with storytelling content.

Second, future work may explore dynamic analysis of VNs, complementing the static parsing approach presented here. Capturing runtime behavior could reveal conditional paths and hidden branches not observable through static analysis alone.

Third, the dataset can support research in procedural narrative generation, particularly for training or evaluating models that generate branching stories.

Finally, we aim to investigate authoring patterns and design practices, identifying recurring structural motifs across games. Such insights could inform tools for assisting writers in creating interactive narratives.

Acknowledgments

The authors are grateful to the Eldorado Research Institute.

References

Allen, F. E. (1970). Control flow analysis. *ACM Sigplan Notices*, 5(7):1–19.

Camingue, J., Carstensdottir, E., e Melcer, E. F. (2021). What is a visual novel? *Proceedings of the ACM on Human-Computer Interaction*, 5(CHI PLAY):1–18.

Danescu-Niculescu-Mizil, C. e Lee, L. (2011). Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd workshop on cognitive modeling and computational linguistics*, pages 76–87.

Drachen, A., Thurau, C., Togelius, J., Yannakakis, G. N., e Bauckhage, C. (2013). Game data mining. In *Game analytics: Maximizing the value of player data*, pages 205–253. Springer.

GitHub, Inc. Github: Where the world builds software. `https://github.com/`. Accessed: 2026-03-21.

itch.io. itch.io: Open marketplace for independent digital creators. `https://itch.io/`. Accessed: 2026-03-21.

Kybartas, B. e Bidarra, R. (2016). A survey on story generation techniques for authoring computational narratives. *IEEE Transactions on Computational Intelligence and AI in Games*, 9(3):239–253.

Mateas, M. e Stern, A. (2003). Façade: An experiment in building a fully-realized interactive drama. In *Game developers conference*, volume 2, pages 4–8.

Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, L., Kohli, P., e Allen, J. (2016). A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849.

Murray, J. H. (2017). *Hamlet on the Holodeck, updated edition: The Future of Narrative in Cyberspace*. MIT press.

Ren'Py Team (2004). Ren'py visual novel engine. `https://www.renpy.org/`. Accessed: 2026-03-21.

Riedl, M. O. e Young, R. M. (2010). Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research*, 39:217–268.

Young, R. M., Ware, S. G., Cassell, B. A., e Robertson, J. (2013). Plans and planning in narrative generation: a review of plan-based approaches to the generation of story, discourse and interactivity in narratives. *Sprache und Datenverarbeitung, Special Issue on Formal and Computational Models of Narrative*, 37(1-2):41–64.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., e Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.