# How quantifying the shape of stories predicts their success

**Olivier Toubia**[a,1,2], **Jonah Berger**[b,1], **and Jehoshua Eliashberg**[b]

[a]Marketing Division, Columbia Business School, Columbia University, New York, NY 10027; and [b]Marketing Department, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved May 4, 2021 (received for review June 7, 2020)

**Narratives, and other forms of discourse, are powerful vehicles for informing, entertaining, and making sense of the world. But while everyday language often describes discourse as moving quickly or slowly, covering a lot of ground, or going in circles, little work has actually quantified such movements or examined whether they are beneficial. To fill this gap, we use several state-of-the-art natural language-processing and machine-learning techniques to represent texts as sequences of points in a latent, high-dimensional semantic space. We construct a simple set of measures to quantify features of this semantic path, apply them to thousands of texts from a variety of domains (i.e., movies, TV shows, and academic papers), and examine whether and how they are linked to success (e.g., the number of citations a paper receives). Our results highlight some important cross-domain differences and provide a general framework that can be applied to study many types of discourse. The findings shed light on why things become popular and how natural language processing can provide insight into cultural success.**

discourse | natural language processing | cultural success | cultural analytics

Narratives and other forms of discourse are powerful vehicles for informing, entertaining, maintaining social order, and making sense of the world (1–5). People watch movies, read books, and consume other narratives, and politicians, journalists, and even academics craft discourse when communicating and sharing ideas.

But why are some narratives, or other types of discourse, more successful? And could a simple set of measures help explain variation in success in different domains?

Across disciplines, researchers have long been interested in features of narratives (6–9). While some narratives seem to move faster, for example, others seem to move slower (10, 11). Similarly, some stories are described as "covering lots of ground" and some narratives are described as "going in circles" (i.e., returning to similar themes again and again). But while researchers and laypeople alike often describe narratives as expressing movement in some abstract space, little empirical work has actually attempted to measure such movements (8, 9, 12). Further, even less work has examined whether such movements have any impact (13, 14). Might certain ways of unfurling a set of ideas increase their success? Are movies that cover a lot of ground, for example, evaluated more positively?

Note that these questions are not restricted to narratives. Narratives usually involve temporality (15, 16), or a sequence of events and actions, but similar questions could be asked of other types of discourse. Some academic papers or legal arguments, for example, seem to cover more ground than others and some textbooks move quickly through disparate ideas while others move more slowly. Might these features shape success (e.g., the number of citations an academic paper receives) and, if so, how?

Attempts to answer such questions have been hindered by quantification. It is difficult to measure, for example, whether one text moves quickly or slowly. While manually coding such aspects might be possible for a small number of texts (17, 18), it is often subjective and difficult to scale.

We fill this gap using natural language processing and machine learning. In any given text, some content appears earlier in the text and other content later. Using several state-of-the-art techniques, we plot chunks of texts as sequential points in a multidimensional space and extract features of the semantic path (i.e., speed, volume, and circuitousness). We examine tens of thousands of texts from a variety of domains (i.e., movies, TV shows, and academic papers) and test how speed, volume, and circuitousness relate to success (e.g., evaluations or citations).

Importantly, we do not mean to suggest that academic papers are narratives or that what makes a movie successful is the same as what makes an academic paper successful. In fact, our findings suggest the drivers are quite different. Rather, our goal is to provide simple measures that help quantify the semantic progression of texts and illustrate how such measures relate to success in different domains.

## Measures

To quantify semantic progression, we take texts (e.g., movies or academic papers), break them into approximately equal-sized chunks or windows, plot each chunk in a high-dimensional semantic space, and examine the path between chunks (see ref. 8 for related work using topic modeling). To do so, we use word embeddings (19), a technique that transforms words into high-dimensional numerical vectors such that the relationship between vectors captures the semantic relationship

## Significance

Why are some narratives (e.g., movies) or other texts (e.g., academic papers) more successful than others? Narratives are often described as moving quickly, covering lots of ground, or going in circles, but little work has quantified such movements or tested whether they might explain success. We use natural language processing and machine learning to analyze the content of almost 50,000 texts, constructing a simple set of measures (i.e., speed, volume, and circuitousness) that quantify the semantic progression of discourse. While movies and TV shows that move faster are liked more, TV shows that cover more ground are liked less. Academic papers that move faster are cited less, and papers that cover more ground or are more circuitous are cited more.

Author contributions: O.T., J.B., and J.E. designed research; O.T. and J.B. performed research; O.T. analyzed data; and O.T. and J.B. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1] O.T. and J.B. contributed equally to this work.

[2] To whom correspondence may be addressed. Email: ot2107@gsb.columbia.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2011695118/-/DCSupplemental.

Published June 25, 2021.

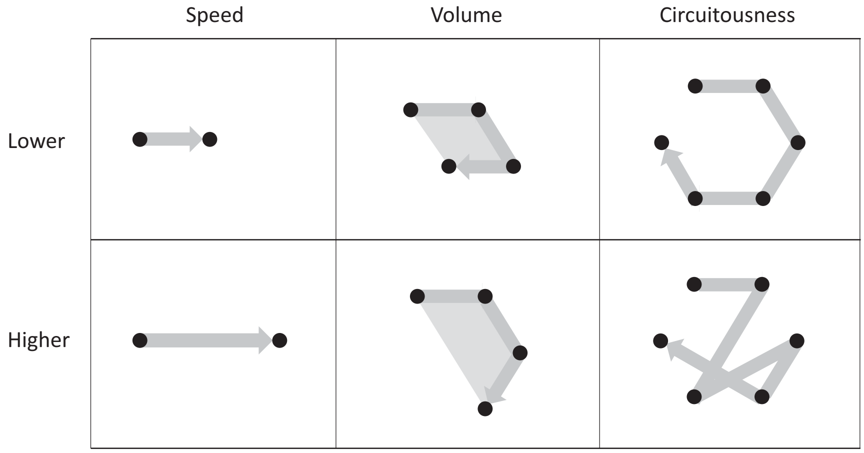

**Fig. 1.** Stylized illustration of the measures. Note that higher speed means more distance was covered in the same number of periods. Higher volume means that more ground was covered in the same number of periods. Higher circuitousness means that a less direct route was taken between a set of points.

between words. We take each word $w$, represented by $x_w$, a 300-dimensional vector; index windows by $t$; and define the average word embedding vector $x_t$ for the words in each window $t$. We denote as $T$ the number of windows in the document, and each text is represented by a sequence of $T$ points, $x_1, x_2, \ldots x_T$, in the 300-dimensional latent word embedding space (see *SI Appendix* for more detail).

From these sequences of points, we calculate new measures that characterize each text's semantic path. A natural first measure of progression is speed or pacing (10). Just as a car can move slower or faster (i.e., covering a smaller or larger distance in the same period), content can move slower or faster (i.e., dwelling on semantically related concepts or moving a larger distance between content that is less semantically related). To capture this, we measure the distance texts travel between consecutive chunks (Fig. 1). Word embeddings capture semantic similarity (20–22) (see *SI Appendix* for additional validation), so consecutive chunks that are farther away are more likely to discuss different topics or themes. We compute the Euclidean distance between consecutive points: $distance(t) = \|x_{t+1} - x_t\|$.[*] Normalizing total distance by text length generates the text's average speed: $speed = \frac{\Sigma_{t=1}^{T-1} distance(t)}{T-1}$.

Speed presents a tradeoff. Larger semantic shifts should make content more engaging and exciting (4), but require additional cognitive effort to process and connect (23). More difficult textbooks, for example, tend to have less semantic similarity between paragraphs (24), which should require greater processing to understand (25). Consequently, the excitement that speed generates likely comes at a (cognitive) cost. As such, speed may have a positive or negative relationship with success, depending on the context.

While speed is useful, it does not provide a complete picture. Two texts could cover the same distance with quite different semantic trajectories (e.g., one goes out and back, while the other goes out and then out even farther). Further, speed focuses only on consecutive points, but the meaning of content is often interpreted from the entire path (2).

To begin to capture these nuances, we measure the volume that a text covers (Fig. 1). While some content is described as covering a lot of ground or touching on many themes, other content is seen as covering less ground (26). We measure such volume by approximating points $\{x_1, x_2, \ldots x_T\}$ with an ellipsoid by solving an optimization problem that finds the minimum volume ellipsoid containing all of these points (27). Normalizing this by the dimensionality of the ellipsoid captures a text's volume and ancillary analyses show that this automated measure is correlated with human perceptions of ground covered (*SI Appendix*). Similar to speed, volume presents a tradeoff. Covering a lot of ground allows audiences to see and connect a wide range of topics but may increase the cognitive burden.

Volume captures the ground covered, but not how these points are covered, so to further quantify the path taken we measure circuitousness (see Fig. 1 and *SI Appendix* for a less simplified illustration). We identify the shortest path a text could have taken, given the first point $x_1$, the last point $x_T$, and the other set of points $\{x_2, \ldots x_{T-1}\}$ "visited" during the text. This optimization problem is a modified version of the well-known traveling salesman problem (28). After solving this, we quantify the extent to which the actual sequence $\{x_1, \ldots x_T\}$ deviates from optimal. Circuitousness is defined as the ratio of actual distance traveled to the shortest possible path. That is, $circuitousness = \frac{\Sigma_{t=1}^{T-1} distance(t)}{length\ of\ shortest\ path}$.

This measure captures human perceptions of circuitousness (*SI Appendix*). While circuitousness might seem undesirable, it may allow the audience to create new and deeper connections between previously explored themes (29).

## Results

We examine the relationship between these measures and success in three domains (i.e., movies, television shows, and academic papers). These domains were chosen based on data availability (*SI Appendix*), but the same approach could be applied to other types of texts (e.g., books or speeches). In addition to a standard set of control variables (*SI Appendix*), we control for textual content by including 100 topic intensities estimated by latent Dirichlet allocation (30). This ensures that our results are not driven by certain topics (e.g., love or social identity) being linked to success.

Examining over 4,000 movies finds that narratives that move faster (i.e., travel farther in consecutive periods, on average) are evaluated more favorably (Table 1, column 1).

Examining over 12,000 TV show episodes finds a similar result (Table 1, column 2). Given that distant points are less similar, they should be more surprising or unexpected. This result is consistent with the suggestion that rapid storyline changes can make narratives more engaging (4). TV show episodes that cover less volume are also evaluated more favorably. While one could

[*] Euclidean distance is highly correlated with another common measure, cosine similarity (correlations $>0.9$ in our datasets—see *Materials and Methods*).

**Table 1. Link between semantic progression and success**

| | Movies | TV show episodes | Academic papers |
|---|---|---|---|
| Average speed | 0.048* | 0.072* | −0.125* |
| Normalized volume | 0 | −0.082* | 0.095* |
| Circuitousness | 0 | 0.006 | 0.070* |
| Controls | | | |
| Year fixed effects | Yes | | Yes |
| Genre fixed effects | Yes | Yes | |
| Movie duration | Yes | | |
| TV channels fixed effects | | Yes | |
| Journal fixed effects | | | Yes |
| No. of pages | | | Yes |
| Log(words in document) | Yes | Yes | Yes |
| Log(sentences in document) | Yes | Yes | Yes |
| Topic intensities | Yes | Yes | Yes |
| No. of parameters | 169 | 148 | 158 |
| No. of observations | 4,118 | 12,336 | 29,300 |
| Mean-squared error | 0.711 | 0.793 | 1.066 |
| $R^2$ | 0.306 | 0.326 | 0.364 |

Note that all independent variables for which coefficients are reported are standardized. The dependent variable is not standardized. Parameters are estimated using a lasso regression. Confidence intervals are obtained via bootstrapping. *The 95% confidence interval does not include 0. Dependent variable is IMDB ratings for movies and TV show episodes and log(1 + citations) for academic papers.

interpret this as driven by TV show episodes being shorter than movies, note that volume is normalized by the number of chunks of text, indicating that even for text of the same length, TV show episodes that cover too much ground are evaluated less favorably. This may be driven by what audiences look for when they consume content from different mediums. While high-volume movies may fit audiences' expectations of being transported through a narrative, TV shows may be consumed as a quick diversion, and thus volume may have a more negative effect. Note that average speed and normalized volume are highly positively correlated in TV show episodes (*SI Appendix*), so each coefficient captures the effect of changing that variable, holding the others constant.

Examining citations of 29,000 academic papers published in 22 journals reveals a distinctly different pattern (Table 1, column 3). First, speed has the opposite effect; papers that move faster are cited less. Rapid changes should increase the effort required to follow an argument, which may reduce citations. Second, volume has the opposite effect; papers that cover more ground are cited more (consistent with the finding that papers that link disconnected areas of knowledge receive more cites, ref. 31). Finally, papers that are more circuitous receive more citations. Consistent with the fact that "spiral" curriculums that revisit similar topics help students learn (32), by repeatedly touching on similar themes, circuitousness may make it easier to integrate disparate information. Given average speed and circuitousness are highly correlated in academic papers (*SI Appendix*), each coefficient should be interpreted as capturing the effect of changing that variable, holding the others constant.

These effects are not trivial: A 1-SD increase in speed is associated with an approximately 12% decrease in citations [as noted in Table 1, the dependent variable for citations is log(1 + citations), so the coefficient should be interpreted carefully].[†] Volume and circuitousness show a similar effect (10 and 7%, respectively). Ancillary analyses (*SI Appendix*) provide further context, comparing these variables' explanatory power to noncontent variables shown to impact citations. The effect of a 1-SD change in average speed, for example, is comparable to the effect of a 1-SD change in institution prestige. Effects for TV show episodes and movies, which are not on a log scale, are more modest: A 1-SD increase in speed in movies (TV show episodes) is associated with an increase in average rating of 0.048 (0.072), on a 10-point scale with an SD of 1.01 (1.08). This is not surprising, given that movies and TV shows involve many nontextual factors (e.g., visual and audio elements).

Ancillary analyses also begin to examine how distance and volume change over the course of a text (*SI Appendix*). Some texts, for example, might move at a consistent speed while others have more variation. Some might cover a lot of volume early on but less so as the content evolves. Given that recent experiences (e.g., the end of a movie) can have a larger impact on evaluations (33), one could imagine that end effects are particularly important. To capture these aspects, we calculate how much each new period adds to the text's distance and volume and measure variation, trend, and end effects for incremental changes in both distance and volume (*SI Appendix*). Results are identical for movies and academic papers, but provide a more nuanced picture for TV show episodes: The positive effect of speed and the negative effect of volume are driven by changes that happen toward the end of the text.

## Discussion

While many have theorized about features of narratives, less work has formalized these intuitions, or tested whether certain features of discourse are linked to success. This paper provides a set of measures to quantify the semantic progression of texts and the ground they cover. In particular, we examined speed, volume, and circuitousness and how they relate to the success of movies, TV show episodes, and academic papers.

Results suggest that the features that make a successful movie may be different from those that make a successful TV show or academic paper, and future work might examine the roots of these cross-domain differences. The type of discourse (e.g., narrative vs. exposition), goal (e.g., to entertain vs. impart knowledge), modality (e.g., video vs. written), outcome measure (e.g., liking vs. citations), and audience expectations may all be important factors. Future work might also examine other types of texts (e.g., books, speeches, or documentaries). A preliminary analysis of 564 fiction books, for example, suggests that the measures reported here may also be helpful in understanding the success of books (*SI Appendix*).

These measures could also be applied to personal narratives. People often use narratives to explain and understand their own lives (34). Just as creative people have more distance (i.e., less semantic relatedness) between their thoughts (35), semantic progression in personal narratives may provide insight into the writer's personality or even how the act of writing impacts wellbeing (36).

Note that we focus on the semantic relation between chunks of text, not on the content of each specific chunk. Two movies may have completely different content (i.e., characters and setting) but have similar speed, volume, or circuitousness. The structures we examine are also different from, and complementary to, dramatic structure (6) or emotional trajectories (9, 12). Rather than examining how sentiment changes across the course of a narrative, for example, or where the climax occurs, we focus on the semantic relationship between different points (i.e., whether content moves quickly between disparate ideas or covers a set of points that are semantically less similar).

This work makes several theoretical contributions. First, it contributes to cultural analytics and understanding why cultural items succeed and fail. While some work suggests that

[†] Adding −0.125 to log(1 + citations) is equivalent to adding log(0.88), meaning that (1 + citations) is multiplied by 0.88, or decreased by 12%.

SOCIAL SCIENCES

cultural success is difficult, if not impossible, to predict due to dynamics of social influence (37), our paper finds that success is not completely random and that item characteristics may also play an important role. While this does not negate the importance of social dynamics, it highlights that with the right tools, researchers can extract features of cultural items that shed light on their success (31, 38–41).

Second, and along those lines, this work also highlights the value of natural language processing to study culture (42). Researchers have long been interested in quantifying narratives and cultural dynamics, but measurement has been a key challenge. Natural language processing, however, provides a reliable method of extracting features and doing so at scale (43, 44). Consequently, it opens up a range of interesting avenues for further research. These tools may be particularly useful for researchers in philosophy, English, and other disciplines who are interested in quantifying aspects of discourse. Researchers in the digital humanities have recently made a number of interesting advances (8, 9, 13, 45, 46), and with the right tools, hopefully scholars can begin to quantify features of culture only dreamed about previously.

Third, our findings dovetail with recent work on how psychological processes shape collective outcomes. A great deal of research has demonstrated that sociocultural background shapes individual-level psychological process (e.g., cognition and attribution) (47). But the reverse is also true; when shared across individuals, psychological processes can act as a selection mechanism, shaping the content of collective culture (48–50). In this case, how people process information, and evaluate content, may shape which movies, shows, and academic papers are more successful.

Future work might examine the underlying cognitive and social processes that underlie these effects. As noted, desire for stimulation or for surprise, cognitive complexity, or processing ease, and a number of other aspects may all play a role. While it is difficult to test psychological mechanisms in field data, subsequent experimental investigations can hopefully manipulate different aspects directly and examine the underlying processes in greater detail.

Work might also examine the consequences of these features for other downstream outcomes (e.g., comprehension, memory, and persuasion). Readers might learn more from content that covers more volume, for example, although covering too much ground too quickly may mitigate this effect. Similarly, circuitousness may, at least in some cases, improve memory by connecting new ideas to previously explored themes. Semantic progression may also impact the persuasiveness of things like political speeches or legal arguments.

In conclusion, narratives and other forms of discourse offer a fertile ground to study features of content that shape success, and their psychological underpinnings. Emerging natural language-processing tools should open up a range of interesting directions for further study.

## Materials and Methods

**Data Preprocessing.** For each document (i.e., each movie, TV show episode, or academic paper), we tokenize the text (i.e., extract individual words from the script), transform each word to lowercase, and look up the embedding of each word $w$, denoted as $x_w$, a 300-dimensional vector. We use the word2vec word embedding model trained by ref. 19, which represents approximately 1 million words as real vectors in a 300-dimensional latent space. Other embedding approaches (e.g., Glove) yield similar results (*SI Appendix*).

We split each document into nonoverlapping windows of approximately equal size. Based on prior work (51), we use the same target window size of 250 words across our three datasets, but to avoid breaking up sentences, some windows are slightly larger than 250 words. For example, if the first 10 sentences contain 240 words and the 11th sentence contains 15 words, we include all 11 sentences and end up with 255 words in the window. Results are similar for windows of other sizes (*SI Appendix*). We index windows by $t$ and define the average word embedding vector $x_t$ for the words in each window $t$, $x_t = \frac{\Sigma_{w \in c_t} x_w}{|c_t|}$, where $c_t$ is the set of words in window $t$.

**Detailed Description of Volume Calculation.** To calculate volume, we start by finding the minimum-volume enclosing ellipsoid containing points $\{x_1, \ldots x_T\}$. The problem may be written as follows (27):

Maximize$_{d,A}$ det($A$)
subject to:
$(x_t - d)^\top A(x_t - d) \leq 1,\ t = 1, \ldots, T$
$A$ is a positive definite matrix.

When the rank of the subspace spanned by the vectors $\{x_t\}$ is equal to 300 and there are at least 301 points, we solve the above problem directly using matlab code made available by ref. 27.

However, if the number of points is less than or equal to the dimensionality of the word embedding space (300), or if the subspace spanned by the vectors $\{x_t\}_{t=1,\ldots T}$ is not full rank, then the above problem is degenerate, as it is possible to cover all points with a "flat ellipse" that lives in a subspace of dimension <300. For example, two points in a three-dimensional space may be covered by a line segment, and three points may be covered by an ellipse in the two-dimensional plane that contains these three points, which has a volume of 0 in the original three-dimensional space (see *SI Appendix*, Fig. S1 for an illustration). In these cases, we find the minimum-volume ellipsoid that contains all of the points $\{x_t\}$, in the corresponding subspace (instead of the entire word embedding space). See *SI Appendix* for details.

Once the minimum-volume enclosing ellipsoid has been computed, we find the eigenvalues of the (positive definite) matrix that defines this ellipsoid. The lengths of the axes of the ellipsoid are given by the inverse of the square root of the eigenvalues.

The volume of the minimum enclosing ellipsoid that contains a set of points is equal to the volume of the unit sphere, multiplied by the product of the lengths of its axes (52). Therefore, the product of the length of the axes gives us a measure of volume relative to a unit sphere. To compare texts of different lengths, we normalize this measure by the dimensionality of the ellipsoid; i.e., we use the geometric mean (rather than the product) of the lengths of the axes of the minimum-volume ellipsoid corresponding to points $\{x_1, \ldots x_T\}$, as our normalized measure of volume. This measure may be interpreted as the ground covered by the text.

**Data and Empirical Approach.** The first dataset examines Internet Movie Database (IMDB) ratings of 4,118 movies based on their subtitles. The second dataset examines IMDB ratings of 12,401 episodes of TV shows based on closed captions. The third dataset examines the citations received by 29,300 academic articles published in 22 different journals in psychology, economics, sociology, political science, and anthropology between 1990 and 2019. See *SI Appendix* for more detail.

To reduce overfitting, limit the number of nonzero coefficients, and reduce the effects of multicollinearity, we use lasso regression (53). Results are almost identical using ordinary least-squares regression (*SI Appendix*), but lasso seemed more appropriate because it is known to address multicollinearity. We log transform average speed, circuitousness, and normalized volume. Only observations for which all variables are available are included in the analysis. See *SI Appendix* for details.

**Data Availability.** Some study data are available upon request.

**ACKNOWLEDGMENTS.** Outstanding research assistance was provided by Reihane Boghrati, Yanyan Li, Lan Luo, Ahmed M'rad, Sanjana Rosario, and Xinyu Wei, and Grant Packard designed the figure. Velibor Misic provided insightful comments on this paper.

1. R. F. Baumeister, L. Zhang, K. D. Vohs, Gossip as cultural learning. *Rev. Gen. Psychol.* **8**, 111–121 (2004).
2. J. S. Bruner, *Acts of Meaning* (Harvard University Press, 1990), vol. 3.
3. R. Dunbar, R. I. M. D. Dunbar, *Grooming, Gossip, and the Evolution of Language* (Harvard University Press, 1998).
4. K. J. Gergen, M. M. Gergen, *Narrative Form and the Construction of Psychological Science* (Praeger Publishers/Greenwood Publishing Group, 1986).
5. R. A. Mar, K. Oatley, The function of fiction is the abstraction and simulation of social experience. *Perspect. Psychol. Sci.* **3**, 173–192 (2008).
6. G. Freytag, *Freytag's Technique of the Drama: An Exposition of Dramatic Composition and Art* (Scholarly Press, 1896).
7. K. Vonnegut, *Palm Sunday: An Autobiographical Collage* (Dial Press, 1999).
8. B. M. Schmidt, "Plot arceology: A vector-space model of narrative structure" in *2015 IEEE International Conference on Big Data (Big Data)* (IEEE, 2015), pp. 1667–1672.

9. J. Gao, M. L. Jockers, J. Laudun, T. Tangherlini, "A multiscale theory for the dynamical evolution of sentiment in novels" in *2016 International Conference on Behavioral, Economic and Socio-Cultural Computing (BESC)* (IEEE, 2016), pp. 1–4.
10. K. Hume, Narrative speed in contemporary fiction. *Narrative* **13**, 105–124 (2005).
11. J. E. Cutting, The evolution of pace in popular movies. *Cognit. Res. Principles Implications* **1**, 30 (2016).
12. A. J. Reagan, L. Mitchell, D. Kiley, C. M. Danforth, P. S. Dodds, The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Sci.* **5**, 31 (2016).
13. J. Archer, M. L. Jockers, *The Bestseller Code: Anatomy of the Blockbuster Novel* (St. Martin's Press, 2016).
14. R. L. Boyd, K. G. Blackburn, J. W. Pennebaker, The narrative arc: Revealing core narrative structures through text analysis. *Sci. Adv.* **6**, eaba2196 (2020).
15. P. Ricoeur, Narrative time. *Crit. Inq.* **7**, 169–190 (1980).
16. W. Labov, Some further steps in narrative analysis. *J. Narrat. Life Hist.* **7**, 395–415 (1997).
17. A. Hillier, R. P. Kelly, T. Klinger, Narrative style influences citation frequency in climate change science. *PloS One* **11**, e0167983 (2016).
18. B. Freeling, Z. A. Doubleday, S. D. Connell, Opinion: How can we boost the impact of publications? Try better writing. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 341–343 (2019).
19. T. Mikolov, É. Grave, P. Bojanowski, C. Puhrsch, A. Joulin, "Advances in pre-training distributed word representations" in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, S. Goggi, H. Mazo, Eds. (European Language Resources Association, Luxembourg, 2018), pp. 52–55.
20. S. Bhatia, Associative judgment and vector space semantics. *Psychol. Rev.* **124**, 1 (2017).
21. N. Garg, L. Schiebinger, D. Jurafsky, J. Zou, Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E3635–E3644 (2018).
22. A. C. Kozlowski, M. Taddy, J. A. Evans, The geometry of culture: Analyzing the meanings of class through word embeddings. *Am. Sociol. Rev.* **84**, 905–949 (2019).
23. J. L. Monahan, S. T. Murphy, R. B. Zajonc, Subliminal mere exposure: Specific, general, and diffuse effects. *Psychol. Sci.* **11**, 462–466 (2000).
24. P. W. Foltz, W. Kintsch, T. K. Landauer, The measurement of textual coherence with latent semantic analysis. *Discourse Process.* **25**, 285–307 (1998).
25. D. S. McNamara, E. Kintsch, N. B. Songer, W. Kintsch, Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognit. Instruct.* **14**, 1–43 (1996).
26. J. Herbert, *Journalism in the Digital Age: Theory and Practice for Broadcast, Print and Online Media* (CRC Press, 1999).
27. N. Moshtagh *et al.*, Minimum volume enclosing ellipsoid. *Convex Optimization* **111**, 1–9 (2005).
28. G. B. Dantzig, *Linear Programming and Extensions* (Princeton University Press, 1998), vol. 48.
29. J. Loewenstein, C. Heath, The repetition-break plot structure: A cognitive influence on selection in the marketplace of ideas. *Cognit. Sci.* **33**, 1–19 (2009).
30. D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003).
31. B. Uzzi, S. Mukherjee, M. Stringer, B. Jones, Atypical combinations and scientific impact. *Science* **342**, 468–472 (2013).
32. R. M. Harden, What is a spiral curriculum? *Med. Teach.* **21**, 141–143 (1999).
33. D. Kahneman, B. L. Fredrickson, C. A. Schreiber, D. A. Redelmeier, When more pain is preferred to less: Adding a better end. *Psychol. Sci.* **4**, 401–405 (1993).
34. D. P. McAdams, K. C. McLean, Narrative identity. *Curr. Dir. Psychol. Sci.* **22**, 233–238 (2013).
35. K. Gray *et al.*, "Forward flow": A new measure to quantify free thought and predict creativity. *Am. Psychol.* **74**, 539 (2019).
36. J. W. Pennebaker, Expressive writing in psychological science. *Perspect. Psychol. Sci.* **13**, 226–229 (2018).
37. M. J. Salganik, P. S. Dodds, D. J. Watts, Experimental study of inequality and unpredictability in an artificial cultural market. *Science* **311**, 854–856 (2006).
38. C. A. Bail, Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 11823–11828 (2016).
39. J. Berger, G. Packard, Are atypical things more popular?. *Psychol. Sci.* **29**, 1178–1184 (2018).
40. O. Toubia, G. Iyengar, R. Bunnell, A. Lemaire, Extracting features of entertainment products: A guided latent Dirichlet allocation approach informed by the psychology of media consumption. *J. Market. Res.* **56**, 18–36 (2019).
41. O. Toubia, A Poisson factorization topic model for the study of creative documents (and their summaries). *J. Market. Res.*, 10.1177/0022243720943209 (2020).
42. J. Berger, G. Packard, Using language to understand people and culture. *Amer. Psych.*, 10.1037/amp0000882.
43. J. W. Pennebaker, M. R. Mehl, K. G. Niederhoffer, Psychological aspects of natural language use: Our words, our selves. *Annu. Rev. Psychol.* **54**, 547–577 (2003).
44. A. Rule, J.-P. Cointet, P. S. Bearman, Lexical shifts, substantive changes, and continuity in state of the union discourse, 1790–2014. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 10837–10844 (2015).
45. T. J. Gray, A. J. Reagan, P. S. Dodds, C. M. Danforth, English verb regularization in books and tweets. *PloS One* **13**, e0209651 (2018).
46. M. Sims, D. Bamman, Measuring information propagation in literary social networks. https://arxiv.org/abs/2004.13980v2 (6 October 2020).
47. H. R. Markus, S. Kitayama, Culture and the self: Implications for cognition, emotion, and motivation. *Psychol. Rev.* **98**, 224 (1991).
48. C. Heath, C. Bell, E. Sternberg, Emotional selection in memes: The case of urban legends. *J. Pers. Soc. Psychol.* **81**, 1028 (2001).
49. M. Schaller, C. S. Crandall, *The Psychological Foundations of Culture* (Psychology Press, 2003).
50. Y. Kashima, A social psychology of cultural dynamics: Examining how cultures are formed, maintained, and transformed. *Soc. Personal. Psychol. Compass* **2**, 107–120 (2008).
51. J. Berger, Y. D. Kim, R. Meyer, What makes content engaging? How emotional dynamics shape success, *J. Cons. Res.*, 10.1093/jcr/ucab010 (2021).
52. A. J. Wilson, Volume of n-dimensional ellipsoid. *Sci. Acta Xaveriana* **1**, 101–106 (2010).
53. R. Tibshirani, Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* **58**, 267–288 (1996).