Using Large Language Models for Understanding Narrative Discourse

Andrew Piper
McGill University

Sunyam Bagga
McGill University

Abstract

In this study, we explore the application of large language models (LLMs) to analyze narrative discourse within the framework established by the field of narratology. We develop a set of elementary narrative features derived from prior theoretical work that focus on core dimensions of narrative, including time, setting, and perspective. Through experiments with GPT-4 and fine-tuned open-source models like Llama3, we demonstrate the models' ability to annotate narrative passages with reasonable levels of agreement with human annotators. Leveraging a dataset of human-annotated passages spanning 18 distinct narrative and non-narrative genres, our work provides empirical support for the deictic theory of narrative communication. This theory posits that a fundamental function of storytelling is the focalization of attention on distant human experiences to facilitate social coordination. We conclude with a discussion of the possibilities for LLM-driven narrative discourse understanding.

1 Introduction

For the purposes of narrative understanding, the distinction between "story" (what happened) and "discourse" (how it is told) is fundamental (Bal and Van Boheemen, 2009; Hühn et al., 2009). This bipartite schema was updated by Genette (1980) to include a third dimension, known as the *narrating instance*. For Genette (1980), "narrative discourse" includes the stylistic qualities of how the narrator's voice influences both the story and its structure. In this framework, narrative discourse is not limited to the structural dimensions of storytelling (seen in the bottom right node of Fig. 1). Rather, it encompasses interactions between all three nodes.[1]

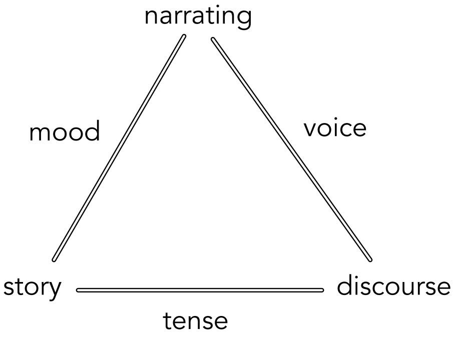

Figure 1: Gérard Genette's classic narrative triangle.

Considerable work in NLP has focused on understanding the two original nodes of Genette's triangle. For the task of *story* understanding (i.e. the lower left node), work has focused on key areas such as the detection of character types (Stammbach et al., 2022; Bamman et al., 2014), event types (Parekh et al., 2023; Chambers and Jurafsky, 2009), and story lines (Caselli et al., 2015)). Similarly, narrative *structure* (i.e. the lower right node), has been amply addressed in concepts such as plot arcs (Reagan et al., 2016; Fudolig et al., 2023), turning points (Ouyang and McKeown, 2015), and non-linearity (Piper and Toubia, 2023).

In this paper we test the affordances of large language models for the analysis of *narrative discourse*, understood here as the three key linking functions between the primary nodes in Genette (1980)'s classic narratological framework (Fig. 1). The value of doing so is to support our broader understanding of the nature and function of storytelling within diverse social and cultural contexts.

As we will see, some of the individual components of narrative discourse have been the subject of NLP research for some time (e.g. dialogue, entity, and tense detection), while some are more novel (e.g. emotionality, conflict, eventfulness, etc.). The principal aim in our work is to bring together these

[1] Confusingly, "discourse" is traditionally used in English to refer to the structural aspects of narrative (lower right node) even though Genette used the term "récit (narrative)" in his original work. A better solution would be to use the term "structure" for the node and "discourse" for the interaction of the nodes.

Proceedings of the The 6th Workshop on Narrative Understanding, pages 37–46
November 15, 2024 ©2024 Association for Computational Linguistics

different strands under a unified theoretical framework to facilitate future benchmarking of language model performance. As Radford and Joseph (2020) have argued through the concept of "theory in, theory out," theory is essential for guiding both model construction and model interpretation.

The use of large language models can potentially address core challenges facing the field of computational narrative understanding. First, they can help narrow the distance between the linguistic features captured by traditional methods in NLP and the theoretical constructs they are meant to capture. The intrinsic language-understanding demonstrated by LLMs can potentially map more directly onto higher-level theoretical constructs.

Second, LLMs can be a powerful way of detecting narrative features at large-scale where we lack abundant training data. As a relatively nascent field with a diverse array of dimensions, we do not yet have a robust infrastructure of already annotated data for a variety of narrative detection tasks.

Third, LLMs can be useful *pragmatically* as a means of bundling diverse computational procedures under a single prompting framework to facilitate greater access and make different approaches more commensurable. Computational narrative understanding is by nature an interdisciplinary undertaking that touches on a range of fields (health, economics, cognitive science, communication, literary studies, sociology and more). Facilitating access can facilitate the wider adoption of common methods for the understanding of narrative communication. That being said, LLMs also introduce their own novel problems of interpretability and generalization and therefore will require extensive testing and validation as is already well underway in numerous areas.

In what follows, we address: 1) prior computational work in narrative understanding as it relates to the two core nodes of Genette's framework; 2) our translation of the concept of "narrative discourse" into a set of natural language prompts; 3) the validation of multiple different models by human annotators; 4) and the insights gained from our models as it relates to understanding the distinctive qualities of narrative discourse. Our aim is to illustrate the ways in which LLMs can contribute to our understanding of narrative communication. We conclude with a discussion of potential limitations and areas for future investigation.

2 Prior Work

A robust literature in NLP addresses two of the key poles of Genette's triangle in Fig. 1 (story and structure). In terms of narrative "structure" (the lower right node, i.e. "how it is told"), a number of pieces have modeled narrative as a structural arc. Schmidt (2015) modeled changes in topic distributions over narrative time in a collection of 80,000 television episodes, while Reagan et al. (2016) and Jockers (2017) have modeled arcs using sentiment detection as a proxy for narrative fortune. This work has been explored in greater depth in Elkins (2022) and newly expanded using ousiometric features such as fear and danger by Fudolig et al. (2023).

Other work has attempted to model narrative structure through the detection of scene changes (Zehe et al., 2021) and narrative "levels," i.e., when stories are imbedded inside of other stories (Reiter et al., 2019). Ouyang and McKeown (2015) have modeled narrative "turning points," based on the theory that narratives are defined by a sense of linear transformation (Bruner, 1991). Piper and Toubia (2023) used word embeddings to model narrative non-linearity through the heuristic of the traveling salesman problem.

On the story side (lower left node), a number of works have modeled different dimensions of story content ("what happened"). Stammbach et al. (2022) have modeled character "roles" (hero, villain, victim) using LLMs, while Rahimtoroghi et al. (2017) and Lukin et al. (2016) have looked at the prediction of character goals in stories built off of prior work encoding semantic relationships in stories (Elson and McKeown, 2010). Goyal et al. (2010) have modeled plot "units," and Jockers and Mimno (2013) have modeled novels as high-level themes using topic modeling. Causality mining has been identified as another core aspect of story understanding by establishing inter-event relationships at the story level (Hu et al., 2017; Meehan et al., 2022; Sun et al., 2024).

In this paper, we seek to integrate the relationships between the three poles of narrative as a set of elementary discursive features. Where prior work has importantly focused on detection tasks related to the individual areas of *story* and *structure*, here we aim to develop a set of features that cover the three core linking functions shown in Figure 1 as described by Genette (1980) and later developed by Herman (2009).

3 Implementation

3.1 Theoretical Framework

In his principal work, *Narrative Discourse*, Genette (1980) introduced three key linking functions between the primary narrative poles, which he named *tense*, *mood*, and *voice*. These functions capture aspects of time and the ordering of events (*tense*); the relationship between events, description, and place (*mood*); and perspectival issues such as point of view, dialogue, interiority, and focalization (*voice*).

Genette's framework has since been updated by Herman (2009) to include three related functions: *sequentiality*, *world building*, and *qualia*, or "what it is like."[2] One can observe how Herman's categories map neatly onto Genette's: *tense-sequentiality, mood-worldbuilding, voice-qualia*.

From this classical tripartite framework, we develop a set of fifteen narrative features, which we then translate into natural language prompts as shown in Table 1. These statements were designed to be elementary in nature with their exact wording refined over multiple rounds of interaction and testing with one of our language models (GPT-4). Some, though not all, of these features have been addressed in prior work (agent detection, dialogue detection, tense, etc). The goal here is to bundle these features within a single theoretical framework and utilize a unified prompting framework for their assessment. Additionally, we introduce new features that have eluded measurement, such as anachrony detection and narrative conflict.

Note that we also translate Genette's somewhat confusingly chosen terms, Tense, Mood, and Voice into the more colloquial terms Time, Setting, and Point-of-View (POV), to facilitate intelligibility.[3] Finally, we also include one non-sensical "honeypot" feature to test whether our models are randomly guessing. The answer to this question should never be positive.

For the first category, "POV (Point of View)," we foreground the experiencing agent as our principal unit. Thus we focus not only on the presence of agents, but also Herman (2009)'s notion of how narrative discourse conveys the "qualia" of experience, i.e. "what it is like." For Herman, narrative discourse aims to illustrate "the pressure of events on a real or imagined consciousness" (14), which nicely captures Genette's idea of "voice." Accordingly, we implement prompts designed to represent the potential foregrounding of sensual and/or emotional experience of characters along with communicative dimensions like dialogue.

For our second category of "Time," we focus on aspects of temporality in narrative, including the use of tense (past/present), anachrony (temporal disorder manifested through flashforwards (prolepsis) or flashbacks (analepsis)), as well as temporal specificity itself, i.e. how explicitly the narrative discourse is located in time. The focus on "event sequences" and "eventfulness" (i.e. how reliant the narrative discourse is on action rather than description, qualia or dialogue) are derived from Herman (2009) and Hühn (2009) respectively and are designed to further capture dimensions of time. The emphasis on conflict in this category stems from narrative theories that foreground the quality of "change" and resolution as essential for narrative communication (Prince, 2012; Bruner, 1991; Herman, 2009; Gottschall, 2012).

For our third category, "Setting," we assess the degree to which narrative discourse situates the reader not only within a definite location ("location"), but also a realized and tangible space ("concreteness"). Symbolism and abstraction capture the inverse, where language removes us from an experiencable location and towards language used to convey disembodied ideas, either abstractly or figuratively.

Note that in every instance we are not attempting to catalogue specific narrative contents, i.e. story-level phenomena. Where story-driven analysis aims to detect plot elements specific to a given story (such as themes, events, locations, or character types), we are interested in the narrative discourse underlying such elements (e.g. the presence of characters, dialogue, qualia, or setting, etc.) In our model we care less about capturing, for example, the specific location or time frame or emotional valence of a story, and instead focus on the extent to which discursive techniques related to temporality, locatability, and perspective are used to convey the events of the story.

3.2 Prompting Framework

We incorporate the sixteen statements listed in Table 1 into the following prompting framework to deliver our questions to the model. We prompt the models to output a three-point ordinal scale based

[2] Herman includes a fourth dimension, *situatedness*, which relates to the social dynamics of narrative and which is beyond the scope of this model.

[3] Genette's terminology faced criticism for its eclectic usage of linguistic terminology so we accordingly adapt it to the general narrative concepts they were aimed to capture.

Category	Feature	Statement
POV	Agents	This passage focuses on the experience of one or more characters.
POV	Emotionality	This passage focuses on the characters' emotions.
POV	Perception	This passage lets you see the world through the eyes and bodies of the characters.
POV	Dialogue	The passage contains dialogue.
TIME	Temporal Specificity	This passage uses specific markers of time.
TIME	Event Sequences	This passage focuses on a series of sequential actions.
TIME	Eventfulness	This passage is very eventful.
TIME	Pastness	This passage is mostly written in the past tense.
TIME	Presentness	This passage is mostly written in the present tense.
TIME	Anachrony	This passage tells of events that occur out of order.
TIME	Conflict	This passage focuses on some kind of conflict or problem.
SETTING	Location	This passage focuses on description of a specific location.
SETTING	Concreteness	This passage focuses on specific concrete details, like objects, places, and surfaces that one can imagine seeing and feeling.
SETTING	Abstraction	This passage focuses on abstract ideas and concepts.
SETTING	Symbolism	This passage uses symbolic or metaphorical language.
HONEYPOT	Emotional Meteorology	This passage focuses on how the emotional states of characters influence the weather.

Table 1: Our features that aim to capture different dimensions of narrative discourse as modeled by Genette (1980) and Herman (2009).

on the degree of presence of a given narrative feature, which we describe below. We use the models listed in Table 3 to compare performance.

Our prompting framework thus consists of the following elements: role prompt, framing question, ordinal scale, narrative feature, and individual passage. Here is an example of our implementation:

> *Today, you are an expert story interpreter. I will give you a passage from a story and ask you a question about it. Here is a passage: [Insert passage.] Can you tell me if the following feature is present? This passage focuses on some kind of conflict or problem. Answer only with a number where 2=strongly present, 1=weakly present, or 0=not present.*

3.3 Data

We use the manually annotated data openly available from Piper and Bagga (2022). In this work, the authors collect 13,543 passages drawn from 18 different genres, roughly split between narrative and non-narrative texts. This data contains passages from contemporary novels, historical novels, short stories, folk tales, and more experimental works of flash fiction. It also includes genres from narrative non-fiction like memoirs, biographies, histories and stories from AskReddit (Ouyang and McKeown, 2015).

These passages have been shown to elicit a high degree of separation when used to train traditional text-based classifiers (F1 = 0.936), even when controlling for different genres in the train and test sets.

Included in this data is a small subset of 394 manually annotated passages for their "narrativity" score. The authors use the construct of "narrativity" to capture the degree to which a given passage engages in the act of narration (Giora and Shen, 1994; Herman, 2009; Pianzola, 2018). We run our experiments on the subset of confirmed narrative passages in the manually annotated data that received a score > 3.0 (on a 5-point Likert scale) and that were initially drawn from the "narrative" genres. This leaves us with 188 sample sentences.

Here we provide examples of low and high rated passages according to their narrativity scores.

High (Avg. Score = 5.0)

Last night I did clinical paperwork and slept while my friends shot whiskey in the living room. Tonight, they're at a party playing beer pong and I'm sipping hot chocolate on the gray couch, the one Simon gave me that's so old the leather has dissolved into wrinkles. Miles the Siamese cat stalks my hair while I read the pharmaceuticals textbook. Tomorrow I imagine more of the same and I'm not sure who, in 10 years, will be sorriest: my impoverished friends, my rich high-living high-blood pressure high-balling self, or the cat, who will be dead. I guess the cat.

Low (Avg. Score = 3.33)

Bored. Displaced. "And what do you think happens to a chigger if nobody ever walks by his weed?" her granny asked, heading for the house with that sidelong uneager unanswered glance, hoping for what? The surprise gift of a smile? Nothing.

3.4 Fine-tuning open-source Models

In addition to GPT-4 (gpt-4-0125-preview), we also experiment with three open-weight LLMs: Llama3 (8B parameters), Mistral (7B parameters), and Mixtral (56B parameters). We fine-tune Llama3 and Mixtral using model distillation from GPT-4 generated annotations.

3.4.1 Training Data

In order to annotate training data for our open-source models, we use GPT-4 (gpt-4-0125-preview) to annotate a dataset of 4,800 passages drawn from the original Piper and Bagga (2022) dataset. Training passages were not drawn from the test dataset. We experiment with modified prompts to optimize training (included with the model documentation).

3.4.2 Implementation Details

All experiments are run on a single A100 40G GPU on Google Colab. We utilize Low Rank Adaptation (LoRA), a parameter-efficient finetuning approach that can significantly reduce GPU memory requirements and the number of trainable parameters (Hu et al., 2021). We use a LoRA rank of 32, LoRA alpha of 16 and a dropout rate of 0.05. Due to memory constraints, we use 8-bit quantization and 4-bit quantization for Llama3 and Mixtral respectively. The models are trained for 2 to 3 epochs using a learning rate of 3e-4 with a decay of 0.001. We observed major performance gains when masking out the instructions and training on only completions. We make publicly available our finetuned Llama3-8B model which performs at par with GPT-4 (gpt-4-0125-preview) and can be run free of cost using a platform such as Google Colab.[4]

3.5 Validation

We use both automated and manual annotation approaches towards validating our models. We only apply automated measures towards our best model, while we measure all model performance against our manual annotations.

We create a validation set drawn from the 188 sample passages in Piper and Bagga (2022). We manually annotate all features (minus the honeypot) using 10 random passages each (for a total of 150 passages).

Replication. We run 15 iterations on a 50% subset of the validation data.

Honeypot. We measure the frequency of a single feature that should never be the right answer (see Table 1) to assess the extent to which our best model may be randomly guessing.

Human Annotation. We employed a group of three student coders who have prior training in text annotation and who were presented the identical prompts as our models' received. To assess agreement among annotators, we report Fleiss' Kappa and the percentage of annotations that resulted in universal agreement.

To assess model accuracy, we report F1 under two conditions: majority vote and minimum match, where we use as reference any human answer that matches the LLM's output regardless of whether it is in the minority. We find upon inspection that given the subjectivity of the ordinal scale that if one trained human annotator approved of a rating then this could reasonably be considered valid.

4 Results

4.1 Validation

Replication. We find that replication occurs in 96.5% of all cases for our best model.

Honeypot. The honeypot answer was labeled 0 (not present) in 100% of cases in our best model.

[4] https://huggingface.co/sbagga/llama3-narrative. Model outputs on the annotated data: https://doi.org/10.6084/m9.figshare.26764231.v1

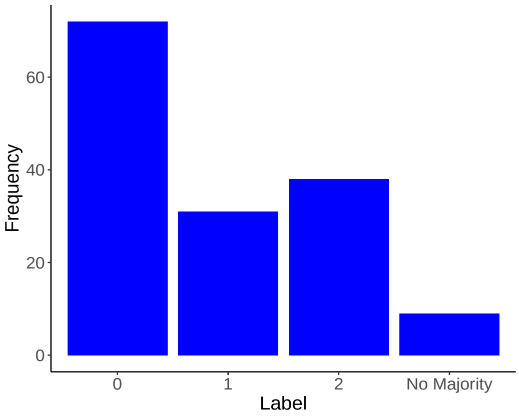

Figure 2: Distribution of majority labels in our annotated data.

Inter-Annotator Agreement. We observe only "fair" levels of agreement between annotators, with a Fleiss's $kappa$ = 0.38 and a universal agreement rate of 43%. We do not observe any dependence between the passage's narrativity score and agreement (i.e. higher narrativity does not produce greater agreement). The distribution of labels is shown in Figure 2.

Model Performance. As we can see in Table 2, GPT-4 was our best performing model, while the fine-tuned Llama3 model using GPT-4 annotated training data achieved proximate performance.

LLM	Majority	MinMatch
GPT4	0.79	0.95
Llama3 8B FT	0.76	0.93
Mixtral 8x7B FT	0.74	0.90
Mixtral 8x7B	0.72	0.87
Llama3 8B	0.51	0.72
Mistral 7B	0.28	0.45

Table 2: Summary of weighted-average F1 scores by model under two reference conditions: majority labels and minimum match where the model matched at least one annotator.

In Table 3, we present the F1 score per feature for our two best models along with the fraction of universal annotator agreement for that feature. As we can see there is considerable variance among tasks when it comes to matching the majority vote, but high performance across the board if we include minority annotations. We find that annotator agreement correlates strongly (r=0.64) with model performance suggesting that the lower performance can be partially attributed to the uncertainty faced by annotators, also supported by the relatively high minority matching scenario across the board.

Feature	Majority	Minmatch	3Agreement
Dialogue	1.0	1.0	0.8
Event Sequences	1.0	1.0	0.5
Emotionality	1.0	1.0	0.5
Anachrony	1.0	1.0	0.7
Pastness	0.95	0.95	0.7
Presentness	0.94	1.0	0.1
Location	0.90	1.0	0.6
Symbolism	0.74	0.83	0.6
Temporal Spec.	0.73	1.0	0.3
Abstraction	0.67	1.0	0.3
Perception	0.64	1.0	0.4
Agents	0.61	0.88	0.6
Eventfulness	0.58	0.85	0.2
Conflict	0.51	0.72	0.2
Concreteness	0.42	0.89	0.0

Table 3: F1 scores by feature for the majority and minority labeling conditions, including the fraction of examples that exhibited universal agreement among annotators.

4.2 Full Data

We present the results of our full prompting experiment in Figure 3 and Figure 4 with respect to our best model. In Figure 3, we query each feature in Table 1 for all narrative passages in our data for a total of 3,008 queries. The figure shows the mean strength score for each feature for all passages. Confidence intervals are calculated by multiplying the standard error for each feature by the z-score for that feature. While Figure 3 only shows results from our best model, we find that our fine-tuned open-source models are strongly correlated with these results as would be expected given our approach of using model distillation for the fine-tuning (as seen in Table 4).

In Figure 4, we show the results of a classification experiment to identify the most distinctive features for predicting narrative passages. Where Figure 3 shows the most common features associated with narrative communication, Figure 4 identifies those features which most distinguish narrative communication from non-narrative. In this experiment, we query each feature in Table 1 for all narrative and non-narrative passages in our data for a total of 342 passages and 5,318 queries. We use a Random Forests classifier with a 75/25 train/test split, which achieves an F1 = 0.95. Figure 4 shows the ranked feature weights for the model.

5 Discussion

The results of our experiments provide valuable information for assessing the discursive priorities of

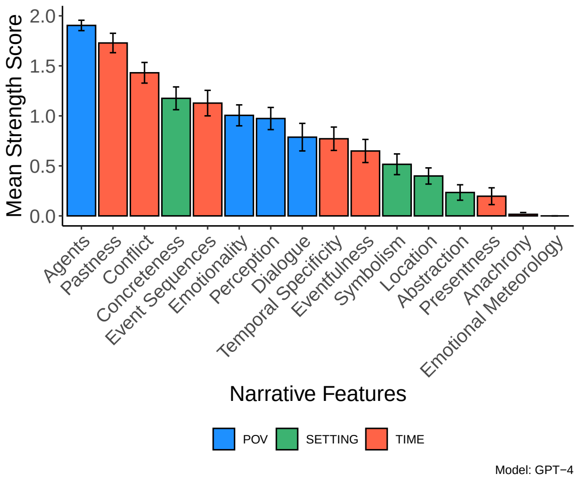

Figure 3: The most common features of narrative passages using our best model (GPT-4).

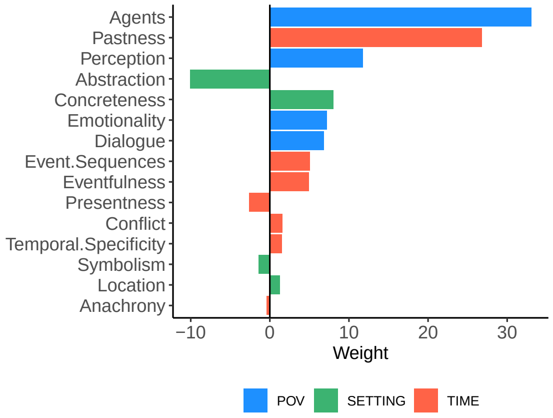

Figure 4: Feature weights for predicting narrative passages. Positive values equal positive predictors and vice versa.

narrative communication. Most notably, they offer further confirmation of the findings of earlier empirical work towards the "deictic theory" of narrative communication (Piper and Bagga, 2022). According to this theory, the principal function of narrative is to focus our attention on the experience of individual agents at a distance, in both time and space. Narrative has a pointing function (i.e. deixis) that furthers goals of social cooperation by creating a framework of "joint attentionality," which cognitive scientists argue is the foundation of developing shared intentions (Tomasello, 2010).

This theory is supported by the prioritization of agents as well as the act pf perception ("seeing the world through the eyes and bodies of the characters"), both of which contribute to the dimension of focalization, of drawing our attention to the *particular* experiences of individuals. As Fludernik (2002) has argued, "There can be narratives without plot, but there cannot be any narratives without a human (anthropomorphic) experiencer of some sort." Interestingly, where prior work had identified perception as a very weak predictor of narrative, the use of LLMs suggests that it plays a much more central role than formerly theorized.

Model Name	ρ
Llama3 8B FT	0.97
Mixtral 8x7B FT	0.98
Mixtral 8x7B	0.94
Llama3 8B	0.72
Mistral 7B	0.14

Table 4: Spearman's rank correlation (ρ) for the open-source LLMs with GPT4 feature ranks.

Concretization and *pastness* similarly work together to construct a distant reality in both time and space. Building a concrete world that one can see and feel is crucial towards constructing that sense of joint attention. The preference for setting these actions in the past tense also helps focalize attention on the "not now." We can see how different discourse features work towards pushing and pulling the mind of the story reader or listener towards somewhere else and away from the present (also crucial for autobiographical narrative where we construct a different self).

In the opposite direction, we see how aspects like *abstraction*, *symbolism*, and *anachrony* are the least associated with narrative discourse, but only abstraction plays a role in discriminating narrativity. When it comes to storytelling, figurative language plays a much more subordinate role to concrete and sensory-based language. The prior emphasis on narrative disorder (*anachrony*) by Genette (1980) appears overstated when looking at a broader sample of text types when compared to deictic techniques of pastness, concretization, and perception.

Of further note is the way the discrimination experiment foregrounds one notable difference between the features' ranks. Where "conflict" has long been theorized as a common feature of narrative (Bruner, 1991), our classification exercise suggests that it is also present within non-narrative communication. In other words, human communication *in general*, at least as represented by the 18 genres in our data, appears to gravitate towards the discussion of conflict rather than this being a unique quality of narrative.

This is yet another way that LLMs have expanded our understanding of narrative communication: as Piper and Bagga (2022) indicate they struggled to model narrative conflict prior to LLMs. Thus its relative importance has remained largely theoretical. That being said, we also note that it indicates one of the lowest levels of agreement with our human annotators and also exhibited very low levels among our annotators. "Conflict" clearly remains a challenging narrative construct worth further study, especially given the importance ascribed to it by narrative theory.

Finally, we note the way in which our classification experiment did not result in a strong clustering of any one of our higher-level classes (POV, setting, time) within the feature ranks. Rather, it appears to be the case that one of the distinguishing features of narrative communication is a reliance on multiple dimensions of discourse (i.e. an intermixing of all three of Genette's linking functions). We observe for example that just under 90% of all narrative passages utilize at least one feature from each of our three classes (POV, setting, time), while non-narrative passages do this just 25% of the time. Narratives are 3.5x more likely to utilize all three types of discourse suggesting both the importance of each class to narrative communication and the importance of multi-dimensionality, i.e. that the mixture of discourse types is essential for narrative communication.

6 Conclusion

In this paper, we have endeavored to frame the concept of narrative discourse as a multi-dimensional aspect of narrative communication. Drawing on the long-established theoretical frameworks of Genette (1980) and Herman (2009), narrative discourse at its highest level consists of three key linking functions that include *time*, *space*, and *perspective* (or tense, mood, voice in Genette's original terminology, see Fig. 1). *Time* links story events with the order in which they are told; *setting* links story events with narrative perspective (of what we see and feel); and *perspective* or *voice* links narrative perspective with narrative structure (characters, dialogue, emotions and other techniques of focalization).

Given the features that we test here, our models provide strong confirmation of prior work emphasizing storytelling's function as a mechanism of developing "joint attentionality" between storytellers and audiences (Tomasello, 2010; Piper and Bagga, 2022). Additionally, the use of LLMs allow us to capture features that previous methods struggled to represent, revising some prior theory and expanding our understanding of narrative discourse more fully. We also provide novel insights into the multi-dimensional nature of narrative communication, i.e. the way it utilizes all three-linking functions to focus our attention on some distant world.

Our work thus suggests that frontier-model LLMs like GPT-4 can be valuable tools for the detection of elementary components of narrative discourse, especially in cases where we lack robust training data for more supervised approaches. Whether as stand-alone applications or as fine-tuning resources for open-weight models, LLMs like GPT-4 indicate reasonable levels of accuracy across a variety of different tasks related to narrative discourse understanding.

Nevertheless, we also observe variable levels of accuracy of our models with respect to different dimensions of narrative discourse. As we note above, much of this appears to be due to annotator disagreement, indicating the subjectivity or ambiguity of the task. Future work will want to delve more deeply into this issue of ambiguity around concepts like "conflict," "eventfulness," or "concreteness," to better understand model limitations and the variance of human responses. For now, we note that with loosened matching criteria models approximate at least some readers' judgments very well.

Based on these experiments, we see LLMs as a valuable addition to the existing tools available for the larger project of computational narrative understanding. Our work provides an initial implementation of the theoretical framework underpinning narrative discourse. Our hope is that future work will continue to expand and revise this approach to achieve deeper understanding of the nature and function of human storytelling.

Acknowledgments

We wish to thank the Social Sciences and Humanities Research Council of Canada for funding to support this research.

Limitations

One of the principal limitations facing our work is the unbounded nature of narrative discourse as a theoretical construct. While we test and validate fifteen constructs that derive from three higher-level categories (time, setting, point-of-view), there may be facets to narrative discourse that are missing from our model. Future work will want to continue to test, expand, and refine the range of narrative dimensions related to narrative discourse.

Second, the use of proprietary LLMs like GPT-4 pose problems with respect to replicability. While we show the same model produces near identical outputs on multiple runs, there is no guarantee that this will be the case with future iterations of the model. Open-weight models thus provide a valuable resource for benchmarking and replicability.

Finally, our work is limited by the need for further cultural breadth in our measurement and validation of narrative discourse. Narrative communication is universally present across all recorded time periods and human cultures, suggesting potential cross-cultural consistency when it comes to the nature of the features of narrative discourse. Nevertheless, our validation of narrative features and our models' ability to approximate them are limited by the culturally specific knolwedge of our annotators and authors. Future work will want to explore the variation not only in the rates of narrative features but also the validity of the features themselves for narrative understanding.

References

Mieke Bal and Christine Van Boheemen. 2009. *Narratology: Introduction to the theory of narrative*. University of Toronto Press.

David Bamman, Ted Underwood, and Noah A Smith. 2014. A bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 370–379.

Jerome Bruner. 1991. The narrative construction of reality. *Critical inquiry*, 18(1):1–21.

Tommaso Caselli, Marieke van Erp, Anne-Lyse Minard, Mark Finlayson, Ben Miller, Jordi Atserias, Alexandra Balahur, and Piek Vossen, editors. 2015. *Proceedings of the First Workshop on Computing News Storylines*. Association for Computational Linguistics, Beijing, China.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 602–610.

Katherine Elkins. 2022. *The Shapes of Stories: Sentiment Analysis for Narrative*. Cambridge University Press.

David K. Elson and Kathleen R. McKeown. 2010. Building a bank of semantically encoded narratives. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Monika Fludernik. 2002. *Towards a 'natural' narratology*. Routledge.

Mikaela Irene Fudolig, Thayer Alshaabi, Kathryn Cramer, Christopher M Danforth, and Peter Sheridan Dodds. 2023. A decomposition of book structure through ousiometric fluctuations in cumulative wordtime. *Humanities and Social Sciences Communications*, 10(1):1–12.

Gérard Genette. 1980. *Narrative discourse: An essay in method*, volume 3. Cornell University Press.

Rachel Giora and Yeshayahu Shen. 1994. Degrees of narrativity and strategies of semantic reduction. *Poetics*, 22(6):447–458.

Jonathan Gottschall. 2012. *The storytelling animal: How stories make us human*. Houghton Mifflin Harcourt.

Amit Goyal, Ellen Riloff, and Hal Daumé III. 2010. Automatically producing plot unit representations for narrative text. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 77–86.

David Herman. 2009. *Basic elements of narrative*. John Wiley & Sons.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Zhichao Hu, Elahe Rahimtoroghi, and Marilyn Walker. 2017. Inference of fine-grained event causality from blogs and films. In *Proceedings of the Events and Stories in the News Workshop*, pages 52–58.

Peter Hühn. 2009. Event and eventfulness. *Handbook of narratology*, 19:80.

Peter Hühn, Jan Christoph Meister, John Pier, Wolf Schmid, and Jörg Schönert. 2009. The living handbook of narratology. *Hamburg: Hamburg University. URL: http://www.lhn.uni-hamburg.de (Retrieved on 06.03. 2024)*.

Matthew Jockers. 2017. Package 'syuzhet'. *URL: https://cran. r-project. org/web/packages/syuzhet*.

Matthew L Jockers and David Mimno. 2013. Significant themes in 19th-century literature. *Poetics*, 41(6):750–769.

Stephanie Lukin, Kevin Bowden, Casey Barackman, and Marilyn Walker. 2016. Personabank: A corpus of personal narratives and their story intention graphs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1026–1033.

Margaret Meehan, Dane Malenfant, and Andrew Piper. 2022. Causality mining in fiction. In *Text2Story@ ECIR*, pages 25–34.

Jessica Ouyang and Kathleen McKeown. 2015. Modeling Reportable Events as Turning Points in Narrative. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2149–2158, Lisbon, Portugal. Association for Computational Linguistics.

Tanmay Parekh, I-Hung Hsu, Kuan-Hao Huang, Kai-Wei Chang, and Nanyun Peng. 2023. Geneva: Benchmarking generalizability for event argument extraction with hundreds of event types and argument roles. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3664–3686.

Federico Pianzola. 2018. Looking at narrative as a complex system: The proteus principle. In *Narrating complexity*, pages 101–122. Springer.

Andrew Piper and Sunyam Bagga. 2022. Toward a data-driven theory of narrativity. *New Literary History*, 54(1):879–901.

Andrew Piper and Olivier Toubia. 2023. A Quantitative Study of Non-linearity in Storytelling. *Poetics*, 98:101793.

Gerald Prince. 2012. *Narratology: The form and functioning of narrative*, volume 108. Walter de Gruyter.

Jason Radford and Kenneth Joseph. 2020. Theory in, theory out: the uses of social theory in machine learning for social science. *Frontiers in big Data*, 3:18.

Elahe Rahimtoroghi, Jiaqi Wu, Ruimin Wang, Pranav Anand, and Marilyn Walker. 2017. Modelling protagonist goals and desires in first-person narrative. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 360–369.

Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. 2016. The Emotional Arcs of Stories are Dominated by Six Basic Shapes. *EPJ Data Science*, 5(1):1–12.

Nils Reiter, Marcus Willand, and Evelyn Gius. 2019. A shared task for the digital humanities chapter 1: Introduction to annotation, narrative levels and shared tasks. *Journal of Cultural Analytics*, 4(3).

Benjamin M Schmidt. 2015. Plot Arceology: A Vector-Space Model of Narrative Structure. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 1667–1672. IEEE.

Dominik Stammbach, Maria Antoniak, and Elliott Ash. 2022. Heroes, villains, and victims, and GPT-3: Automated extraction of character roles without training data. In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 47–56, Seattle, United States. Association for Computational Linguistics.

Yidan Sun, Qin Chao, and Boyang Li. 2024. Event causality is key to computational story understanding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3493–3511, Mexico City, Mexico. Association for Computational Linguistics.

Michael Tomasello. 2010. *Origins of human communication*. MIT press.

Albin Zehe, Leonard Konle, Lea Katharina Dümpelmann, Evelyn Gius, Andreas Hotho, Fotis Jannidis, Lucas Kaufmann, Markus Krug, Frank Puppe, Nils Reiter, et al. 2021. Detecting scenes in fiction: A new segmentation task. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3167–3177.