Article

Journal of Language and Social Psychology
1–19

DOI: 10.1177/0261927X18824859
journals.sagepub.com/home/jls
SAGE

What Words Are Worth: National Science Foundation Grant Abstracts Indicate Award Funding

David M. Markowitz[1]

Abstract

Can word patterns from grant abstracts predict National Science Foundation (NSF) funding? In an analysis of over 7.4 million words covering 19,569 proposals, this article presents evidence that the writing style of NSF grant abstracts corresponds to the amount of money received for the award. The data describe a clear relationship between word patterns and funding magnitude: Grant abstracts that are longer than the average abstract, contain fewer common words, and are written with more verbal certainty receive more money from the NSF (approximately $372 per one-word increase). While such language patterns correspond to award amount, they largely contradict the NSF's call to communicate science in a plain manner, suggesting an inconsistency between the injunctive norms of the NSF and the descriptive norms of science writing. Broadly, the results support a tradition of research that uses big text data to evaluate social and psychological dynamics.

Keywords

grant funding, big text data, National Science Foundation, open data, automated text analysis

The rise of big text data has motivated a new frontier of research into the relationship between language patterns and social and psychological dynamics (Lazer et al., 2009). Records from social media (e.g., Twitter; Golder & Macy, 2011; Murphy, 2017), blogs (e.g., LiveJournal; Cohn, Mehl, & Pennebaker, 2004), and professional databases

[1]University of Oregon, Eugene, OR, USA

Corresponding Author:
David M. Markowitz, School of Journalism and Communication, University of Oregon, Allen Hall, Eugene, OR 97401, USA.
Email: dmark@uoregon.edu

(e.g., annual reports to the U.S. Securities and Exchange Commission; Li, 2008) provide naturally occurring language data that can reveal how word patterns correlate with processes that were once observable only in the laboratory. For example, early studies on emotional contagion, or the transfer of positive or negative emotion from one person to another, required tightly controlled experiments to discover the effect through nonverbal behavior (Hatfield, Cacioppo, & Rapson, 1993), but the phenomenon can also be observed in large online networks through language (Kramer, Guillory, & Hancock, 2014). Across a variety of settings, evidence suggests that social and psychological information can be collected from the words that people communicate (Pennebaker, 2011).

Several recent studies highlight how language patterns are socially and psychologically rich, from indicating intelligence and complex thinking (e.g., high rates of articles and prepositions correlate with high academic performance in college; Pennebaker, Chung, Frazee, Lavergne, & Beaver, 2014), to social status among group members (e.g., nonleaders use more first-person singular pronouns than leaders; Kacewicz, Pennebaker, Davis, Jeon, & Graesser, 2014; Markowitz, 2018), and political affiliation (e.g., conservatives display a different trace of negative emotions than liberals; Robinson, Boyd, & Fetterman, 2014). Beyond description, language patterns can also be predictive. For example, recalling a trauma (e.g., sexual assault) and using death-related words predict worse psychological outcomes for people who suffer from posttraumatic stress disorder (Alvarez-Conrad, Zoellner, & Foa, 2001), stability in a romantic relationship is often predicted by partners who use consistent rates of function words (e.g., articles, pronouns; Ireland et al., 2011), and the popularity of a song is predicted by how atypical its lyrics are from the music genre (J. Berger & Packard, 2018).

The current study adopts this approach and examines text data from an archive of National Science Foundation (NSF) grant proposals to investigate if the writing style of a grant abstract corresponds to the amount of funding received for the award. After controlling for several factors, including the award directorate, grant duration, and award year, the data show that grant abstracts written in complex manner (e.g., abstracts that contain more words, fewer common words) with high rates of verbal confidence (e.g., more certainty, less tentativeness) are awarded more money.

Language Patterns and Financial Funding

The idea that language patterns from a grant proposal can relate to the magnitude of NSF funding is feasible given prior work that has used word patterns to examine online peer-to-peer loan requests, where people create a text profile to persuade lenders about their creditworthiness. In one analysis of approximately 200,000 loan requests from Prosper.com (Larrimore, Jiang, Larrimore, Markowitz, & Gorski, 2011), verbosity and the expression of certainty positively correlated with funding success. That is, an increase of 60 words in a person's loan request increased the probability of funding by approximately 1%. Language patterns can also predict if people will default on their loan (e.g., words related to religion negatively associate with loan repayment; see Stephens-Davidowitz, 2017). People who write more, convey a more confident loan request description, and focus on their loan details often appear more credible, and this improves the probability of receiving money.

Peer-to-peer loan requests and grant proposals have similarities that can motivate predictions about how the word patterns in a grant abstract may associate with funding received for an award. First, the discourse objective is similar: written requests for money. Second, grant and loan reviewers are typically unknown to the authors. While grant authors may be familiar with NSF directorates, it is unclear who will appraise each project. This suggests that grant writers must provide enough detail in the proposal to be considered an expert, but the writing style must also be approachable for readers in other disciplines to find value in the research (Feng & Shi, 2004). For loan requesters, such details include information about how the loan will be used, when it will be repaid, and relevant credit history, while grant writers may highlight the novelty of their ideas and publication record to emphasize their productivity. Finally, peer-to-peer lending and grant writing are both high-stakes settings. People can ask for an online loan for many reasons (e.g., paying off credit card debt, wedding expenses), and convincing lenders of creditworthiness can substantially change a person's financial outlook. Similarly, persuading reviewers about the value of a research program can lead to an increase in academic resources and an enhanced reputation as a scholar (Connor & Mauranen, 1999).

Predictions

Two sets of independent variables are used to evaluate the amount of financial grant funding received from the NSF: (1) complexity of the discourse and thinking style of the grant writers and (2) verbal expressions of confidence in the science proposal. This article leverages empirical evidence from other studies to motivate predictions about the relationship between grant abstract writing and funding amount. It also explores how other language dimensions may correlate with funding to uncover new patterns that could not be anticipated, a priori.

Discourse and Thinking Style Complexity. Prior peer-to-peer lending research (Larrimore et al., 2011) suggests that a more complex writing structure (e.g., text that contains more words) can lead to a greater probability of funding. Other studies have found similar effects, such as lengthier Amazon reviews predict perceived review helpfulness (Mudambi & Schuff, 2010), longer product descriptions on eBay predict a higher number of bids and greater selling prices of consumer products (Flanagin, 2007), and longer and more complex online medical advice positively relates to the perceived expertise of a doctor (Toma & D'Angelo, 2015). In these cases, verbosity and increased verbal complexity lead to positive appraisals of a target often because of a credibility heuristic. Complex discourse patterns can make communicators appear trustworthy and can help to reduce uncertainty among strangers in environments with limited cues (Flanagin, 2007; Toma & D'Angelo, 2015). Therefore, the more information that is provided in a grant abstract, the more that readers can plausibly infer the credibility of the authors.

Although longer grants should be positively associated with the amount of NSF funding received for an award, prior work also observed that there are limits to how much verbosity is associated with funding success (Larrimore et al., 2011). Lengthy loan descriptions can hurt a person's chances of receiving funding, while increasing the word count of a short description can lead to an increase in funding probability. Taken together, grant proposals with longer than average word counts should receive more funding, though excessively long abstracts should receive less funding (as reflected by a quadratic term for word count).

In addition to the prior empirically grounded prediction, exploratory complexity dimensions were evaluated to understand how other language patterns may relate to the amount of grant funding received. The first dimension, words per sentence, provides an approximation of the structural complexity of the text. Sentences that contain more words are often more difficult to read and comprehend than sentences with lower word counts (Flesch, 1948; Markowitz & Hancock, 2016). Therefore, consistent with the prediction made for the total number of words per grant abstract, text with more words per sentence should also associate with more funding for NSF proposals.

The second dimension, common words, is the percentage of words in an abstract that are considered common English terms. The NSF "is committed to writing new documents in plain language" in order for members of the public to understand the science that was funded (NSF, 2018). This ideal is consistent with the Plain Writing Act of 2010, a federal law enacted to ensure that government writing is "clear, concise, well-organized, and follows other best practices appropriate to the subject or field and intended audience" (p. 2681). Prior work has used the rate of common words to evaluate social and psychological dynamics in science writing, such as the language patterns of fraudulent science articles. Papers by scientists who committed data fraud contained fewer common words from an empirically validated English dictionary (Pennebaker, Booth, Boyd, & Francis, 2015) than papers based on genuine data (e.g., using the word *occupation* instead of *job* often reflects less clear text; Markowitz & Hancock, 2016). It is therefore reasonable to suggest that the NSF will reward proposals that follow institutional guidelines for plain writing. Grant abstract content that is described with more common words, consistent with the NSF's message, should receive more award money.

A final exploratory dimension considers the complexity of the writers' thinking as reflected by function words. Function words (e.g., articles, negations) are small style words (e.g., *a, not*, respectively) that have little meaning outside of a sentence but hold significant social and psychological value, such as indicating personality traits and social cohesion (see Chung & Pennebaker, 2007, for a review). In English, there are approximately 400 function words, but they comprise nearly 50% of the words that people use in writing and conversation. On average, high rates of articles and prepositions reflect complex and analytic thinking, while high rates of pronouns, adverbs, conjunctions, and negations reflect storytelling and narrative thinking that are comparatively simple (Pennebaker et al., 2014). Therefore, consistent with the prior rationale relating common words and the amount of grant award money received, NSF funding should be associated with a simple thinking style with low rates of jargon and discourse patterns that tell a story.

In sum, the relationship between complexity and the amount of funding awarded by the NSF (in dollars) leads to the following hypothesis:

> **Hypothesis 1a-d (H1a-d):** Grant abstracts that contain (a) more words, (b) more words per sentence, (c) more common words, and (d) a narrative writing style will associate with more money awarded by the NSF.

Confidence in the Science Proposal. Prior work has also observed that fully funded loan requests contain more verbal certainty (e.g., high rates of words such as *absolute, definite, essential*) and a trend toward less tentativeness (e.g., low rates of words such as *if, hope, guess*; Larrimore et al., 2011). This idea is consistent with the confidence heuristic in judgment and decision-making research (Price & Stone, 2004), which suggests that people who appear confident and certain are appraised positively (e.g., they are viewed as more knowledgeable) compared to people who do not appear confident. For example, using phrases such as "I am not sure . . ." leads to negative evaluations of advice-givers relative to expressions of certainty (e.g., stating "I am very confident that . . .") or the absence of certainty (Gaertig & Simmons, 2018). Recent work also suggests that written expressions of certainty positively associate with favorable perceptions of a research project during a first stage of review for European Research Council grants (van den Besselaar, Sandström, & Schiffbaenker, 2018). Together, this evidence presupposes that grant proposal abstracts with more certainty and less tentativeness should be appraised favorably and receive more financial funding from the NSF.

A third measure of confidence is exploratory and considers how well the authors have explained their proposal. One way to describe why the science is interesting, exciting, and important is to explain how elements of the study are related, why the proposal is worthy of funding, and how the issues in the proposal are timely. These goals can be achieved by increasing the rate of causal language in the grant writing. Causal claims are foundational to science as an institution (Marini & Singer, 1988), and writing in a causal manner (e.g., using words such as *infer, intend*, or *solve*) may encourage grant readers to suspect that the proposal is more rigorous or that the authors have sufficiently explained the rationale for funding their science. Grant proposals are often considered acts of persuasion (Connor & Wagner, 1998), and causal claims can bolster the persuasiveness of the writing. Together, the amount of grant funding should also increase as the rate of causality increases.

> **Hypothesis 2a-c:** Grant abstracts that contain (a) more verbal certainty, (b) less tentativeness, and (c) more causal language will associate with more money awarded by the NSF.

Method

Award abstracts from funded NSF proposals are public (https://www.nsf.gov/awardsearch/), and the awards in this study were limited to grants submitted from the United States. An initial search produced over 200,000 award abstracts, though records from the late 20th century are often expired and incomplete. Therefore, several selection criteria were used to create a refined data set of NSF proposal abstracts.

First, a maximum of 3,000 abstracts per state, from active grants with start dates between 2010 and 2018, were included based on data extraction limits. Second, awards must be labeled as a Standard Grant from the NSF and not a continuing grant or fellowship. After removing abstracts with less than 25 words, the final data set contained a sample of 19,569 award abstracts, covering 7,405,295 words and over $8.6 billion in awarded grant funding. The data and code are publicly available on the Open Science Framework (https://osf.io/n29yv/).

Database Descriptive Statistics

Awards were distributed across all NSF directorates: Biological Sciences (n = 1,688), Computer and Information Science and Engineering (n = 3,438); Engineering (n = 5,159); Geosciences (n = 1,518); Mathematical and Physical Sciences (n = 3,416); Social, Behavioral and Economic Sciences (n = 1,816); and Education and Human Resources (n = 2,356), including those from the Office of the Director (n = 178). The average amount of money awarded for each proposal was $440,091.02 (*SD* = $475,979.32; Q1 = $199,396, *Mdn* = $341,798, Q3 = $500,000). The award amount was not normally distributed (skewness = 4.51, kurtosis = 35.31) and therefore natural log-transformed, which substantially reduced distribution issues (natural log-transformation skewness = −1.05, kurtosis = 1.46). Data were gathered from all 50 U.S. states and the District of Columbia.

Automated Text Analysis

Award abstracts were analyzed using Linguistic Inquiry and Word Count (LIWC: Pennebaker et al., 2015). This automated text analysis program counts words from a piece of text as a percentage of the total word count and increments an internal dictionary of social, psychological, and syntactic categories. For example, the phrase "In this project, we will evaluate" contains six words and increments several LIWC categories, including but not limited to: pronouns (*this, we*; 33.33% of the total word count) and verbs (*will*; 16.67% of the total word count). LIWC's word counting method is a valid approach to quantify word patterns for spoken and written text (Boyd & Pennebaker, 2015; Tausczik & Pennebaker, 2010), with hundreds of studies using this tool for social and psychological evaluations of language data across disciplines. Language dimensions in this study were drawn from the standard LIWC2015 dictionary unless noted. Nontraditional characters and HTML tags were removed to the best of the author's ability.

Language Predictors: Discourse and Thinking Style Complexity

The first set of word patterns to predict funding from the NSF related to discourse and thinking style complexity, operationalized by word count, the percentage of common words in English, words per sentence, and analytic thinking variables.

Word Count. Word count considers the total number of words in each abstract. On average, award abstracts contained 378.42 words (*SD* = 117.68 words; Q1 = 295 words, *Mdn* = 368 words, Q3 = 447 words). Word count squared was also entered into the model to evaluate if lengthy abstracts associate with less funding (Larrimore et al., 2011). Akaike information criterion (AIC) was reduced after including the quadratic term for word count in the full model (AIC with WC^2 = 41,927.73, AIC without WC^2 = 41,971.82), suggesting a better model fit when word count squared was added as a predictor.

Words per Sentence. The words per sentence category uses punctuation markers to count the average number of words in each sentence of the abstract (*M* = 28.28 words per sentence, *SD* = 5.77 words per sentence; Q1 = 24.40 words per sentence, *Mdn* = 27.50 words per sentence, Q3 = 31.24 words per sentence). More words per sentence suggests that the text is more complex, since longer sentences are more difficult to read and comprehend than shorter sentences (Flesch, 1948). To more accurately capture words per sentence, recurring words with punctuation marks that do not typically designate the end of a sentence (e.g., U.S.A., Dr.) were replaced with conceptually equivalent values (e.g., USA, Dr).

Common Words. The percentage of common words was operationalized by the LIWC dictionary category (*M* = 72.06%, *SD* = 5.88%; Q1 = 67.85%, *Mdn* = 71.75%, Q3 = 76.29%), which is a proxy for the number of everyday words in English (Markowitz & Hancock, 2016; Tausczik & Pennebaker, 2010). Recall, the LIWC dictionary contains a range of dimensions that tap social (e.g., words related to friends, family), psychological (e.g., emotion terms, words that indicate cognitive processes), part of speech (e.g., function words such as pronouns, articles), and behavioral characteristics (e.g., words that indicate power, motion, or risk). Words inside the LIWC dictionary are common in everyday communication while words outside of the LIWC dictionary are often considered jargon, or specialized terms (Markowitz & Hancock, 2016).

Analytic Speech. Analytic speech considers the complexity of the writer's thinking style through function words (Chung & Pennebaker, 2007) and is measured on a scale of 0 to 100. A high score on this index suggests that the writers use more articles and prepositions relative to pronouns, conjunction, negations, and adverbs (*M* = 96.25, *SD* = 2.96; Q1 = 95.24, *Mdn* = 97.09, Q3 = 98.24). Articles make references to nouns and prepositions help to form relationships between objects; high rates of these language features are often associated with concrete, complex, and analytic thinking (Pennebaker et al., 2014). On the other hand, pronouns, conjunctions, negations, and adverbs are storytelling words often found in narratives. These language categories reflect dynamic storytelling discourse that is often cognitively simple relative to discourse with high rates of articles and prepositions (Markowitz, 2018).

Language Predictors: Confidence in the Science Proposal

The second set of word patterns consists of three language dimensions—certainty terms, words related to tentativeness, causal terms—which describe how authors verbally express confidence in their science.

Certainty Terms and Tentativeness. Words related to certainty describe the writer's confidence in an idea (M = 0.74%, SD = 0.59%; Q1 = 0.32%, Mdn = 0.64%, Q3 = 1.03%). Terms such as *absolute* or *exact* convey precise thinking and clear expectations, while tentativeness (e.g., words such as *almost, doubt, hope*) expresses caution or hesitation (M = 1.25%, SD = 0.99%; Q1 = 0.56%, Mdn = 1.05%, Q3 = 1.70%).

Causal Terms. Causal terms describe why the science is important (M = 3.13%, SD = 1.50%; Q1 = 2.04%, Mdn = 2.96%, Q3 = 4.01%). Words such as *affect* or *because* make causal explanations between variables and can explain a rationale for funding.

A bivariate correlation matrix for all primary variables in this study is located in Table 1.

Results

The data were analyzed using a mixed effects regression with the lme4 package (Version 1.1-19) in R (Version 3.5.1). Each model contained standardized dependent and independent variables, predicting the grant award amount (natural log-transformed) from language patterns and including several control variables as random effects. The first control, *award duration*, accounted for the number of days between the end and start date of the award. The second control, *directorate*, was categorical and accounted for variability among awards from different NSF directorates, since areas of study may have inconsistent budgets and funding conventions. Third, the NSF has a dynamic budget (e.g., the fiscal year 2010 NSF budget was \$6.926 billion and the fiscal year 2017 NSF budget was \$7.472 billion), and since the amount of available funding is not consistent across years, this variability must be accounted for. Note, AIC was lower when *year* was added as a random effect instead of a fixed effect (AIC random = 41,927.73, AIC fixed = 42,071). Finally, a random effect for principal investigator controlled for the nonindependence of awards from the same lead researcher. The fixed and random effects explained a large proportion of the variance for money awarded by the NSF (R^2c = 0.70).

The number of abstracts in the dataset exceeds the minimum number of subjects-per-variable for overfitting concerns (see Austin & Steyerberg, 2015). With less than 20 variables in the a priori model, there are nearly 1,000 cases per variable. This is

Table 1. Bivariate Correlation Matrix of Primary Dependent and Independent Variables.

	Award amount	WC	WC^2	WPS	Common words	Analytic writing	Certainty	Tentativeness	Causal terms
Award amount	—								
WC	.150**	—							
WC^2	.133**	.974**	—						
WPS	.114**	.142**	.134**	—					
Common words	−.125**	−.099**	−.087**	.015*	—				
Analytic writing	.033**	−.004	.004	.144**	−.140**	—			
Certainty	.048**	.022**	.016*	−.076**	−.006	−.056**	—		
Tentativeness	−.114**	.017*	.015*	−.153**	.119**	−.198**	.150**	—	
Causal terms	.051**	.080**	.059**	−.093**	−.005	−.220**	.002	.093**	—

Note. WC = word count, WPS = words per sentence. Award amount = natural log-transformed amount of financial grant funding.
$^{*}p < .05$. $^{**}p < .01$. Calculations are two-tailed (N = 19,569).

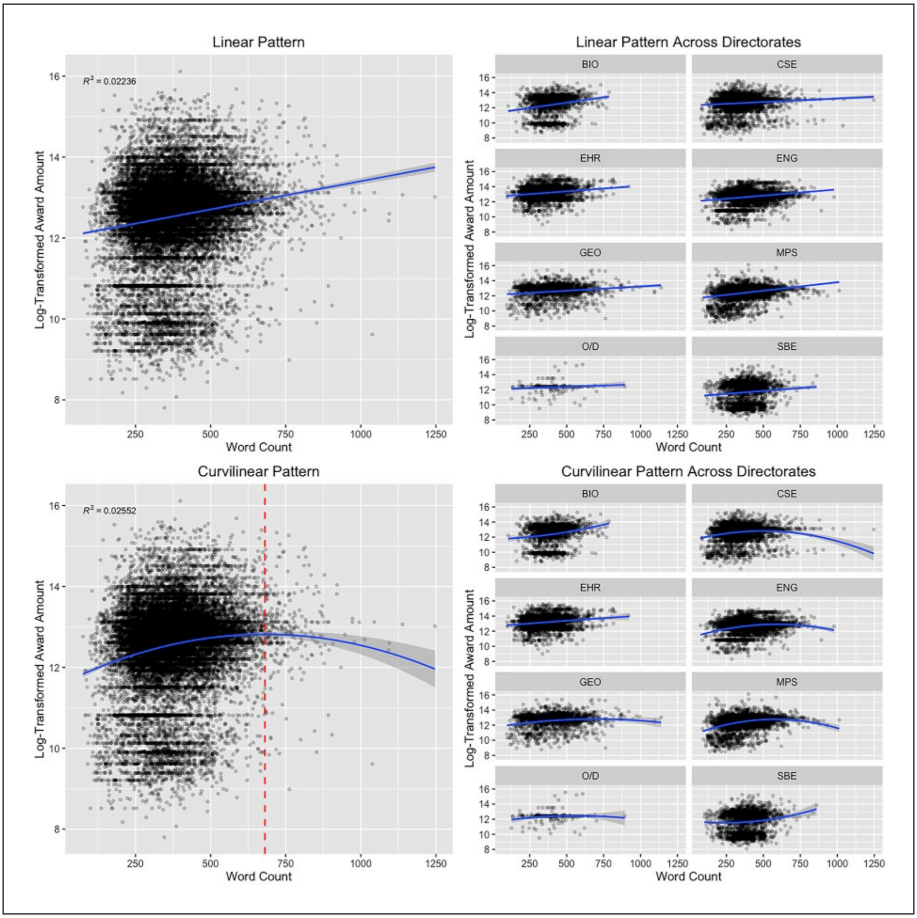

Figure 1. Scatterplots of word count and log-transformed award amount.
Note. BIO = Biological Sciences; CSE = Computer and Information Science and Engineering; EHR = Education and Human Resources; ENG = Engineering; GEO = Geosciences; MPS = Mathematical and Physical Sciences; O/D = Office of the Director; SBE = Social, Behavioral, and Economic Sciences. *Top left panel*: Scatterplot depicting a linear relationship between variables. *Top right panel*: Scatterplots depicting linear relationships across directorates. *Bottom left panel*: Scatterplot depicting a curvilinear relationship between variables. Vertical red line represents $x = 681$. *Bottom right panel*: Scatterplots depicting curvilinear relationships across directorates.

over 20 times the recommended number of cases per predictor based on conservative thresholds.

Discourse and Thinking Style Complexity

The top left panel of Figure 1 displays that on average, grant abstracts with more words receive more financial award funding ($\beta = .24$, $SE = .02$, $p < .001$). Specifically, a one-word increase in abstract length corresponds to a \$372[1] increase in the amount of money received after accounting for award duration, NSF directorate, year, and multiple awards from the same researcher. Despite potential reporting convention differences between directorates, the positive association between word count and the amount of financial award funding is largely consistent across areas (top right panel of Figure 1).

However, as the significant and negative word count squared coefficient indicates ($\beta = -.16$, $SE = .02$, $p < .001$), excessively long abstracts tend to receive less funding. The bottom left panel of Figure 1 displays the nature of the curvilinear relationship between word count and money awarded by the NSF, with an extremum at 681 words (680.61), derived from the following equation: $f(x) = 11.58 + 3.644e^{-3}(x) - 2.677e^{-6}(x^2)$. This pattern is reflected across all directorates except for Biological Sciences, Education and Human Resources, and Social, Behavioral, and Economic Sciences, which instead demonstrate a relatively consistent increase in financial NSF funding as word count increases (bottom right panel of Figure 1). Note, the overall curvilinear model explains nearly an additional 0.3% of variance relative to the overall linear model, and this ΔR^2 was statistically significant, $F(1, 19{,}566) = 63.40$, $p < .001$. This suggests the curvilinear model is indeed superior to the linear model in predicting money awarded by the NSF from grant abstract word count. These data provide partial support for H1a.

As the number of words per sentence increases, funding amount also increases ($\beta = .04$, $SE = 0.01$, $p < .001$), a finding consistent with H1b. Therefore, after accounting for the other predictors in the model, grants with more structurally complex abstracts tend to receive more financial award funding.

Does the complexity of the grant abstract content and thinking style of its writers relate to funding magnitude? Table 2 shows that using fewer common words in English associates with receiving more award money ($\beta = -.07$, $SE = .01$, $p < .001$). According to the data, there is little financial incentive to write according to the NSF's guideline for plain and simple language. Though inconsistent with H1c, these results are in line with the prior complexity findings, as the increased complexity of verbal structure (e.g., words per sentence) and content (e.g., fewer common words) associates with more money for grant writers.

A writing style that is similar to a narrative instead of an analytic report (e.g., contains more pronouns, conjunctions, adverbs, and negations relative to articles and prepositions) associates with more grant money ($\beta = -.02$, $SE = 0.01$, $p < .001$). Therefore, while the authors' writing structure and content should be complex, *how* they communicate the science should be dynamic as the negative relationship indicates. A personal touch, with more pronouns relative to articles or prepositions, is associated with receiving more money from the NSF. This finding is consistent with H1d and the NSF's plain writing objective, as a simpler narrative thinking style (e.g., with high rates of storytelling language such as pronouns and adverbs but low rates of complex categorical markers) corresponds to more NSF funding.

Table 2. Mixed Effect Regression Results Predicting NSF Award Funding From Language Patterns.

	Predictors	β	*SE*	*t*	*p*	95% CI
	Intercept	−.06	.15	−0.42	.68	[−0.352, 0.254]
Complexity	Word count	.24	.02	10.63	$<2.0 \times 10^{-16}$	[0.197, 0.286]
	Word count squared	−.16	.02	−7.21	5.97×10^{-13}	[−0.205, −0.117]
	Words per sentence	.04	.01	6.85	7.74×10^{-12}	[0.026, 0.047]
	Common words	−.07	.01	−11.21	$<2.0 \times 10^{-16}$	[−0.080, −0.056]
	Analytic writing	−.02	.01	−3.39	<.001	[−0.029, −0.008]
Confidence	Certainty	.02	.01	4.43	9.48×10^{-06}	[0.013, 0.033]
	Tentativeness	−.06	.01	−11.06	$<2.0 \times 10^{-16}$	[−0.070, −0.049]
	Causal terms	.06	.01	11.69	$<2.0 \times 10^{-16}$	[0.052, 0.073]

Note. NSF = National Science Foundation; AIC = Akaike information criterion. AIC = 41,927.73. *N* = 19,569. Dependent and independent variables were standardized. The model includes several controls, such as year, directorate, duration of the award, and a random effect for the principal investigator. The 95% confidence intervals (CIs) are based on the profile likelihood method, and Wald-type CIs produced consistent results.

Confidence in the Science Proposal

Grant abstracts written with more verbal certainty ($\beta = .02$, $SE = .01$, $p < .001$) and less tentativeness ($\beta = -.06$, $SE = .01$, $p < .001$) also correlate with more grant funding from the NSF. The exploratory causality dimension also received support as more causal language is associated with more award funding ($\beta = .06$, $SE = .01$, $p < .001$). These patterns provide evidence in support of H2a-c, though the effect sizes are small (Cohen's $f^2s < 0.014$).

Exploratory Robustness Checks

Data from both active and expired awards within a more diverse time period (1981-2009, to exclude cases from the current data set) evaluated how funding is associated with word patterns. Across 53,170 proposals covering over 13.6 million words, all reported language patterns from Table 2 were maintained with no substantive changes in significance levels.

To uncover additional relationships between language patterns and money awarded by the NSF, all standardized language features (excluding individual punctuation dimensions) from LIWC (plus word count squared) and the prior controls were entered into a mixed effects regression model. To reduce the chance of obtaining Type I errors from multiple exploratory tests, *p* values were corrected by multiplying each unadjusted significance level by the number of LIWC fixed effects (83) in the model (the Open Science Framework contains all exploratory results).

In total, 30 language dimensions were statistically significant including four features from Table 2, which operated in a consistent direction with the a priori model:

word count ($p < .001$), word count squared ($p < .001$), words per sentence ($p < .001$), and tentativeness ($p < .001$).

Other notable results suggest that a more upbeat verbal tone ($p = .002$), more verbal expressions of motivation and drive ($p < .001$), and more adjectives correlate with an increase in grant funding ($p < .001$). These data indicate that detailed and enthusiastic science writing can associate with more money awarded by the NSF. They pair well with the a priori model (Table 2), which suggested that more scientific certainty and less tentativeness were indicative of increased funding magnitude. However, psychological features such as the overall rate of affect (e.g., using words such as *excellent, inferior;* $p > .80$) were unrelated to funding amount. This suggests that after accounting for other language parameters, emotional appeals may not correlate with more grant funding, on average.

Discussion

The data from this study present a clear and consistent picture of how the writing patterns of grant abstracts are related to funding from the NSF. The structure, content, and writing style of NSF grant abstracts, plus the confidence communicated by grant writers, correlate with the amount of money awarded to researchers. A writing style that is structurally complex with fewer common words, but is written like a story and expresses more scientific certainty, correlates with receiving more money from the NSF. Less grant funding is associated with abstracts that are too long and contain a tentative style. The curvilinear representation of word count best describes the data and suggests that abstracts longer than 681 words, on average, receive less grant money. Together, the results support a well-established literature that suggests word patterns can reflect meaningful behavioral outcomes.

Why was complexity associated with greater funding support? Although speculative, grants with longer than average abstracts may have received more money because reviewers were convinced of the authors' ability to carry out the research and their expertise. Prior work suggests that increased word count and discourse complexity can reduce the uncertainty that a perceiver feels about the communicator (C. R. Berger & Calabrese, 1975; Flanagin, 2007). Therefore, providing more details about the proposal's science may encourage readers to believe that the writers are more dependable. Writing in a complex manner with more words and words per sentence provides only a structural cue of credibility to provide a positive first impression, however. The other language data in this study suggest that *how* people are communicating, particularly with style words (e.g., pronouns) and expressions of confidence (e.g., more certainty and causal language, less tentativeness), relates favorably to grant writers who want to receive more money from the NSF. These data are largely consistent with prior evidence (van den Besselaar et al. 2018), which has used language as a lens to understand stages of the grant review process and grant application outcomes.

Another possible explanation for the relationship between complexity and the amount of award funding is that complexity in science is normative and complex writing patterns match the reviewers' expectations for a successful proposal. Social norms research (Cialdini, 2006) suggests that norms are typically separated into perceptions about what people believe others should do (injunctive norms) and what people actually do (descriptive norms). The uptake of these different norm types can have important consequences for how people interpret and perceive a target (Lapinski & Rimal, 2005).

For example, in most cases, the target of complex and disfluent writing is appraised negatively (Oppenheimer, 2006). However, when people make judgments by comparing one target to another, a complex stimulus (e.g., jazz or classical music) can be rated positively over a relatively simple stimulus (e.g., pop music) when the task is appropriately challenging (Norman, 2010). Therefore, grants with structurally complex abstracts may receive more funding because complexity is a descriptive norm of science and is consistent with how people write science proposals. An injunctive norm of the NSF and other government agencies, on the other hand, encourages simple writing, which does not match how science is written in practice. Together, the NSF's call for plain writing may be an injunctive norm that is at odds with how science is actually communicated and then funded financially. An important area of future work may be to experimentally test the prior idea to better understand the connection between expectations, social norms, and financial support.

How do the present results also inform our understanding of grant writing and reviewing? A related analysis by Hume, Giladi, and Chung (2015) observed that funded grants often outperform unfunded grants in two ways: (1) the proposal's execution plan and (2) grantsmanship (e.g., the art and ability to obtain peer-reviewed funding; Kraicer, 1997). The authors found that unfunded grants were ill-prepared for submission, poorly written, or incomplete relative to funded grants (Hume et al., 2015). Such findings can extend to the current study and offer a post hoc explanation about why, among funded grants, some projects are awarded more money than others. Another aspect of grantsmanship may also be word patterns of complexity and confidence that signal the research's value to reviewers. Abstracts provide a window into the writing style of NSF proposals at large, but they also offer valuable first impression information about the authors, such as their credibility. Prior work suggests that language patterns can serve as credibility markers (Mitra, Wright, & Gilbert, 2017; Toma & D'Angelo, 2015) and also indicate the likelihood of follow-through with a written plan that has financial stakes (e.g., repayment of a loan; Larrimore et al., 2011). Therefore, this evaluation of NSF-funded grant abstracts offers that what and how authors write are meaningful to reviewers when considering award money. Stricter tests investigating the relationship between language patterns, awarded funding received from institutions, and credibility should be performed experimentally to further substantiate this idea.

Finally, it is important to highlight several advantages associated with the approach taken in this article. One of the principal benefits of using naturally occurring text data to indicate real-world outcomes is the improved external validity that is achieved over laboratory data. Laboratory studies often place subjects in situations that are unnatural or force them to respond in unnatural ways. For example, prior work has evaluated how people change their word patterns when lying or telling the truth in the lab (Burgoon & Qin, 2006; Hancock, Curry, Goorha, & Woodworth, 2008; among others), though participants in these studies often produce sanctioned lies that are not ecologically or externally valid (Levine, 2018). Analyzing data from online repositories allows for observations that are unaffected by constraints imposed by the research process. Another advantage stems from the idea that online language data are often accompanied with outcome variables that are high-stakes and meaningful. For example, in an analysis of U.S. presidential debates from 1976 to 2012, candidates who matched their opponents' language style on function words also performed better in the polls relative to those who failed to match their opponents (Romero, Swaab, Uzzi, & Galinsky, 2015). The current study observed meaningful associations between grant abstract writing and financial funding. These patterns indicate that writing style and word choice matter to understand how money is awarded in science.

Limitations and Future Research

There are several limitations of this study worth noting. First, the findings are correlational and do not imply a causal relationship between word patterns and NSF funding. Second, the writing data included only grant abstracts from funded proposals since full proposals are not public. Abstracts are often written after reviews and discussion about the award, though it is likely that the text reflects how researchers write in general. Future work should seek a large repository of funded and unfunded proposals to evaluate how awards contain distinct language profiles as well (e.g., see Hume et al., 2015; van den Besselaar et al., 2018). The language variables evaluated here also serve as proxies for social and psychological attributes. For example, words per sentence are often a measure of syntactic complexity because more words per sentence is typically associated with less readable text. However, as noted by an anonymous reviewer, complexity can also take the form of intellectual complexity, where the ideas of the writer are complicated but may be written efficiently for an expert audience (e.g., fewer words per sentence). Future work should take a more nuanced view of language patterns to understand their relationship to social and psychological processes that may also be multidimensional, such as complexity and confidence.

Finally, while the statistical tests report strong significance levels, the effect sizes are relatively small (Cohen's $f^2s < .014$). The effect sizes are largely consistent with other evaluations of language patterns taken from natural data sources, however (Kramer et al., 2014), and the ability to observe these relationships likely benefited from scale. Therefore, while one of the advantages of computational social science is the ability to discover new trends in data, the utility of the findings may be limited to the data source and the methods used to evaluate them. Future research should continue to apply computational approaches to big text data, with social science theory and empirical rationales, to guide predictions about how language patterns and psychological processes relate.

Conclusion

The amount of grant funding offered by the NSF is correlated with language patterns in grant proposal abstracts. This study used a database of publicly available grant abstracts to discover that grants with more structurally complex writing and more confidence tend to receive more money from the NSF. These results stand in partial contrast to proposal writing guidelines by the NSF, but they are supported by prior work that suggests positive financial outcomes can be associated with greater verbal output, more complex writing structures, and increased verbal certainty.

Acknowledgments

I would like to thank Jeff Hancock and Dan Cosley for feedback on this article and the two anonymous reviewers for their thoughtful improvements to the article.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Note

1. The specific funding amount, \$371.90, was calculated by regressing the raw award amount on word count and including random effects for award duration, directorates, year, and principal investigator.

References

Alvarez-Conrad, J., Zoellner, L. A., & Foa, E. B. (2001). Linguistic predictors of trauma pathology and physical health. *Applied Cognitive Psychology*, *15*, S159-S170.

Austin, P. C., & Steyerberg, E. W. (2015). The number of subjects per variable required in linear regression analyses. *Journal of Clinical Epidemiology*, *68*, 627-636.

Berger, C. R., & Calabrese, R. J. (1975). Some explorations in initial interaction and beyond: Toward a developmental theory of interpersonal communication. *Human Communication Research*, *1*, 99-112.

Berger, J., & Packard, G. (2018). Are atypical things more popular? *Psychological Science*, *29*, 1178-1184.

Boyd, R., & Pennebaker, J. W. (2015). A way with words: Using language for psychological science in the modern era. In C. Dimofte, C. Haugtvedt, & R. Yalch (Eds.), *Consumer psychology in a social media world* (pp. 222-236). New York, NY: Routledge.

Burgoon, J. K., & Qin, T. (2006). The dynamic nature of deceptive verbal communication. *Journal of Language and Social Psychology*, *25*, 76-96.

Chung, C., & Pennebaker, J. W. (2007). The psychological functions of function words. In K. Fiedler (Ed.), *Social communication* (pp. 343-359). New York, NY: Psychology Press.

Cialdini, R. B. (2006). *Influence: The psychology of persuasion*. New York, NY: Harper Business.

Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic markers of psychological change surrounding September 11, 2001. *Psychological Science*, *15*, 687-693.

Connor, U., & Mauranen, A. (1999). Linguistic analysis of grant proposals: European Union research grants. *English for Specific Purposes*, *18*, 47-62.

Connor, U., & Wagner, L. (1998). Language use in grant proposals by nonprofits: Spanish and English. *New Directions for Philanthropic Fundraising*, *1998*, 59–74.

Feng, H., & Shi, L. (2004). Genre analysis of research grant proposals. *LSP and Professional Communication*, *4*, 8-32.

Flanagin, A. J. (2007). Commercial markets as communication markets: Uncertainty reduction through mediated information exchange in online auctions. *New Media & Society*, *9*, 401-423.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*, 221-233.

Gaertig, C., & Simmons, J. P. (2018). Do people inherently dislike uncertain advice? *Psychological Science*, *29*, 504-520.

Golder, S. A., & Macy, M. W. (2011). Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, *333*, 1878-1881.

Hancock, J. T., Curry, L. E., Goorha, S., & Woodworth, M. (2008). On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, *45*, 1-23.

Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1993). Emotional contagion. *Current Directions in Psychological Science*, *2*, 96-99.

Hume, K. M., Giladi, A. M., & Chung, K. C. (2015). Factors impacting successfully competing for research funding: An analysis of applications submitted to the Plastic Surgery Foundation. *Plastic and Reconstructive Surgery*, *135*, 429e-435e.

Ireland, M. E., Slatcher, R. B., Eastwick, P. W., Scissors, L. E., Finkel, E. J., & Pennebaker, J. W. (2011). Language style matching predicts relationship initiation and stability. *Psychological Science*, *22*, 39-44.

Kacewicz, E., Pennebaker, J. W., Davis, M., Jeon, M., & Graesser, A. C. (2014). Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, *33*, 125-143.

Kraicer, J. (1997). *Art of grantsmanship*. Retrieved from http://www.hfsp.org/funding/art-grantsmanship

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*, 8788-8790.

Lapinski, M. K., & Rimal, R. N. (2005). An explication of social norms. *Communication Theory*, *15*, 127-147.

Larrimore, L., Jiang, L., Larrimore, J., Markowitz, D. M., & Gorski, S. (2011). Peer to peer lending: The relationship between language features, trustworthiness, and persuasion success. *Journal of Applied Communication Research*, *39*, 19-37.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., . . . Van Alstyne, M. (2009). Computational social science. *Science*, *323*, 721-723.

Levine, T. R. (2018). Ecological validity and deception detection research design. *Communication Methods and Measures*, *12*, 45-54.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics*, *45*, 221-247.

Marini, M. M., & Singer, B. (1988). Causality in the social sciences. *Sociological Methodology*, *18*, 347-409.

Markowitz, D. M. (2018). Academy Awards speeches reflect social status, cinematic roles, and winning expectations. *Journal of Language and Social Psychology*, *37*, 376-387.

Markowitz, D. M., & Hancock, J. T. (2016). Linguistic obfuscation in fraudulent science. *Journal of Language and Social Psychology*, *35*, 435-445.

Mitra, T., Wright, G. P., & Gilbert, E. (2017). A parsimonious language model of social media credibility across disparate events. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing—CSCW '17* (pp. 126-145). Portland, OR. Retrieved from http://comp.social.gatech.edu/papers/cscw17-cred-mitra.pdf

Mudambi, S. M., & Schuff, D. (2010). What makes a helpful online review? A study of customer reviews on Amazon.com. *MIS Quarterly*, *34*, 185-200.

Murphy, S. C. (2017). A hands-on guide to conducting psychological research on Twitter. *Social Psychological and Personality Science*, *8*, 396-412.

National Science Foundation. (2018). *Plain language*. Retrieved from https://www.nsf.gov/policies/nsf_plain_language.jsp

Norman, D. A. (2010). *Living with complexity*. Cambridge: MIT Press.

Oppenheimer, D. M. (2006). Consequences of erudite vernacular utilized irrespective of necessity: problems with using long words needlessly. *Applied Cognitive Psychology*, *20*, 139-156.

Pennebaker, J. W. (2011). *The secret life of pronouns: What our words say about us*. London, England: Bloomsbury Press.

Pennebaker, J. W., Booth, R. J., Boyd, R. L., & Francis, M. E. (2015). *Linguistic inquiry and word count: LIWC2015*. Austin, TX: Pennebaker Conglomerates.

Pennebaker, J. W., Chung, C. K., Frazee, J., Lavergne, G. M., & Beaver, D. I. (2014). When small words foretell academic success: The case of college admissions essays. *PLOS One*, *9*, e115844.

Plain Writing Act of 2010. 946, 146 Cong. Rec. 2861-2863 § Sec. 3. Definitions. (2010).

Price, P. C., & Stone, E. R. (2004). Intuitive evaluation of likelihood judgment producers: Evidence for a confidence heuristic. *Journal of Behavioral Decision Making*, *17*, 3957.

Robinson, M. D., Boyd, R. L., & Fetterman, A. K. (2014). An emotional signature of political ideology: Evidence from two linguistic content-coding studies. *Personality and Individual Differences*, *71*, 98-102.

Romero, D. M., Swaab, R. I., Uzzi, B., & Galinsky, A. D. (2015). Mimicry is presidential: Linguistic style matching in presidential debates and improved polling numbers. *Personality and Social Psychology Bulletin*, *41*, 1311-1319.

Stephens-Davidowitz, S. (2017). *Everybody lies: Big data, new data, and what the Internet can tell us about who we really are*. New York, NY: Dey Street Books.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*, 24-54.

Toma, C. L., & D'Angelo, J. D. (2015). Tell-tale words: Linguistic cues used to infer the expertise of online medical advice. *Journal of Language and Social Psychology*, *34*, 25-45.

van den Besselaar, P., Sandström, U., & Schiffbaenker, H. (2018). Studying grant decision-making: A linguistic analysis of review reports. *Scientometrics*, *117*, 313-329.

Author Biography

David M. Markowitz is an assistant professor in the School of Journalism and Communication at the University of Oregon. He uses language data from natural repositories to make inferences about what people are thinking, feeling, and experiencing psychologically. A large part of his research focuses on how deception affects language, such as how fraudulent scientists write their research papers compared to genuine scientists. His work has appeared in outlets such as the *Journal of Communication, Communication Research*, and *Journal of Language and Social Psychology*. He received his PhD from Stanford University and his undergraduate and master's degrees from Cornell University.