Towards a Data-Driven Theory of Narrativity

Andrew Piper and Sunyam Bagga

New Literary History 53.4 (2022): 879-901

For well over half a century, arguably beginning with the publication of the special issue of the journal Communications dedicated to the "structural analysis of narrative" in 1966, literary scholars have developed a myriad of theories to describe narrative as a distinct form of human communication.[1] However different the theoretical or methodological orientations, what unites them is a shared belief in the linguistic commonalities that comprise the practice of narration across time and space. They share a commitment, in the words of Gérard Genette, to understanding "the oppositions through which narrative defines and constitutes itself in the face of various nonnarrative forms."[2]

And yet this very theoretical richness now poses a distinct research challenge. Scholars continue to generate theoretical frameworks largely by extrapolating from a small number of documents to explain large-scale narrative behavior within a variety of cultural settings. What has largely been missing from this research is the important next step of testing the degree to which these different theories fit different collections of documents within different cultural settings. It's time to move from using a few enigmatic examples to engaging in the process of experimental testing using much broader collections of documents.

In this essay, we wish to provide a framework for the empirical testing of narrative theory using the process of machine learning and predictive modeling.[3] Generally understood, predictive modeling as it relates to literary studies involves the process of testing different kinds of textual features to predict different qualities pertaining to large numbers of literary

documents.[4] These qualities might be levels of prestige or cultural capital,[5] the persistence of genre coherence,[6] the demarcations of historical time-periods,[7] or readers' sense of transport or emotional investment.[8] Predictive modeling allows us to ask, Which features and which theoretical frameworks have more predictive power when it comes to understanding a given textual quality across a diverse array of documents? Do we see commonalities among predictive features across different document types? Or do we see a strong degree of variability, suggesting that the quality we are interested in varies significantly depending on the context or situation?

The probabilistic framework behind predictive modeling means that we do not need to subscribe to the idea that there are universal qualities surrounding narrative communication -- that there are a fixed set of necessary conditions that occur always and everywhere. Nor does it mean that we have to subscribe to the negative theory that there are no intrinsic qualities at all, i.e. that anything can be narrative. Instead, predictive models allow us to conceptualize narrative communication as a function of the likelihood of certain features being present that can occur in a variety of potential configurations and that trigger readers' beliefs to varying degrees.

We thus see predictive modeling as a valuable tool in the literary critical toolkit to address the tripartite problems of theory validation, theory reduction, and theory development. Predictive models can help us, first and foremost, to validate different theoretical frameworks with respect to cultural behavior. By testing theories against real-world practices, we can observe how well a theory describes the world. At the same time, predictive models can also help us reduce the cacophony of competing theories, to bring more clarity to our understanding of cultural behavior. Predictive models will help us foreground more viable theories and discard less tenable ones. Finally, predictive models can help us develop new theories derived from observing considerably larger collections of documents and in considerably different ways using

computational techniques. We might call this "born-digital theory": theories of literary expression that emerge from computational approaches to texts rather than bibliographic modes of engagement. We expect a great deal of new theories to emerge in the coming years that derive from exploratory computational approaches to textual understanding.

Our goal in this project is to develop a first step towards what we call a minimal theory of narrativity, to identify what David Herman calls "the elements of narrative" that cohere across as diverse array of cultural settings as possible.[9] One of the principal shifts to occur in the field of narratology over the past several decades has been an emerging understanding of narrative as a matter of degree rather than of kind.[10] "Narrativity" according to these theories is a quality that can best be understood not as a global binary class (a document either is or is not narrative), but as a local, multi-dimensional scalar property.[11] As Ochs and Capps write, "We believe that narrative as genre and activity can be fruitfully examined in terms of a set of dimensions that a narrative displays to different degrees and in different ways."[12] In this sense, a narrative document, such as a novel, may exhibit greater or lesser degrees of narrativity at different moments, just as ostensibly non-narrative documents, such as scientific reports, may also exhibit degrees of narrativity at certain points in the text.

As we will show with respect to a collection of over 13,000 documents drawn from an array of different genres and time periods, our models suggest that there is a very small set of features that are highly predictive of narrativity across document types and that these features strongly align with readers' judgments. Narrativity, according to our models, is a highly predictable form of communication, one that is reducible to a select few core dimensions (with many subsidiary nuances). As we discuss at the close of our essay, these dimensions can be described by what we call the "distant worlds theory," where narrativity is most strongly and

consistently identified through the concretized actions of an agent at a distance to the teller. We explore the implications of these findings with respect to existing narratological theory as well as highlight the limitations of what our models cannot tell us. As we make clear, these are preliminary findings that we look forward to validating with future work drawn from increasingly diverse cultural settings.

Modeling Narrative, Part 1: Feature Space

How do we know when someone is telling a story? Consider, for example, the following two passages.

A.

> One time when I was in 6th grade I was getting ready for school in the morning. I remember getting breakfast and sitting down at the kitchen table. I heard footsteps going down to the basement. At first I thought it was my dad because usually he would wake up to me going downstairs. But I heard my dad's loud snoring from the bedroom.

B.

> When philosophers, well-known for being averse to silence, enter into a conversation, then they should speak as if they were being proved wrong, but in a manner which convicts the opponent of untruth. The point is not to generate cognitions which are

> absolutely correct, bulletproof and watertight – these run unavoidably into tautology – but rather those which direct the question of their correctness towards themselves.

Most readers would readily agree that the first passage has a high degree of narrativity and the second does not. (The first is drawn from the AskReddit thread, "What's your creepiest (REAL LIFE) story?" while the second is from Adorno's Minima Moralia). If asked to explain themselves, readers might emphasize different features, like the prevalence of first-person pronouns in A or the frequency of abstract nouns in B or the locational information in A (sixth grade, kitchen table, basement, bedroom, downstairs). They may focus on the actions as well, where A focuses more on past-tense sensory inputs (sitting, hearing, snoring) while B focuses far less on concrete actions (philosophers are well known, the point is not, they should speak). Finally, readers may focus on higher-level concepts such as urgency, concreteness, or suspense to describe their response.

As Michael Gavin has highlighted, the foundation of any reading model, whether cognitive or computational, is a combination of what he calls "feature space" and "document space."[13] As readers, we understand new texts based on the texts we have already encountered just as we understand these new texts through the qualities or features we choose to focus on. As we age, we encounter more and more texts and develop, in neurotypical cases, increasingly sophisticated frameworks for understanding and relating to them. The features we focus on manifest theories about the textual qualities that matter, in the process creating particular representations (or views) of texts. The choice of texts used to inform our understanding similarly manifest theories about the kinds of texts that matter, creating representations of larger social categories out in the world. Both are manifestations of prior beliefs, i.e. theoretical

frameworks, and both represent the constraints on what a model or mind can "learn" and thus what we can know through them.

One of the challenges of computationally modeling texts is that we do not have a priori information about the dimensionality and qualities of this so-called "feature space" (also true as we will see about document space). While not infinite, the number of potentially meaningful features in a given passage is extraordinarily large. At the same time, the number of likely meaningful features, i.e. those that help us make distinctions and draw judgments about the types of texts we are reading, may be quite small.

This is the first way that theory matters for computation. Theory provides the belief-system that constrains the plenitude of potential features that can be used to describe textual meaning. Theory may be wrong, which is why we test it, but it offers a framework and a set of boundaries for building models, which can then be iteratively updated.[14] If we wish to model the weather, theories of weather patterns will help us make decisions about better or worse choices with respect of meteorological features.

In this section, we describe the features used to represent our documents, where our choices are guided by prior narrative theory, beginning with the work of Gerard Genette and moving through more recent work by David Hermann and Monica Fludernik.

In his foundational work on narrative discourse, Genette builds on the Russian formalist distinction between story and discourse to add a third dimension of narrating, emphasizing the perspectival aspect of narrative (Fig. 1).[15] Genette links these "big three" dimensions with three further subsidiary qualities, which he calls tense, mood, and voice, drawing on linguistic terminology. Tense for Genette captures the relationship between the order of events (story) and the order of their telling (discourse) and is thus fundamentally concerned with notions of time.

Mood on the other hand represents the relationship between the narrating instance and the events of the story. For Genette this can be productively translated into Plato's terms diegesis and mimesis, where diegesis (or "pure narrative [haplé diégésis]") represents the events proper and mimesis captures scenic dimensions like setting and details (descriptions) as well as indications of communication (saying that someone said something). Mood thus captures concepts related to both "setting" and "eventfulness." Finally, voice refers to the relationship between the narrator and the narrative discourse, what Genette calls "the generating instance of narrative discourse" (213). This can be understood to capture simple concepts like "person" (first, second and third), levels or scenes (intradiegetic differences among narrators), as well as more complex constructs like focalization. Voice thus represents the domain of "perspective" and "personhood" as it inflects language use. For ease of comparison, translate Genette's terminology of tense, mood, voice to time, setting, and perspective.

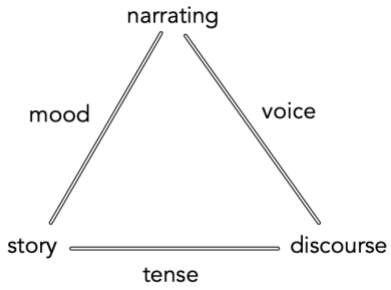

Fig. 1 Genette's narrative triangle

Over the ensuing decades, numerous theorists would complement, modify and update Genette's initial model. Herman for example, building on Fludernik's emphasis on "experientiality" as a core quality of narrative,[16] would introduce the tripartite concepts of event sequences (time), world building (setting), and feltness (perspective). One can see how Herman's

categories align reasonably well with Genette's initial model.[17] Herman importantly introduces a fourth category, which he calls "situatedness," emphasizing that the social context in which narrative occurs will have an impact on the nature and interpretation of the act of communication. Other theorists have in turn elaborated particular dimensions more fully, where for example Hühn has developed a theory of event types,[18] Sternberg, Brewer, and Ricoeur an emphasis on temporality,[19] and Laure-Ryan a theory of fictional worlds.[20]

In Table 1, we present an overview of the features we use to capture these broader theoretical frameworks, aligning each under the tripartite headings of time, setting, and perspective. While we describe each feature in greater detail in the supplementary material, as one can see from the table our effort here has been to construct features that capture each of the three primary dimensions of Genette's triangle and Herman's four-part schema (minus situatedness, which we capture through our data selection). What is important to emphasize is that these features are neither canonical nor exhaustive. There may be many more (and potentially "better") ways to capture, for example, narrative time or setting or perspective. We think far more work can be done in terms of feature engineering to better capture existing theoretical frameworks with respect to narrative. As one can also see from the table, we also include features that lie outside of existing theoretical frameworks, i.e. features for which we do not have theoretical explanations for their meaning or role, though as we show later we can potentially incorporate some of their dimensions within existing schemas.

Finally, it is important to emphasize that in our models we do not capture questions of "change" which have been central to narrative theory since Aristotle. For our purposes, the value of the concept of narrativity is that it can be understood as a local textual property that is manifested through particular linguistic practices, whereas notions of "change of state" can be

understood as a broader document-level property (or at least sub-document, i.e. longer stretches of text). We thus see our work as a complement to emerging computational research on modeling narrative "change."[21]

Category	Name	Description
	Temporality	NER Tags for “time” “date” “duration”
Time	Temporal Order	TimeML TLINKs
	Setting	NER Tags “location”, “organization” (pobj), Wordnet places lexicon
Setting	Concreteness	Brysbaert's lexicon
	Saying	bookNLP supersense “verb.communication”
	Eventfulness	rate of non-helping verbs / rate of non-stative verbs
	Feltness	perceptual, cognitive, emotion lexicon
Perspective	Agenthood	animacy detection
	Agency	nsubj <–> vb distance
	Coherence	cross-sentence entity recurrence
	pos	part-of-speech ngrams (1-3)
Other	dep	dependency relationship ngrams (1-3)
	word	word ngrams (1-3)

Table 1. Feature space employed in our models.

Modeling Narrative, Part 2: Document Space

The next component of our model is the document space that will be learned and tested. Our data includes passages from 18 different genres that capture a diverse array of communicative "situations," including classical narrative genres like short stories, fairy tales from around the world, and novels (historical and contemporary) along with less conventional fictional genres like "flash fiction" and non-fictional genres like AskReddit posts that tell stories in response to prompts such as "What's your creepiest (REAL LIFE) story?" On the non-narrative side, we include genres such as scientific abstracts and legal contracts in addition to more ambiguous non-narrative genres such as book reviews, aphorisms, academic articles from the humanities, and US Supreme Court decisions.

Our goal in assembling this data is to generate as heterogenous a collection of texts as possible in order to better understand the variability or consistency of narrativity across diverse communicative situations. Capturing all possible narrative "situations" is impossible. What we can do with our models, however, is begin to gain confidence about the generalizability of our findings across multiple settings. Are there document types for which our insights about narrativity mutate considerably? Are there document types that when excluded from the learning process have a significant impact on the consistency of our model's judgments when it comes to identifying the core qualities of narrativity? These are the types of questions that the fungibility of computational modeling allows us to ask. While we can never be certain that the qualities we describe here are good descriptions of all narrative situations ever, we can gain confidence about our ability to describe the likelihood of qualities present in narrative situations.

A second aspect of our document space is the length of each document. Because our interest is in local "narrativity," i.e. the extent to which a span of tokens expresses narrative communication rather than describing document-level properties (this is/is-not a narrative), we represent our documents as randomly selected sequences of a fixed number of sentences, in our case 3-6 sentences, with a vast majority being 5 sentences in length. This too derives from narrative theory through the emphasis on narrativity as a local reading experience that can change over document time. In total, our data encompasses 13,543 passages as described in Table 2.

	NARRATIVE			**NON-NARRATIVE**	
CODE	**DESCRIPTION**	**# Docs**	**CODE**	**DESCRIPTION**	**# Docs**
BIO	Biographies	898	ABSTRACT	Scientific Abstracts	950
FABLE	Fables	258	APHORISM	Aphorisms	462
FAIRY	Fairy tales	740	BREVIEW	Book reviews	776
FLASH	Flash fiction	390	LEGAL	Contracts	1054
HIST	Popular History	979	LITSTUDY	Academic articles (Lit)	468
MEM	Memoirs	935	PHIL	Academic Articles (Phil)	519
NOVEL_CONT	Contemporary Novels	776	SCOTUS	Supreme Court	942
NOVEL_19C	19C Novels	998			
REDDIT	AskReddit	971			
ROC	Artificial Stories	976			
SHORT	Short Stories	451			

Table 2. Overview of our 18 genres used in this study.

One of the challenges in building predictive models is the quality of the annotated data. The model learns based on the classes that it is supplied (which is why it is known as "supervised learning"). While we have selected passages from genres where we have strong beliefs about their narrativity, we cannot be absolutely certain that each example is perfectly aligned with its label or that all passages behave equally with respect to their investment in narrativity. Indeed, we assume that at least some examples of non-narrative passages will exhibit some degree of narrativity and vice versa.

To address this problem, we manually annotate a small set of data using a team of three student annotators. The full process and codebook is described in the supplementary material. After several weeks of training, we annotated 394 passages according to a 5-point Likert scale capturing the degree of a passage's narrativity. To assess inter-rater agreement, we use the average deviation index as suggested by Thomas O'Neill.[22] Doing so, we find a median deviation of 0.37 and a mean of 0.41 (+/- 0.31). This suggests that readers were on average within less than half a point on the Likert scale across all annotations, suggesting to us reasonable levels of agreement.[23]

We present examples here of high, medium, and low scoring passages according to our readers' assessment of narrativity along with the average deviation between reader scores (where a value of 0.5 would mean on average readers disagreed by half a point on our five point Likert scale). All of the annotations are included in the supplementary material.

Avg Score 5.0, Avg Deviation = 0. In the center of the town, the Mercedes stopped a second time, outside a charcuterie and an adjoining boulangerie. Again Keller sped past, but Gabriel managed to conceal himself

in the lee of an ancient church. There he watched as the woman climbed out of the car and entered the shops alone, emerging a few minutes later with several plastic sacks filled with food.

Avg. Score = 3.0, Avg. Deviation = 0.84. There were other dramatic glitches, too. Despite Cornell's love for the part, she was not suited to it. While Anouilh's Antigone epitomized the enfant terrible, Cornell was in her early fifties and brought to the role a calm, dignified strength, making it harder for the audience to feel that she was imperiled. Photographs of the production reveal her imposing, statuesque presence, precisely the opposite of "la petite maigre" called for by Anouilh.

Avg. Score = 1.2, Avg. Deviation = 0.38. To understand a thing is to discover how it _operates_. The eternal forms of things are laws of natural action. Such are the law of gravitation, the laws of optics or of chemical combination. A static picture unless so interpreted must be at once valueless and meaningless. It follows that Thought and Discourse, in furnishing us with Knowledge, must themselves be active, and must in some way or other reproduce the activity of Nature.

Results

The aim of our project is to identify the smallest set of features that exhibit the highest predictive accuracy when it comes to estimating a passage's narrativity and to assess the stability and consistency of these features across a variety of conditions. In other words, as with any theory we want to test the generalizability of our models to capture the idea of "narrativity" under varying conditions. We thus create a workflow where we assess the potential influence of different factors on the generalizability of our models. We summarize results here with details and further visualisations provided in the supplementary material. In the next section, we discuss the implications of these findings for narrative theory, where we develop our "distant worlds" theory of narrativity more fully.

A. Does the choice of algorithms effect our models?

We begin by testing the extent to which different machine learning algorithms impact the accuracy of our models across different sets of features. We experiment with three popular machine learning algorithms for text classification (logistic regression, support vector machines, and random forests) and feature spaces that can be broadly categorized into three groups: lexical features (lexical uni/bi/trigrams), syntactical features (part-of-speech uni/bi/trigrams and dependency-relationship tags uni/bi/trigrams), and our higher-level custom features described in Table 1. We validate our models in these experiments using a process of five-fold cross validation on the experimental data.

Table 3 shows the average F1 score for different feature combinations and algorithms (a full list is in the supplementary material). In each case, the task is to predict the narrativity of a passage. As we can see, the best performing model achieves an F1 score of 0.936, meaning that in more than 9 cases out of 10, with just a few sentences our model can detect whether a string of text belongs to the class "story" across 18 different genres. The best performing model combines all three of our custom features described above along with part-of-speech unigrams (for a total of 48 features). More granular features, such as bigrams and trigrams, perform less well when we cap the overall feature space at 100 variables (to avoid overfitting). Third and finally, we can see that choice of algorithm has little effect on the accuracy of the model.

Feature Set	**LogReg**	**SVM**	**Random Forests**
POS1_TimeSettingPerspective	0.933	0.932	0.936
POS1	0.922	0.923	0.925

Word1	0.909	0.911	0.914
POS3	0.895	0.894	0.895
DEP3	0.842	0.842	0.835
WORD3	0.809	0.807	0.806

Table 3. Predictive accuracy of different feature sets using three machine-learning algorithms. Values represent average F1-scores using five-fold cross validation. POS1 / WORD3 refer to the n-grams used for that particular feature.

On the one hand, this degree of accuracy confirms prior narrative theory that has a strong belief in narrative's distinctive linguistic behavior.[24] Narrativity is a very recognizable form of communication. On the other hand, this finding runs counter to assumptions in literary studies about the complexity or ineffability of literary behavior. The highly predictable nature of narrative communication aligns with other data-driven findings that underscore the importance of legibility when it comes to literary behavior, whether in the form of genre, readership, or even fictionality.[25] We can see a consistent picture emerging of cultural practices whose primary function appears to be the overt signaling of their categorical status.

B. Do our models reflect reader judgments?

The next aspect of our models that we explore is the relationship between our models' predictions and our readers' judgments. What kind of confidence can we have in the relationship between the predicted probability of a passage being narrative (the model's "confidence") and

our readers' judgments of the passage's "narrativity"? One of the central tenets of this project is that narrativity is a scalar property. Can our models capture this experience of "degree"?

When we compare reader judgments with the model's predicted probabilities of a passage's narrativity (Fig. 2), we observe a correlation coefficient of 0.79 using Spearman's rho. This translates into an overall F1 score of 0.89 if we treat documents labeled with above 3.0 as narrative and those with 3.0 or below as non-narrative, indicating similar performance with our model tested on the large labeled data using cross-validation. We thus do not see a meaningful decline in performance when we condition on reader judgments compared to our less precisely labeled data.

We can also report that while false positives (cases where our model thought the passage was narrative and our readers did not) are relatively evenly distributed across different kinds of documents, the false negatives (cases where our model that the passage was not narrative but our readers did) are dominated by "book reviews" (accounting for 17 of 24 misclassifications) suggesting that this one genre poses uniquely confounding problems for our model (or our readers). Were we to remove book reviews altogether our recall would increase considerably to 0.97, giving us an overall F1 of 0.93 and a rho of 0.82. We also note that our false positives are largely clustered between 2 and 3 on the Likert scale (with most being greater than 2.5), suggesting that discrepancies between our model and readers is mostly occurring in areas where readers are weakly confident about a passage's non-narrativity.

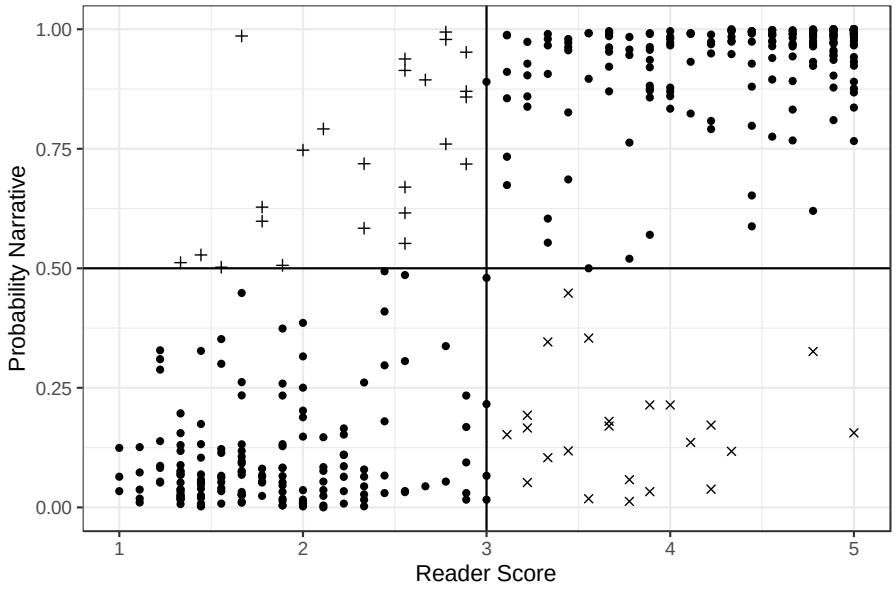

Fig. 2 Correlation between reader judgments of narrativity and our best model's predicted probability of being narrative. Dark circles represent overlap between reader judgments and the model, while +'s and x's indicate false positives and false negatives respectively.

C. Which features have the most predictive power?

Our larger goal in this project is to derive what we are calling a "minimal theory" of narrativity, i.e. we want to build the most parsimonious model possible that largely retains its predictive power (it's generalizability across numerous document types). To do so, we begin by ranking the feature weights associated with our best model. To better understand just how few features we would need to achieve close to our maximum predictive accuracy, we then test each feature beginning with the top most weighted feature and adding one feature at a time as we move down the list and rerunning our model. Keep in mind that we use the large dataset described in Table 2 as our training data and the reader annotated data as our test data, i.e. we validate our model's accuracy on reader judgments. And as with all subsequent experiments, we remove book reviews from the reader data for reasons discussed above.

As we can see in Figure 3, with just a few features (anywhere between 2-5), we can achieve close to our maximum F1 score. On the y-axis we present the model's accuracy, and on the x-axis we present individual features that we add together one step at a time (i.e. we begin with just one feature (agenthood) and then add one feature at a time measuring the F1 score at each step). According to this process, the five features most predictive of narrativity are:

Positive Predictors:

1. agenthood (the rate of animate entities)
2. vbd (the rate of past tense verbs)
3. concreteness (the rate of concrete vocabulary)

Negative Predictors

4. nn (the rate of nouns)
5. vbz (the rate of participles)

We will return to the significance of these features in our next section, but underscore again the surprising parsimony of the model.

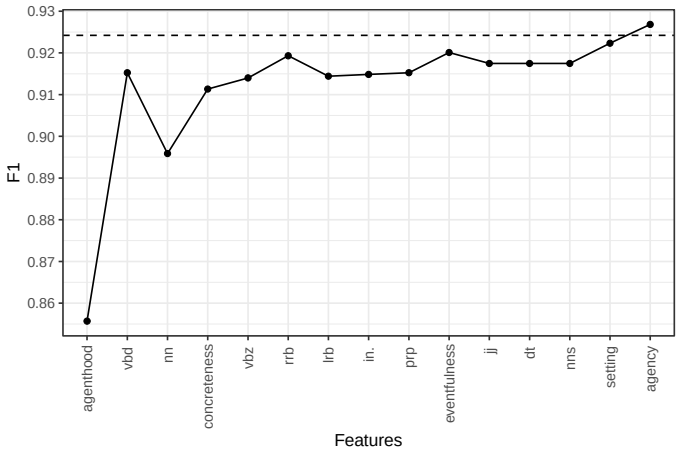

Fig. 3 Each step on the x-axis is equal to the addition of a feature to all prior features. The y-axis reports the F1 score for those feature combinations. The model is validated on the reader annotated data. The dashed line represents the F1 score using all 48 features.

D. What effect does genre have on our models?

Before we turn to a deeper examination of our features, one further question we want to ask is whether the model's performance is tied to the specific configuration of the training data. For example, is our model simply learning something about genre rather than narrativity? As with the above tests, more detailed explanations along with graphs of our experiments are included in the supplementary material.

One way of manipulating the data to test the interaction between genre and the features that predict narrativity is to selectively remove genres from the training data and then calculate the feature weights for each new hypothetical model. Are the leading features invariant to different combinations of genres used in the training data? To test this, we systematically remove combinations of two genres from our training data and then test our model on the reader data. We find that the top 5 features described above account for 95% of all occurrences of the top five weighted features in our manipulated data. In other words, even when we remove different pairs of genres we almost always end up with the same five most predictive features regardless of the types of texts that are learned. This suggests that these features' predictive power appears to be largely independent of the genres being learned. No matter what combination of genres we remove, we continually see these features as the most strongly weighted features in our model.

A second way we can test the influence of genre on our model is to calculate the accuracy of the model after pairs of genres have been removed. Rather than look at the consistency of features, we can look at the consistency of the model's accuracy. Thus we are asking how well our model can predict the narrativity of flash fiction even if it has never learned what a passage of flash fiction looks like. Doing so, we find that when any two genres are removed from the training data model performance only degrades by a maximum of 2%. The lowest performing models in this scenario all have in common that the genre of scientific "abstracts" have been removed suggesting that the model's understanding of non-narrativity may be somewhat dependent on the inclusion of this genre, though the decline in performance is still small (~2%).

E. Testing for alternative features

Finally, while our feature weights and genre tests allow us to observe the robustness of our small group of top features to generalize across document types, we assume that there may be other combinations of features that have predictive power when it comes to narrativity. In other words, while we can identify the best features (among all the features we have available), it is also useful to understand just how much better they are than any other combination of features.

To test this, we first remove our top five features from our best model. We then iterate through our remaining feature space creating a variety of other four-feature-combinations. While we find that some sets of four-feature-combinations perform within 2-3% of our best model, the bulk are considerably lower (less than 100 combinations are within 5%). Among top models using this method, personal pronouns are overwhelmingly represented, which likely serve as a proxy for agenthood. In addition to pronouns, we see features related to concretization such as

setting and temporality that are also prominent, as are features related to events such as eventfulness and agency (with feltness being a powerful negative predictor). In other words, models that perform less well but reasonably closely to our best model are comprised of similar kinds of features to our best model lending further support to the generalizability of our models.

F. Limitations

As with all models, it is important to underscore the limitations at work here. No model can capture all of reality (the territory and map can never coincide). Our theory of narrativity is constrained by the documents that the model has learned along with the way those documents are represented through the construction of our features. When we conduct permutation tests on our features, we see that the same features continue to be predictively powerful independent of the genre constellation of our data. This suggests that our model should be robust when approaching unseen genres. Furthermore, when we remove our best features we discover other features that align conceptually with our best features, suggesting that it would be surprising if new, untested features told a dramatically different story than our model.

Nevertheless, future work will want to continue to validate our model's generalizability. Are there particular genres where our model fails? When and under what conditions? How cross-cultural or transhistorical are our models? Since our learning has only occurred using a single language largely focused on more contemporary documents, it is possible that other languages and cultural settings may behave differently. Second, how might new, more nuanced features give us a deeper understanding of narrativity? Whether we add new features or represent concepts like "setting" or "eventfulness" differently, we want to be open to the idea that our

understanding of narrativity may change. Finally, if we expand the set of readers who are providing our understanding of narrativity, would we continue to see the same level of agreement between our models and readers' feelings about narrativity? Future work will want to look into validating this reader experience of narrativity more deeply.

The Distant Worlds Theory of Narrativity

Turning now to our "best model," what can we learn from these experiments about the nature of narrative communication? And how do they align with received theoretical frameworks that we discussed above?

Looking at the top three positive predictors of our model, we can say that what is most predictive of narrativity is a highly focalized, concretized, and distantiated form of agency. According to our models, narrativity is associated with:

- the repetitive foregrounding of one or more actors;
- the concretization of actions within the world;
- the temporal distantiation of those actions from the perspective of the teller.

If we were to state this as a theorem, we would say: We are most confident of narrativity when we encounter a highly focalized set of agents set at a distance to the teller who are situated within a particular time and place.

Overall, we find that our models align well with Fludernik's emphasis on the concept of "experientiality" at the heart of narrativity. As Fludernik writes, "Experientiality reflects a

cognitive schema of embodiedness that relates to human existence and human concerns. [...] In my model there can be narratives without plot, but there cannot be any narratives without a human (anthropomorphic) experiencer of some sort."[26] As we can see in the list of most highly weighted features (Fig. 4), agenthood -- the presence of an animate agential entity -- was the single most important dimension of narrativity. Such agency is further underscored by the predictive value of eventfulness, i.e. the rate of actions that narrative encodes, lending support to Hühn's emphasis on events and eventfulness as a key framework for understanding narrative communication.[27]

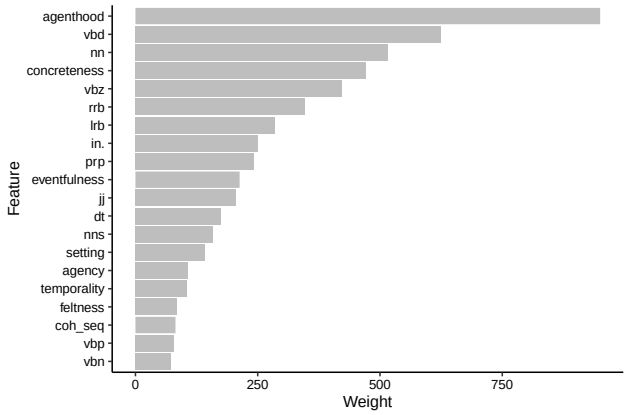

Fig. 4 Features ranked by their predictive weight in our best model (Random Forests).

On the other hand, the emphasis on "sequentiality" posited by Herman as one of the core dimensions of narrativity was less strongly indicated by our models. Indeed, the explicitness of temporal order was considerably less predictive than pastness (vbd), but also invocations of time and the rate of actions (eventfulness). That being said, it is hard to imagine how strings of past tense verbs are not somehow intrinsically sequentially related to one another, a facet that may be

going un(der)detected by our models.[28] Nevertheless, if we are trying to formulate the "elements" of narrative and thinking about the dimension of time, we would argue that temporal distance is far more fundamental than temporal sequence.

When it comes to the question of "setting" (or "world building" in Herman's terms), we observe two important points. First, it is interesting that concretization is a far stronger predictor than location (captured by our feature setting). The world that is constructed by narrative is first and foremost physical and geometric as opposed to geo-located and positioned. As we have argued elsewhere and as the work of Denis Tenen has shown, it lends support to the theory of narrative as a vehicle of embodied cognition where what is being dramatized in narrative is not an entire simulated "world," but rather a continual stream of extended cognition through an available object-world.[29] As Randall Beer puts it, "the focus [in embodied cognition] shifts from accurately representing an environment to continuously engaging that environment with a body so as to stabilize appropriate co-ordinated patterns of behavior."[30] Narrative concretization allows an experiencing agent to keep the sensory channel "open."

Finally, we would argue that our models have an important bearing on the question of "perspective," "consciousness," or "mind" when it comes to narrativity. Herman argues that a core element of narrative is "the pressure of events on a real or imagined consciousness" (14), suggesting that what matters is the way experience registers itself on a consciousness, which he calls "feltness," strongly aligning his work with research on literature and "theory of mind."[31] We would argue that one way to conceptualize this is not in terms of explicit subjective experience (an agent who thinks or feels something), but rather more in terms of thereness, a deictic function of narrative that situates an agent in time and space at a remove from the teller of the story.

According to our models, it's not inward moving, but outward pointing that matters when it comes to narrative.

Seeing narrative's deictic function aligns in interesting ways with the research of Tomasello on the attentional origins of human communication.[32] For Tomasello, whose research is based on extensive reviews of animal studies, human communication is fundamentally oriented around attention-directing. It can be understood as the linguistic equivalent of gestural pointing. This framework is underscored by the fact that our feltness measure, which captures explicit invocations of thinking and feeling is actually negatively predictive of narrative. While it is entirely possible that readers (listeners or viewers) may project feltness onto a narrative agent, they do so not through the explicit invocations of such experience, but through the invocation of concretization, of an agent being explicitly placed in a world in both time and space. Such concretization becomes a key aspect of narrative's communicative foundations. It allows us as readers (or listeners or viewers) to feel through the agent rather than feel in the agent.

To help illustrate these points, we conclude with an example drawn from our data. The passage comes from Dagoberto Gilb's "Winners on the Pass Line," a short story about migrant labour in the U.S.[33] The passage reflects the main character's return home to Mexico and was rated high in narrativity by both our readers (an average score of 5 with no disagreement) and our model (99% probability of being narrative):

> He drank only one beer on the airplane. He sat next to a window in the non-smoking section and no one sat next to him. Sometimes he shut his eyes, but he didn't sleep. When he opened them he'd see the wing of the airplane shudder. There weren't enough clouds

to look at, and the earth seemed only slightly more alive than one of his hundred-dollar bills.

This passage nicely captures the aspect of experientiality discussed above, as each sentence repetitively returns to the embodied experience of the main character (he drank, he sat, he saw...). There is an eventfulness here that establishes the agency of the primary entity. While the actions are partially sequential in nature (drinking beer, looking out the window), one can see how there is also a strong simultaneity to them as well. The experiential ordering of drinking and looking is unclear. Rather, the event of beer drinking exists in its own timeframe, like the shutting of one's eyes as well as looking out the window. Each action rests in its own temporal plane, intersecting but not in a strictly linear way. What arguably matters more to the passage's narrativity is the eventfulness rather than the sequentiality.

The actions also do not stand alone but are highly concretized through an array of physical objects that populate this scene (beer, airplane, window, seat-section, wing, clouds, and hundred-dollar bills). The world of the airplane only assumes meaning through the accumulation of sensory experiences. Finally, the experience of distantiation and deixis that we are arguing is so central to narrativity is doubly in force here: not only do we see it through the accumulation of past tense verbs but through the variety of thematic separations that the passage enacts. No one sits next to him. He shuts his eyes but doesn't sleep. And of course, there is the classic scene of looking out a window at an inaccessible view that is expressly described as incommensurable with the monetary objects in his pocket. The meaning of this scene is constructed through the boundaries it erects between us and it, between Ray, the main character, and the world. The narrative world is a microcosm of the distantiation that narrative itself performs.

As a concluding note, we simply want to underscore that one of the most important aspects of this exercise has been the process itself. We fully expect our models to be refined, updated, and maybe even one day refuted. But the key aspect is that we are using data to test the robustness of our models to explain the behavior of large numbers of documents across numerous social situations rather than relying on the views of a single expert opinion. Modeling can help us gain confidence about the generalizability of our theories about the world as well as develop more parsimonious ways of describing that world. Most importantly, it can help us move in new directions as we work to validate and understand more deeply the insights that data-driven modeling offers. The distant worlds theory raises interesting questions as to why human beings developed a capacity to point at distant agents / distantiated agency through language. Our provocation is that "theory of mind" explanations, i.e. that we tell stories to heighten human empathy, are not fully satisfactory in accounting for the distinct nature of narrative communication that we are showing here. Nor do we find evolutionary accounts that suggest "social learning" as the basis of storytelling to adequately describe what our models are showing us.[34] The question remains, why narrative? What is the value of socializing and ritualizing the communication of distantiated agency? We think there is an opportunity here to reflect further on the cognitive and social value of the deictic theory of narrativity that our models deictically point us toward.

---

[1] For a useful account of the history of the field, see Genevieve Liveley, Narratology (Oxford: Oxford Univ. Press, 2019).

[2] Gérard Genette, “Boundaries of Narrative,” New Literary History 8.1 (1976), 1–13.

[3] All data, code, and supplementary material related to this article is located in the following repository: https://doi.org/10.6084/m9.figshare.21656780.

[4] Ted Underwood, Distant Horizons (Chicago: Univ. of Chicago Press, 2019).

[5] Andrew Piper and Eva Portelance, "How Cultural Capital Works: Prizewinning Novels, Bestsellers, and the Time of Reading," Post-45 (2016). https://post45.org/2016/05/how-cultural-capital-works-prizewinning-novels-bestsellers-and-the-time-of-reading/

[6] Ted Underwood, "The Life Cycles of Genres," Journal of Cultural Analytics 2 (2). https://culturalanalytics.org/article/11061-the-life-cycles-of-genres

[7] Chris Beausang, "Diachronic Delta: A Computational Method for Analysing Periods of Accelerated Change in Literary Datasets," Digital Scholarship in the Humanities, 2021.

[8] Arthur M. Jacobs, "Neurocognitive Poetics: Methods and Models for Investigating the Neuronal and Cognitive-Affective Bases of Literature Reception," Frontiers in Human Neuroscience 9 (2015): 186.

[9] David Herman, Basic Elements of Narrative (Chichester, U.K, Wiley-Blackwell, 2009).

[10] Rachel Giora and Yeshayahu Shen, "Degrees of Narrativity and Strategies of Semantic Reduction," Poetics 22, no. 6 (1994): 447–458; Federico Pianzola, "Looking at Narrative as a Complex System: The Proteus Principle," in Narrating Complexity, ed. Richard Walsh and Susan Stepney (Springer, 2018), 101–22.

[11] Andrew Piper and Sunyam Bagga, "Detecting Narrativity Across Long Time Scales," CHR 2021: Proceedings of the Computational Humanities Research Conference (2021): 319-332.

[12] Elinor Ochs and Lisa Capps, Living Narrative: Creating Lives in Everyday Storytelling (Cambridge, MA: Harvard Univ. Press, 2009), 19.

[13] Michael Gavin, “Is There a Text in My Data? (Part 1): On Counting Words,” Journal of Cultural Analytics 5 (1). https://culturalanalytics.org/article/11830-is-there-a-text-in-my-data-part-1-on-counting-words

[14] This process of iterative model building was first described in Andrew Piper, "Novel Devotions: Conversional Reading, Computational Modeling, and the Modern Novel," New Literary History 46.1 (2016): 63-98.

[15] Gérard Genette, Narrative Discourse: An Essay in Method (Ithaca: Cornell Univ. Press, 1983).

[16] "Experientiality reflects a cognitive schema of embodiedness that relates to human existence and human concerns. [...] In my model there can be narratives without plot, but there cannot be any narratives without a human (anthropomorphic) experiencer of some sort.” Monika Fludernik, Towards a 'Natural' Narratology (London: Routledge, 1996), 9.

[17] Herman, Basic Elements of Narrative, 14.

[18] Peter Hühn, "Event and Eventfulness," in Handbook of Narratology, edited by Peter Hühn et al., Vol. 1 (De Gruyter, 2014): 159-178.

[19] Meir Sternberg, "Telling in time (I): Chronology and narrative theory," Poetics Today 11, no. 4 (1990): 901-948. Brewer, William F., and Edward H. Lichtenstein. "Stories are to Entertain: A Structural-Affect Theory of Stories," Journal of Pragmatics 6, no. 5-6 (1982): 473-486. Paul Ricoeur, "Narrative Time," Critical Inquiry 7, no. 1 (1980): 169-190.

[20] Marie-Laure Ryan, "Possible Worlds," Handbook of Narratology, edited by Peter Hühn et al., Vol. 2 (De Gruyter, 2014): 726-742.

[21] Andrew Piper, "Novel Devotions”; Jessica Ouyang and Kathleen McKeown, "Modeling Reportable Events as Turning Points in Narrative,” Proceedings of the 2015

Conference on Empirical Methods in Natural Language Processing (2015): 2149-2158; Benjamin M. Schmidt, "Plot Arceology: A Vector-Space Model of Narrative Structure," in 2015 IEEE International Conference on Big Data (IEEE, 2015): 1667-1672; Andrew J. Reagan, Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter Sheridan Dodds, "The Emotional Arcs of Stories are Dominated by Six Basic Shapes," EPJ Data Science 5, no. 1 (2016): 1-12.

[22] Thomas A. O'Neill, "An Overview of Interrater Agreement on Likert Scales for Researchers and Practitioners," Frontiers in Psychology 8 (2017): 777.

[23] We also note observable variation between genres, both in terms of between-group averages as well as within group variation. For example, ROC and Fables exhibit average deviations of 0.08 and 0.17 respectively, while contemporary novels and flash fiction deviation averages are both over 0.60. This suggests that some genres elicited consistently more disagreement than others. At the same time, we see less within-group variation in the non-narrative classes. This means that coder disagreement was more consistent within non-narrative passages, suggesting that non-narrativity is easier to agree on than narrativity.

[24] As Herman writes, "A sequence can be processed as a narrative not just because it has a certain form but also because its form cues readers, in structured, nonrandom ways, to interpret the sequence as a narrative." David Herman, "Scripts, Sequences, and Stories: Elements of a Postclassical Narratology," PMLA 112, no. 5 (1997): 1050.

[25] Andrew Piper, "Sense (Fictionality)," Enumerations: Data and Literary Study (Chicago: Univ. of Chicago Press, 2018), 94-117. Piper and Portelance, "How Cultural Capital Works."

[26] Fludernik, Towards a "Natural" Narratology, 9.

[27] Hühn, "Event and Eventfulness."

[28] Recent work on tense clusters may provide some indication why explicit temporal markers are less predictive of narrative communication. In other words, sequentiality may be less informative than tense markers themselves for reasons we discuss later in the paper. See Thomas Bögel, Jannik Strötgen, and Michael Gertz, "Computational Narratology: Extracting Tense Clusters from Narrative Texts," Proceedings of the Ninth International Conference on Language Resources and Evaluation (2014): 950-955.

[29] Andrew Piper and Sunyam Bagga, "A Quantitative Study of Fictional Things," CHR 2022: Proceedings of the Computational Humanities Research Conference (2022): 268-279; Dennis Yi Tenen, "Toward a computational archaeology of fictional space," New Literary History 49, no. 1 (2018): 119-147.

[30] Randall D. Beer, "Dynamical Approaches to Cognitive Science," Trends in Cognitive Sciences 4, no. 3 (2000): 97.

[31] Lisa Zunshine, Why We Read Fiction: Theory of Mind and the Novel (Columbus: Ohio State Univ. Press, 2006).

[32] Michael Tomasello, Origins of Human Communication (Cambridge, MA: MIT Press, 2010).

[33] Dagoberto Gilb, "Winners on the Pass Line," in The Scribner Anthology of Contemporary Short Fiction: Fifty North American Stories Since 1970, ed. Lex Williford and Michael Martone (New York: Scribner Paperback Fiction, 2007), 239-252.

[34] Brian Boyd, On the Origin of Stories : Evolution, Cognition, and Fiction (Cambridge, MA: Harvard Univ. Press, 2009).