Revision of XGE-2021-3239R1 as invited by the action editor, Julie Van Dyke, PhD

**Semantic Progression and Narrative Success**

**Abstract**

Why are some narratives more successful? While this question has ancient roots, studying it empirically has been challenging. We suggest that semantic progression (i.e., semantic similarity between adjoining portions of a narrative) may shape audience responses, but that this role changes over a narrative’s course. Specifically, while slower semantic progression (i.e., greater semantic similarity between adjoining portions) may be beneficial at the beginning of narratives, faster semantic progression may be beneficial towards the end. To test this possibility, we use natural language processing and machine learning to analyze over 40,000 movie scripts. Consistent with our theorizing, Deep-Learning-based embeddings find that movies with slower semantic progression early, and faster semantic progression later, are evaluated more positively. Analysis of over 10,000 TV episodes finds similar results. Overall, these findings shed light on what makes narratives engaging, deepen understanding of what drives cultural success, and underscore the value of emerging computational approaches to understand human behavior.

Narratives are an integral part of everyday life. Early humans shared epic poems and stories of the hunt, and today we read books, watch movies, and consume content online.

But why are some narratives more successful than others?

Everyone from psychologists and philosophers to linguists and narrative theorists have long been interested in quantifying properties of narratives (e.g., Cutting, 2016; Pennebaker, 2018; Propp, 1968; Carroll, 1990). In *Poetics*, Aristotle proposed that stories tend to have common structures and Freytag (1900) later suggested a “dramatic arc” or pyramid of rising and falling action. Campbell (1949) theorized that there was an archetypical hero’s journey and Kurt Vonnegut’s rejected master’s thesis argues that stories could be divided into eight common shapes. Recent work has begun to address this topic empirically, examining variation in sentiment (Reagan, Mitchell, & Kiley, 2016) and categories of language (Boyd, Blackburn, & Pennebaker 2020).

But while research is starting to identify patterns of language across narratives, there has been less attention to whether, and how, particular narrative features might shape their success. Popular perspectives argue that certain ways of writing can make narratives more successful (e.g., McKee, 1997; Coyne, 2015; Snyder, 2005), for example, but the little empirical work that has been done found no relationship between story structures and popularity (Boyd et al., 2020).

Building on research on semantic similarity, however, we suggest that the pace or speed of a narrative’s semantic progression may shape audience response. Textbooks, books, and other texts can be described by their cohesion (Graesser, Singer, and Tranbasso 1994; Graesser, McNamara, Louwerse, and Cai 2004), or how semantically related chunks (e.g., adjoining paragraphs) are to one another (Foltz, 2007). Compared to a paragraph about pine trees, for example, a paragraph about another type of tree should be more semantically similar, or related, than a paragraph about a bank robbery. Along these lines, research finds that adjoining paragraphs of textbooks are more semantically related than paragraphs that are further away and adjoining paragraphs within a chapter are more similar than those on either side of a chapter break (Foltz et al., 1998; see Foltz 2007 for a review). Further, averaging across all adjoining chunks of a text provides insight into the how easy or difficult it should be for a reader to consume that text. More difficult textbooks, for example, tend to involve larger semantic jumps (i.e., lower similarity or greater distance) between adjoining chunks of text (Foltz et al., 1998).

While semantic relatedness is often applied in the context of education and learning (e.g., how coherence impacts comprehension), we suggest that the same ideas may be helpful in understanding the evaluation of narratives. Objects that cover a greater distance in the same amount of time can be described as moving faster. Using the notion of semantic relatedness, the same can be said for narratives. Some narratives move more slowly, dwelling on semantically related concepts for longer periods, while others move more quickly, jumping between content that is less semantically related.

Consequently, semantic progression can be connected to the notion of pacing, or the speed at which a narrative is told (Hume 2005; Turco, 1999). While speed may also conjure up the amount of action, or how quickly things cut between scenes, here we focus specifically on speed at it relates to semantic progression, or the speed at which the content of the narratives unfolds. The speed of semantic progression can be defined as how quickly the content of discourse (e.g., a narrative) moves between adjoining chunks. By comparing equally long chunks of text across narratives (e.g., 250-word blocks), narratives whose adjoining chunks are more semantically related can be described as moving more slowly, while narratives whose adjoining chunks are less semantically related can be described as moving faster.

But which should be liked more: slower semantic progression or faster?

For ideas to make sense, they have to be at least somewhat coherent. Slower semantic progression should require less cognitive work to follow (Monahan, Murphy and Zajonc, 2000) and make it easier for people to track what is happening (McNamara Kintsch, Songer and Kintsch, 1996). That said, faster progression juxtaposes different concepts which could increase surprise and stimulation and make narratives more engaging (Gergen and Gergen, 1986). So which is better?

We suggest that the answer may depend on the part of the narrative that is being considered. While one could argue that faster semantic progression itself is either good or bad (i.e., people like faster or slower paced stories overall), we suggest that the speed of semantic progression *within* a narrative should also impact responses.

At the beginning of a book, movie, or any other narrative, the canvas is blank. The audience doesn’t know anything about the characters or context. Consequently, the beginning must set the stage (Cutting 2016). It must outline these details, and do so in a way that builds a base, or jumping off point, for the rest of the narrative (MacEwan, 1900). Indeed, descriptions of people, places, and things peak at the beginning of stories (McClure and Enderle, 2018), as do prepositions and articles (Boyd et al., 2020), which helps the audience understand what is going on (Morrow, 1990). A story might start by talking about “the house at the end of the road,” for example, but once the reader becomes familiar with the context, simply refer to the house as “it.”

Consequently, we suggest that early on, slower semantic progression (i.e., greater similarity between adjoining narrative chunks) may boost success. Like teammates in a relay race, if the second runner starts slowly enough, the first runner can still catch them and hand off the baton. But if that second runner starts at full speed, the first person will never be able to catch up. The same may be true in narratives. If they more too fast, too early on, the audience may get lost.

A variety of literatures support this prediction. Children’s stories, jokes, and even music often begin by repeating a similar concept (Loewenstein and Heath, 2009; Rozin, Rozin, Appel, and Wachtel, 2006). In the Three Little Pigs, for example, one pig builds a straw house, the wolf blows it down, and something very similar happens to a second pig (i.e., he builds a house of sticks and the wolf blows it down). In jokes, a certain thing happens when a priest walks into a bar, and a similar thing then happens to a nun. This semantic similarity grounds the audience and helps build expectations. Similarly, developmental psychology research suggests that when trying to learn something, children often prefer seeing the same stimulus again and again so they can deepen their understanding. Exposure to the same, or similar content, is easier to process and requires less cognitive effort (Monahan, Murphy and Zajonc, 2000; Zajonc 2001), which may be particularly important at the beginning of a narrative when an audience knows little about the world being created and the characters in it.

But while slower may boost success early on, this may shift as narratives advance. Once the audience has met the characters and understands the context, the plot must progress (MacEwan, 1900). Relationships must develop, things must happen, and challenges must be overcome. Indeed, measures of plot progression (e.g., pronouns and auxiliary verbs) start small but pick up later in narratives (Boyd et al., 2020).

Consequently, we suggest that towards the end of the narrative, faster pacing (i.e., lower semantic similarity between adjoining chunks) may be beneficial. Consistent with this notion, while jokes and folk tales start with repetition to form expectations, these expectations are then broken by the final contrasting event (e.g., but when the wolf tied to blow down the house of bricks…), facilitating surprise and engagement (Lowenstein and Heath, 2009; Lowenstein, Raghunathan, and Heath, 2011; Rozin et al., 2006).[[1]](#footnote-2) Treading the same ground again and again can get boring and once people understand something, they often want to move on to something else (Flavell, Miller, and Miller, 2001). Similarly, work on comprehension and learning finds that textual coherence helps novices, but lower coherence may actually be better for more knowledgeable readers because they already know enough to fill in the blanks (McNamara Kintsch, Songer and Kintsch, 1996).

In sum, we suggest that speed of semantic progression may have different effects on narrative evaluations depending on the point in the narrative. Similar things should be easier to process and require less cognitive effort, and such ease of processing can lead to increased liking and evaluations (Alter and Oppenheimer 2009). Consequently, early on, when characters, setting and everything else are still novel, narratives should be evaluated more positively when the speed of semantic progression is slower (i.e., greater semantic similarity between adjoining chunks). Slower speeds should help the audience understand what is going on, which should increase engagement and evaluations.

Towards the end of the narrative, however, the opposite should occur. Once the audience is already familiar with the characters and setting, they have the cognitive structures and schemas in place that should make it easier to incorporate novel information. More novel things should be more stimulating (Flavell, et al. 2001) and thus faster semantic progression (i.e., lower semantic similarity between adjoining chunks) may be evaluated more positively.

Said another way, there is a tension between novelty and familiarity. Being exposed to the same thing over and over again gets boring, and novelty provides stimulation which can increase evaluations. At the same time though, if something is too novel, it may be difficult to understand. Consequently, while a blend of novelty and familiarity is often beneficial, which aspect is valued more should depend on position in the narrative, At the beginning of a narrative everything is new, so there is less need to provide additional stimulation, and greater semantic similarity should facilitate understanding. At the end of a narrative, however, the characters and setting are already laid out, and so greater semantic leaps may provide beneficial stimulation.

To test whether the speed of semantic progression has different effects at different parts of a narrative, we use Natural Language Processing. We embed the dialogue of more than 40,000 movies into vectors that characterize semantic content. Then, we analyze the relationship between adjoining chunks of narrative and how it relates to cultural success. In addition to providing multiple controls and robustness checks, to test the generalizability of the effects, we also examine whether they hold among 10,000 TV episodes.

Note that we focus on semantic progression rather than specific details of the content. Two narratives might have completely different characters, setting, and plot, for example, but their semantic progression might be quite similar (i.e., slow pacing initially and faster at the end). We test whether such speed of semantic progression might impact audience response.

**Empirical Analysis of Over 40,000 Movies**

**Data**

To collect data on the semantic content of narratives, we used the English OpenSubtitles2018 corpus of OPUS (<https://opus.nlpl.eu/OpenSubtitles-v2018.php> Tiedemann & Lison, 2016). It contains the dialogue, or words actors say throughout a movie’s script. The text files were cleaned to remove irrelevant information (e.g., metadata on what software was used to encode subtitles or opening credits).

To capture audience responses, we collected movie ratings from IMDb.com. Each movie is reviewed by multiple people on a 1-10 scale and IMDb reports the mean score. We focused on popular opinion, rather than critics ratings, because we are interested in general response. The semantic progression of the movie should influence audiences evaluations of it, and online ratings provide a measure of such evaluations

 To control for other movie features which might impact success, we collected metadata like runtime, production year, and movie genre (e.g., action or comedy, see Supplemental Materials Table S1 for prevalence of different genres). To focus on movies, analysis focused on content that was at least 30 minutes long and had at least 2500 words. This resulted in a dataset of 42,472 unique movies.

**Computing Speed of Semantic Progression**

First, we split each movie into chunks, where each chunk (except the last one) has the same number of words. There is no obvious right answer for how long chunks should be, so we relied on prior work as well as the nature of the data. Most movies range from between 3,000 and 20,000 words, and given the need to break them up into at least five parts (Freytag, 1900), and have at least three chunks per part (to compute average semantic progression), we used 250-word blocks for our main analysis. Results are the same, however, for larger and smaller blocks (i.e., 150 and 350 words, see robustness section).

Second, we determine the semantic similarity of adjoining chunks. Early work in this area used latent semantic analysis (see Foltz, 2007 for a review), but recent advances in computer science provide even more advanced approaches. Unsupervised embedders take words, sentences, or even whole documents, and represent them as vectors in a high-dimensional space. Word embeddings, for example, are based on Firth’s (1957) suggestion that “you shall know a word by the company it keeps.” If two words are often surrounded by similar words, they probably have very similar meanings. Embedding algorithms “organize” words in a multi-dimensional space, where each word receives a numerical vector, and words with similar meanings or uses are closer together (Bhatia, 2017). Extensions of these algorithms allow larger chunks of text (i.e., paragraphs or whole documents) to be transformed into numerical vectors with the same interpretation.

Most sentence or paragraph-level embedders use a common approach to word-level embeddings but differ in how they combine them to get a vector for the whole chunk. Given its performance in human scoring benchmarks, we use Google’s Universal Sentence Encoder (Cer et al., 2018) but the results are robust to other approaches (i.e., Doc2Vec, see robustness section). The Universal Sentence Encoder offers a few different Deep-Learning models pre-trained on a variety of large corpi (e.g., English-language Wikipedia and online reviews), and for its simplicity, speed, and robust results, we chose Google’s pre-trained Deep Averaging Network, or DAN (Iyyer, Manjunatha, Boyd-Graber, & Daumé III, 2015). DAN first embeds individual words in a text (as word2vec), then takes their weighted average for classification and loss function computation (See Supplemental Materials for more detail).

Following prior work (e.g., Foltz et al., 1998) we measure the semantic similarity of adjoining blocks of text using cosine similarity (also see Bhatia, 2017; DeFranza, Mishra, and Mishra, 2020). This provides marginal advantages over Euclidean distance with added resilience against a few degenerate cases sometimes seen in text data (though both approaches lead to similar results). It is defined as

$$cosθ=\frac{\vec{a}∙\vec{b}}{\left‖\vec{a}\right‖\left‖\vec{b}\right‖}$$

This yields similarity values between -1 and 1, where two identical vectors would have similarity 1 and larger cosine similarity values indicate greater semantic similarity (as they do in LSA).

Prior work ([Bhatia,](#page10) 2017; [Garg et al.,](#page10) 2018; [Kozlowski et al., 2019)](#page10), as well as a series of validation tests (see below) demonstrate that this measure captures semantic similarity as well as human perceptions of similarity. Within the same movie, for example, adjoining blocks are, on average, scored as more semantically similar than nearby blocks, which are scored as more similar than blocks that are further away, which are scored as more similar than blocks from a completely different movie.

Third, by averaging across pairs of adjoining chunks, we calculate the average speed of semantic progression for different parts of each movie. A great deal of prior work suggests that movies follow a five-act structure (Freytag, 1900), so the main analysis break movies up into five parts, but results are robust to larger or smaller numbers as well (see robustness). For each part (i.e., one-fifth of the movie), we calculate the average similarity between adjacent chunks (i.e., chunk 1 and chunk 2, chunk 2 and chunk 3, and so on) within that part. Taking the opposite of this captures speed of semantic progression. If two chunks are more semantically dissimilar (i.e., lower cosine similarity between them), it means that greater semantic distance was covered in the same amount of time, and thus indicates a faster speed of semantic progression.

Finally, linear regression predicted average movie rating based on the average speed of semantic progression of each part. See Table S2 for a correlation matrix of the variables.

**Results**

Results indicate that while movies with faster semantic progression early on (i.e., in the first part) are evaluated less positively, movies with faster semantic progression at the end (i.e., last part) are evaluated *more* positively (Table 1, Model 1). Speed of semantic progression initially has a negative relationship with evaluations (i.e., greater speed or more distance between adjoining chunks is detrimental), but this relationship increases through the rest of the movie, eventually becoming significantly positive by the end (i.e., greater speed or more distance between adjoining chunks is linked to higher ratings, Figure 1).

*Table 1.* Speed of Semantic Progression and Audience Response

While one could wonder whether these effects are somehow driven by the modeling approach used, estimates for each movie part (i.e., fifth of the narrative) are independent, so no functional form is being forced on the results. Instead, the fact that estimates are negative initially, but reduce and become positive at the end is consistent with the notion that slower pacing boosts success at the beginning of movies while faster pacing helps towards the end.

*Figure 1***:** Audience Response Based on Speed of Semantic Progression Across Movie Parts

Note: Results based on raw coefficients and standard errors from Table 1, model 1

**Controls.**These initial results are consistent with our theorizing, but one could wonder whether they are driven by other factors. Consequently, we include various control variables (i.e., year of release, runtime, genre, budget, and number of ratings) to test alternative explanations and robustness.

First, one might wonder whether release year is somehow driving the results. Maybe more recent movies are both more highly rated, for example, and use certain plot structures. While it is unclear how such an explanation would explain variation in semantic progression *within* movies, to control for this possibility, we control for *Year of Release*. Using continuous form or dummy variables yields similar results, so for simplicity the main results include a continuous version.

Second, one could wonder whether longer movies have faster or slower semantic progression towards the end and might also receive differential ratings. Consequently, we control for *runtime* in minutes.

Third, one could wonder whether certain genres of movies might tend to receive higher ratings and also have different patterns of semantic progression. Consequently, we use dummy variables to control for IMDb’s *genre* tags (e.g. Action, Comedy, or Horror). Some movies were tagged with multiple genres, and thus count in multiple groups.

Fourth, one could wonder whether blockbuster type movies tend to use certain plot structures and receive higher or lower ratings. We control for this possibility two ways: (1) *Budget* and (2) *Number of Ratings* of each movie, controlling for the log of each because it gives a more normal distribution of values.

Even controlling for all these factors, however, the relationships between speed of semantic progression and success persist (Table 1, model 2). Movies with faster semantic progression early on are evaluated less positively, but movies with faster semantic progression towards the end are evaluated *more* positively

**Robustness and Alternative Explanations**

We also test robustness in a number of other ways, including (1) the blocking approach, (2) number of parts, (3) embedding approach, (4) modeling approach, (5) blocks per part, (6) mean similarity across parts, (7) removing stop words, (8) comparing only movies of similar lengths, and (9) production side factors. Across all these different specifications, the results still persist.

First, one could wonder whether the results are somehow driven by the blocking approach used. The main model used 250-word blocks, but to test this possibility, we also examine 150 and 350-word blocks. Results remain the same (Table 2). Movies with faster semantic progression early on are evaluated less positively, but movies with faster semantic progression towards the end are evaluated *more* positively

*Table 2.* Robustness Tests

 

Second, one could wonder whether the results are somehow driven by the number of parts or acts used. Following prior work (Freytag, 1900), the main model used a five-act structure, but some work has suggested seven-act structures and, though less common even four or six act structures. To test the robustness of the effect, we also test a four, six, and seven part structure (i.e., analyzing the speed of semantic progression within quarter of each movie, sixths, or sevenths). Results remain the same (see Table S3).

Third, one could wonder whether the results are somehow driven by the embedding approach used. Our main analyses relied on USE-DAN because it well captures semantic similarity while being simpler, faster, and more robust than alternative methods, but to test robustness, we use Doc2Vec (Quoc & Mikolov, 2014). We used a pre-trained Doc2Vec model (Lau & Baldwin, 2016) to calculate the position of each movie chunk and re-ran our main analyses. Results remain the same (Table 2).

Fourth, one could wonder whether the results are somehow driven by treating each part (i.e., one-fifth of the movie) as independent. Maybe audiences value high similarity at the beginning of a narrative *or* low similarity at the end, for example, but don’t care whether both are present. There might also be other interactions that are stronger than the effects of individual acts or parts. To test this possibility, we cluster movies based on their shapes and examine whether movies with certain shapes are liked more (see Supplemental Materials). Results remain the same.

Along those lines, one could wonder whether it would make more sense to look at change in narrative speed over the entire movie. To examine this alternate measure, we take the slope of the line between act number and average speed in the act for each movie. A positive slope indicates that the speed of semantic progression is generally increasing across the movie. Consistent with the notion that people like slower speed early and faster speed late, results indicate that movies are liked more when the speed of semantic progression increases more over the course of the movie (coefficient = 1.15, SE = 0.29, p < .001).

Fifth, one could wonder whether the method of dividing up the text could be driving the results. The results are robust to different blocking sizes, but the number of blocks or chunk per part can also vary. While a 10,000-word movie would be broken up into five parts (i.e., acts) of 2,000 words, each of which composed of eight 250-word chunks, other lengths are slightly more uneven. A 11,000-word movie, for example, has 44 250-word chunks, which means that while the first four parts would each have nine chunks, the final part would have only eight. To test whether slightly differing numbers of chunks for later parts or acts in some cases could somehow drive the results, we removed the last chunk of parts with higher counts (so all parts have an equal number) and reran the model. Results remain the same (Table 2).

Sixth, one could wonder whether the mean level of semantic progression across parts could be driving the results. Maybe audiences always like the same speed of semantic progression throughout a narrative, but actual speed of semantic progression used tends to be higher at the beginning and lower at the end, leading to the observed relationships (i.e., lower than average is preferred at the beginning and higher than average is preferred at the end). But this is not the case. There is no difference between the mean speed of semantic progression at the beginning and end ($t$(43078) = -0.93, *p* = 0.82). Further the mean level of speed of semantic progression is similar (i.e., within one standard deviation) across the movie. Consequently, variation in actual speed of semantic progression across the movie has trouble explaining the results.

 Seventh, one could wonder whether the results could be driven by different types of language emerging across a story. Narratives tend to move from low to high rates of pronouns (Boyd, et al. 2021), for example, and one could wonder whether the resulting changes in weighting of background function words is driving the effect. To test this possibility, we re-run the main analysis, but remove stop words (e.g., I, me, an, and the) from the text. While the size of the exact coefficients changes slightly, the results remain the same. Movies are liked more when they move slowly at the beginning and faster at the end. This casts doubt on the possibility that language shifts are driving the effect, but future work might examine the consequences of shifts in language across narratives more generally.

Eighth, one could wonder whether the results were somehow driven by comparing movies of different lengths. This is not the case. Even looking at movies of similar lengths (i.e. between 25 and 35 chunks, which is around the mode of the distribution) results remain the same.

Ninth, one could wonder whether production side factors could somehow be driving the effect. We already controlled for production side measures like budget, but maybe certain screenwriters tend to make “better” movies, for example, and also tend to follow certain pacing structures and that is driving things.

But this does not seem to be the case. We identified famous screenwriters that have written a number of films (i.e., George Lucas, Quentin Tarantino, and Ingmar Bergman), found all the movies in the dataset that they had written, and tested whether movies by the same screenwriter tended to have a consistent pattern of semantic progression. They did not. Some of George Lucas’ movies, for example, increased in semantic progression across the course of the narrative, while others decreased. Even Quentin Tarantino, a writer known for his distinctive style, has considerable variety in his semantic progression (see Figure S1).

This variation also casts doubt on the possibility that these findings are already known in the industry. If faster pacing being helpful late but detrimental early was already known, one would expect famous screenwriters to use this approach. But they do not. Further, one could argue that even if screenwriters don’t know it, studios do, and so the invest more money in films that follow that structure. But as shown in Figure S1, not all famous movies follow this structure, casting doubt on that possibility as well.

**Validation**

While the results are consistent with our theorizing, one could wonder whether the measure of semantic progression is truly capturing the similarity between content. A great deal of prior work documents that embeddings distance captures semantic similarity (e.g., [Bhatia,](#page10) 2017; [Garg et al.,](#page10) 2018; [Kozlowski et al., 2019)](#page10), but to underscore this point, we provide a few additional validation tests in our data (see Supplemental Materials for additional comparison with the STS benchmark).

**Semantic Similarity between Different Chunks**. If our measure is truly picking up semantic similarity, as we suggest, then following Foltz et al. (1998), adjoining chunks of a movie should be scored as more semantically similar than chunks that are further away. Similarly, chunks of one movie should be scored as more semantically similar to chunks of that same movie than chunks of a completely diﬀerent movie.

To test these possibilities, we draw 9,000 target chunks randomly (without replacement) across all movies. For each target chunk, we randomly select four comparison chunks: (1) an adjacent chunk, (2) a chunk 5 positions away from the initial chunk, (3) a chunk 5 to 10 positions away, and (4) a chunk from another movie. Then, we compute the distance between the target chunk and each of these four types of chunks and average them for each type across all 9,000 target chunks.

Results support the notion that our measure captures semantic similarity. As predicted, adjacent chunks from the same movie are more semantically similar, on average, than chunks from the same movie within 5 positions, which are more semantically similar than chunks selected anywhere from the same movie, which are more semantically similar than chunks from diﬀerent movies (M**Adjacent** = 0.52 < M**5Away** = 0.47 < M**5-10Away** = 0.44 < M**Diff Movie** = 0.35; *t*s *> 22*, *p*s < .0001).

**Relation to Human Perceptions.** To provide further evidence that this measure captures semantic similarity, we also tested whether it is related to perceived similarity. We randomly picked 50 movies, randomly picked a target chunk from each, and then picked two comparison chunks. To do so, we computed the distance between the target chunk and all other chunks in the same movie, and randomly picked a chunk from the bottom quartile of similarity and a chunk from the top. Then, for each target chunk, three hypothesis-blind research assistants coded which comparison chunk was more similar. Majority rule determined which chunk humans judged as closer.

Results indicate that our measure of semantic similarity was reasonably related with human perceptions of similarity. Human judgment agreed with word embeddings 74% of the time. As a baseline, the joint probability of agreement between judges (i.e., the probability that evaluations from two judges on the same triplet is the same) was 72%. This indicates that our semantic similarity measure is reasonably consistent with human similarity perception and similarly reliable.

**Ancillary Test: Empirical Analysis of Over 10,000 TV Episodes**

Results of the main study are consistent with our theorizing, but one could wonder whether the findings are replicable. Even though we included multiple controls, and robustness checks, maybe there is something unique to movies that is driving the effect. Consequently, to further test robustness, and generalizability, we conduct an ancillary test in another domain. Specifically, we collect over ten thousand episodes from over 300 TV shows. We test whether TV shows are liked more when the speed of semantic progression is slower at the beginning but faster at the end.

 A major global media company provided a dataset of closed captions for English language TV shows. This included everything from multi-season thrillers to episodic children’s cartoons. To capture audience response, as in the main study, we used IMDb.com to collect ratings for each episode of each TV show. To control for other features which might impact success, we collected metadata at both the show level, such as the channel it aired on and genre (e.g., Comedy and Crime), and episode level, such as the season of the episode and episode number within the season. Episode-level information allows us to control for the possibility that later episodes of a show are liked more (or less). This resulted in a dataset of 10,578 unique episodes from 290 TV shows.

As noted in the main study, computing average semantic progression requires at least three chunks per part of the narrative. TV shows are much shorter than movies (M = 4,321 words vs. 8,827), so breaking episodes into five parts and using 250-word chunks of would mean that many episodes would have fewer than three chunks a part. We address this in three ways. First, the main analysis uses 150-word chunks and three parts. This ensures that most narratives have a similar number of chunks per part as movies. Second, to ensure that the results aren’t somehow driven by the chunk size, we also examine 250-word chunks. Third, to ensure that the results aren’t somehow driven by the number of parts, we also examine what happens when more parts are used.

Across all three versions, results are similar to those found for movies (Table 3). While episodes with faster semantic progression early on are evaluated *less* positively, episodes with faster semantic progression at the end are evaluated *more* positively. In the main analysis, for example, speed of semantic progression initially has a negative relationship with evaluations, but this relationship increases through the rest of the show, eventually becoming significantly positive by the end (Model 1). Results are the same using larger chunks (Model 2) or dividing shows into more parts (Model 3)

Table 3: Speed of Semantic Progression in TV Episodes

 

 Analysis of over 10,000 TV episodes underscore the findings of the main study. While TV episodes whose early parts are more semantically similar were evaluated more positively, shows whose later parts are more semantically similar were evaluated less positively. Finding the same results across a different domain speaks to the generalizability of the effect.

**Discussion**

Academics and practitioners alike have long been interested in what makes narratives engaging. From Aristotle’s early ideas about the trajectories of tragedies and comedies, to more recent theories of ways to make screenplays, scripts and stories more popular (e.g., Mckee, 1997), different perspectives have made different propositions. But while these ideas have captured the imaginations of everyone from literary theorists to so called “script-doctors,” actual empirical tests have been few and far between. Consequently, less is known about what makes some narratives more successful.

Machine Learning based Natural Language Processing of over 40,000 movies sheds light on this question. While movies whose early parts are more semantically similar are evaluated more positively, movies whose later parts are more semantically similar are evaluated *less* positively. Ancillary analysis of over 10,000 TV episodes finds similar results. Taken together, the findings suggest that while slower semantic progression may be beneficial at the beginning of narratives, faster semantic progression is beneficial towards the end.

While we focused on movies and tv show due to data availability, similar approaches could be applied to any type of narrative (e.g., plays or books). One important moderator, however, may be the way narratives are consumed. While movies and plays tend to be consumed in one sitting, for example, books tend to be read over multiple occasions. Consequently, the impact of pacing may vary. In books, pacing *within* chapters may be important, as starting slowly may helping readers remember what happened in the last chapter. Research could also examine how the speed of semantic progression within a narrative shapes comprehension and memory.

Future work might also more deeply examine the psychological mechanisms behind these effects. As noted, comprehension, processing ease, and stimulation may all play a role. At the beginning of a narrative, much is already quite novel, and so greater semantic similarity (i.e., slower semantic progression) may help the audience understand what is going on and make the content easier to process, bother of which should increase evaluations. At the same time, at the end of the narrative, the characters and settings are no longer new, so providing greater surprise and stimulation through faster semantic progression may boost evaluations.

Work might also examine how speed of semantic progression shapes other relevant outcomes, like comprehension and memory. Does faster speed of semantic progression, for example, decrease understanding and memory for what occurred? This may be particularly relevant for educational texts.

One could also examine whether the effects extend to other types of discourse. If these effects are specific to narratives, things like product manuals or how-to guides should show different patterns. On the other hand, if slower pacing early, and faster pacing later, make any type of content easier to follow, such semantic progression may be beneficial in a broader set of content. Even “non-narratives” like product manuals often follow a building structure where later points may build on earlier ones, and thus semantic progression may play a role.

Future work might also examine other narrative features. People often talk about narratives as covering a lot of ground, for example, or going in circles. Research could try to quantify these aspects, looking at ground covered (i.e., the space covered by all the chunks of the narrative) or the degree to which it repeatedly returns to common themes and ideas. Two movies might both be fast paced, for example, but one might go back and forth between the same ideas while the other moves on to new topics.

Similar ideas could also be applied to personal narratives. People often use narratives to explain and understand their own lives (McAdams & McLean, 2013). Just as creative people have more distance (i.e., less semantic relatedness) between their thoughts (Gray et al., 2019), faster semantic progression in personal narratives may provide insight into the writer’s personality, or even how the act of writing may impact wellbeing (Pennebaker, 2018).

Importantly, we do not mean to suggest that the script captures everything in a movie. Audio-visual elements such as music and cinematography certainly play an important role. But capturing these aspects, and linking them to audience reactions, is far from trivial. We utilized the largest available movies script database, but there is much less moment-to-moment data available on cinematography or audio-visual aspects. Further, even if one could eventually construct such a dataset, quantification may be challenging. Less is known about how to measure a movie’s visual evolution of a movie (c.f., Cutting 2016), or how to quantify its soundtrack. Hopefully future work can examine these questions in greater detail.

These other aspects also help put the size of the observed effect in context. Unlike laboratory experiments, where everything else is carefully controlled, we focused on real world content evaluations in the noisy field. Thousands of things likely impact such evaluations (e.g., cinematography, music, the actors, acting quality, and the director), and given all of these other contributing factors, it would be quite surprising if speed of semantic progression alone explained a huge portion of the variance. After all, a movie could have no name actors, a terrible plot, bad acting and directing, and still move slowly at the beginning and fast at the end. That said, the fact that an effect of semantic progression emerges even with everything involved speaks to it’s persistence. Further, the fact that TV episodes show similar results speaks to the effect’s generalizability.

We also do not mean to suggest that semantic progression is the only element that contributes to narrative success. The experience of being transported by a story, use of rhetorical devices, and invoking emotion should all also play a role (e.g., Berger, Kim, & Meyer, 2021; Green & Brock, 2000). In addition to these, and potentially other aspects, however, we suggest that the overall narrative structure, and speed of semantic progression in particular, is important to consider.

More generally, these results speak to the role of psychological processes in cultural success. Just as sociocultural background shapes psychological processes (Markus & Kitayama, 1991), the reverse is also true; When shared across individuals, psychological processes can act as a selection mechanism, shaping whether cultural items succeed or fail (Kashima, 2008; Schaller & Crandall, 2004). Just as minimally counterintuitive narratives are more memorable and popular (Norenzayan et al., 2006), in this case, the speed of semantic progression within a narrative may shape its success.

Finally, these findings highlight the value of Natural Language Processing and emerging computational approaches to study human behavior. Language in an integral part of everyday life, and a wealth of related data from social media posts and online reviews to song lyrics, and movie scripts is now available. By quantifying features of textual data, natural language processing can unlock a range of interesting questions.

**Context Paragraph**

 We have recently been using natural language processing to extract behavioral insight from textual data. Some of this work has focused on cultural success. Why are some songs, books, movies, and even academic papers more successful than others? When shared across individuals, psychological processes can act as a selection mechanism, shaping whether cultural items succeed or fail. Consequently, textual features which shape things like memory, evaluation, or social transmission can impact success. Atypical songs, for example, are more popular (Berger & Packard 2019), as are those that use second person pronouns (because they remind people of close others, Packard & Berger 2020). Here we studied how the speed of semantic progression might shape the evaluation of movies (and tv episodes). The speed of semantic progression should impact how easy information is to process, as well as things like surprise, which together in turn should impact evaluations. These findings shed light on why narratives succeed and fail and how natural language processing can be used to study culture.

**References**

Alter, A., & Oppenheimer, D. (2009). Uniting the Tribes of Fluency to Form a Metacognitive Nation. *Personality and Social Psychology Review*.

Berger, J., Kim, Y., & Meyer, R. (2021). What Makes Content Engaging? How Emotional Dynamics Shape Success. *Journal of COnsumer Research, 48*, 235-250.

Bilgin, M., & Senturk, I. (2017). Sentiment analysis on Twitter data with semi-supervised Doc2Vec. *International Conference on Computer Science and Engineering.*

Biswas, A., Bhutani, M., & Sanyal, S. (2017). MRNet-Product2Vec: A Multi-task Recurrent Neural Network for Product Embeddings. *Machine Learning and Knowledge Discovery in Databases. ECML PKDD .*

Booker, C. (2004). *The seven basic plots: why we tell stories.*

Bost, X., Labatut, V., Gueye, S., & Linares, G. (2016). Narrative smoothing: Dynamic conversational network for the analysis of TV series plots. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, (pp. 1111-1118). San Francisco.

Calantone, R., Yeniyurt, S., Townsend, J. D., & Schmidt, J. B. (2010). The effects of competition in short product life-cycle markets: the case of motion pictures. *The Journal of Product Innovation Management*, 349-361.

Canet, F., Valero, M. A., & Codina, L. (2016). Quantitative approaches for evaluating the influence of films using the IMDb database. *Communication & Society*, 151-172.

Carroll, N. (1990). *The Philosophy of Horror.* New York: Routledge.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., & Specia, L. (2017). SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 1-14). Vancouver: Association for Computational Linguistics.

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., St. John, R., . . . Kurzweil, R. (2018). Universal Sentence Encoder for English. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 169–174). Brussels: Association for Computational Linguistics.

Chen, F., Liu, X., Proserpio, D., & Troncoso, I. (2020). Product2Vec: Understanding Product-Level Competition Using Representation Learning. *NYU Stern School of Business*.

Church, K. (2017). Word2Vec. *Natural language engineering, 23(1)*, p.155-162.

Delen, D., Sharda, R., & Kumar, P. (2007). Movie forecast guru: a web-based DSS for Hollywood managers. *Decision Support Systems*, 1151-1170.

Dellarocas, C., Zhang, X., & Awad, N. (2007). Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive Marketing*, 23-45.

Dodds, K. (2006). Popular geopolitics and audience dispositions: James Bond and the Internet Movie Database (IMDb). *Transactions of the Institute of British Geographers*.

Elberse, A., & Eliashberg, J. (2003). Demand and supply dynamics for sequentially released products in international markets: the case of motion pictures. *Marketing Science*, 329-354.

Eliashberg, J., Hui, S., & Zhang, Z. (2014). Assessing Box Office Performance Using Movie Scripts: A Kernel-Based Approach. *IEEE Transactions on Knowledge and Data Engineering, 26*(11), pp. 2639-2648.

Field, S. (1989). *Selling a Screenplay: The Screenwriter's Guide to Hollywood.* New York: Delacorte Press.

Field, S. (2005). *Screenplay: The Foundations of Screenwriting .* Delta.

Flavell, J. H., Miller, P. H., & Miller, S. A. (2001). *Cognitive development.* Upper Saddle River, NJ: Prentice Hall .

Franco-Salvador, M., Plotnikova, N., Pawar, N., & Benajiba, Y. (2017). Subword-based Deep Averaging Networks for Author Profiling in Social Media. In *CLEF (Working Notes)*.

Franzosi, R. (2010). *Quantitative narrative analysis.* London : SAGE.

Freytag, G. (1900). *Freytag's Technique of the drama.* Chicago: Scott, Foresman.

Freytag, G. (1900). *Freytag's Technique of the Drama, An Exposition of Dramatic Composition and Art by Dr. Gustav Freytag: An Authorized Translation From the Sixth German Edition by Elias J. MacEwan, M.A.* Chicago: Scott, Foresman and Company.

Frye, N. (2016). The Secular Scripture: A Study of the Structure of Romance. In *The Secular Scripture and Other Writings on Critical Theory* (pp. pp. 1-124).

Garg , N., Schiebinger, L., & Jurafsky, D. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, (pp. 3635-3644).

Green, M., & Brock, T. (2000). The Role of Transportation in the Persuasiveness of Public Narratives. *Journal of Personality and Social Psychology, 79(5)*, 701-721.

Hauge, M. (1991). *Writing Screenplays that Sell.* HarperCollins Publishers.

Humphreys, A., & Wang, R. (2018). Automated Text Analysis for Consumer Research. *Journal of Consumer Research*, 1274-1306.

Iyyer, M., Manjunatha, V., Boyd-Graber, J., & Daumé III, H. (2015). Deep Unordered Composition Rivals Syntactic Methods for Text Classification. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1681–1691). Beijing: Association for Computational Linguistics.

Jung, J. (2012). Attribute selection-based recommendation framework for short-head user group: An empirical study by MovieLens and IMDB. *Expert Systems with Applications*, 4049-4054.

Lau, J., & Baldwin, T. (2016). An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation. *Proceedings of the 1st Workshop on Representation Learning for NLP*, (pp. 78-86).

Liu, X., Lee, D., & Srinivasan, K. (2019). Large-Scale Cross-Category Analysis of Consumer Review Content on Sales Conversion Leveraging Deep Learning. *Journal of Marketing Research.*, Volume 56, Issue 6, pp. 918-943.

Loewenstein, J., & Heath, C. (2009). The Repetition-Break Plot Structure: A Cognitive Influence. *Cognitive Science* , 1-19.

Lowenstein, J., Raghunathan, R., & Heath, C. (2011). The Repetition-Break Plot Structure Makes Effective Television Advertisments. *Journal of Marketing, 75*, 105-119.

MacEwan, G. (1900). *Freytag's Technique of the drama.* Chicago: Scott, Foresman.

Munroe, R. (2009 ). *Movie narrative charts*. Retrieved from http://xkcd.com/657/

Nadiri, A., Gharekhani, M., & Khatibi, R. (2018). Mapping Aquifer Vulnerability Indices Using Artificial Intelligence-running Multiple Frameworks (AIMF) with Supervised and Unsupervised Learning. *Water Resources Management*, 3023-3040.

Netzer, O., Lemaire, A., & Herzenstein, M. (2019). When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications. *Journal of Marketing Research*.

Nguyen, T. T., & Maxtone-Graham, J. (Directors). (2011). *Jo ha kyu* [Motion Picture].

Packard, G., Moore, S. G., & McFerran, B. (2018). The Impact of Personal Pronoun Use in Customer-Firm Interactions. *Journal of Marketing Research*.

Propp, V. (1968). *Morphology of the Folktale.* Austin: University of Texas Press.

Quoc, L., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *Proceedings of the 31st International Conference on Machine Learning.*

Reagan, A., Mitchell, L., & Kiley, D. (2016). The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, Volume 5, Issue 1, pp. 1-12.

Rocklage, M., & Fazio, R. (2015). The Evaluative Lexicon: Adjective use as a means of assessing and distinguishing attitude valence, extremity, and emotionality. *Journal of Experimental Social Psychology*, 214-227.

Rozado, D. (2019). Using Word Embeddings to Analyze how Universities Conceptualize “Diversity” in their Online Institutional Presence. *Society*, 256-266.

Rozin, P., Rozin, A., Appel, B., & Wachtel, C. (2006). Documenting and Explaining the Common AAB Pattern in Music and Humor: Establishing and Breaking Expectations . *Emotion, 6(3)*, 349-355.

Sanakoyeu, A., Bautista, M., & Ommer, B. (2018). Deep unsupervised learning of visual similarities. *Pattern Recognition*, 331-343.

Shuai, Q., Huang, Y., & Jin, L. (2018). Sentiment Analysis on Chinese Hotel Reviews with Doc2Vec and Classifiers. *IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference.*

Silva, S., & Tehrani, J. (2016). Comparative phylogenetic analyses uncover the ancient roots of Indo-European folktales. *Royal Society Open Science*.

Spüler, M., Rosenstiel, W., & Bogdan, M. (2012). Online adaptation of a c-VEP Brain-computer Interface(BCI) based on error-related potentials and unsupervised learning. *Public Library of Science ONE*.

Sunder, V., Ganesh, L., & Marathe, R. R. (2018). A morphological analysis of research literature on Lean Six Sigma for services. *International Journal of Operations & Production Management*, 149-182.

Tausczik, Y., & Pennebaker, J. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 24-54.

Thompson, K. (1999). *Storytelling in the New Hollywood: Understanding Classical Narrative Technique.* Harvard University Press.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. *Proceedings of the 8th International Conference on Language Resources and Evaluation*, (pp. 2214-2218.).

Tiedemann, J., & Lison, P. (2016). OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. *Proceedings of the 10th International Conference on Language Resources and Evaluation.*

Todorov, T. (1975). *The fantastic : a structural approach to a literary genre.* Ithaca, N.Y. : Cornell University Press.

Toubia, O., Iyengar, G., Bunnell, R., & Lemaire, A. (2019). Extracting Features of Entertainment Products: A Guided Latent Dirichlet Allocation Approach Informed by the Psychology of Media Consumption. *Journal of Marketing Research*, 18-36.

Trottier, D. (1995). *The Screenwriter's Bible.* Silman-James Press.

Valls Vargas, J. (2017). *Narrative Information Extraction with Non-Linear Natural Language Processing Pipelines.*

Vany, A. D., & Walls, W. (1999). Uncertainty in the movie industry: does star power reduce the terror of the box office? *Journal of Cultural Economics*, 285-318.

Vonnegut, K. (1999). *Palm Sunday: An Autobiographical Collage.* New York: Dial Press.

Vonnegut, K. (n.d.). *Shapes of Stories*. Retrieved from https://www.youtube.com/watch?v=oP3c1h8v2ZQ

Vonnegut, K., & Comberg, D. (2004). *Kurt Vonnegut on the Shapes of Stories.* Retrieved from Youtube/dbcomberg: https://www.youtube.com/watch?v=oP3c1h8v2ZQ

Watson, A. (2019, September 30). *Value of the global entertainment and media market from 2011 to 2023*. Retrieved from Statista: https://www.statista.com/statistics/237749/value-of-the-global-entertainment-and-media-market/

Wells, D. (2012). *Writing Excuses 7.41: Seven-Point Story Structure*. Retrieved from Writing Excuses: https://writingexcuses.com/2012/10/07/writing-excuses-7-41-seven-point-story-structure/

Zhang, D., Xu, H., Su, Z., & Xu, Y. (2015). Chinese comments sentiment classification based on word2vec and SVM. *Expert Systems with Applications, 42*(4), 1857-1863.

1. Note that while our work is certainly related to the structures described by Rozin et al. and Lowenstein and Heath, there are some important differences. While they focus just on short content (i.e., jokes and ads), we consider content that is much longer and where similarity is more complex. While they dichotomize similarity (i.e., either things are similar or dissimilar), we allow similarity to be continuous and develop a method for empirically quantifying similarity. [↑](#footnote-ref-2)