Psychology of Popular Media Culture
2014, Vol. 3, No. 3, 154–163


2160-4134/14/$12.00 http://dx.doi.org/10.1037/ppm0000036

# Blog Topic and Word Frequency: What Differentiates Between High and Low Powered Blogs?

Cynthia C. Johnson, Erin M. Buchanan, and Kayla N. Jordan
Missouri State University

Researchers and directories have difficulty classifying blogs into categories resulting from the rate at which blogs are created, as well as the overlap in content. Further, although blog popularity may be apparent in Internet traffic, the determinants of authority or status of a blog have yet to be explored. This article examined how word content affected blog status through several linguistic analyses by examining word usage across 24 high- and low-status informative blogs with differing topics: politics, technology, entertainment, and business. We compared the frequency of parts of speech, the frequency of unique words, and the use of low-frequency words and found blog dissimilarities that may indicate key distinctions in style and linguistic usage that differentiate both topic and status of the blog.

*Keywords:* psycholinguistics, blogs, word frequency



As of December 2011, BlogPulse.com identified >179 million Weblogs in existence, with nearly 100,000 created every 24 hr. Technorati.com (2013a), now the web's leading blog directory, acknowledges the difficulty of defining blogs, but they look for monthly original web content posted and a public Atom or RSS feed. Thus, a good generic description of a blog may be a Web site where a user or users post opinions and information regularly. However, Technorati's directory ranking system does not include social media sites such as Facebook and Twitter, which are often referred to as microblogs and limit users' posts to only a few sentences. The number of traditional blogs began to grow exponentially in the early nineties as accessible self-publishing platforms became popularized by diary bloggers (Allen, 2008). Because of this rapid growth, a variety of suggested blog classifications exist, such as by topic, purpose, author, or community. However, many difficulties are associated with classifying this medium owing to stylized differences. Qu, Pietra, and Poon (2006) have suggested that blogs be classified with a multilayered hierarchy, while others have proposed programs (Liu, Birnbaum, & Pardo, 2009) using pronouns or frequent words (Elgersma & de Rijke, 2008) and a semantic network based on Wikipedia (Ayyasamy, Alhashmi, Eu-Gene, & Tahayna, 2012) to help with classification schemes. Further, humans and directories cannot adequately keep up with the growing number of blogs (Qu et al., 2006), and findings suggest the most difficult to categorize blogs are news and political blogs, which both serve primarily to inform.

Despite the difficulty of categorizing blogs into just one classification, there are, as in most communication media, four accepted primary purposes: to inform, to persuade, to entertain, and, most unique to the blogging medium, to emotionally vent. Herring, Scheidt, Bonus, and Wright (2004) similarly categorized blogs by purpose rather than topic and through content analyses found the following similar categories: news and opinion, information sharing, and "vehicle[s] for self-expression and self-empowerment," (p. 1) with the latter accounting for the majority of the randomly selected blogs in their study. Technorati.com (2012a) analyzed a survey of 4,114 bloggers in 2011 and asked questions concerning who blogs (age, gender, location, socioeconomic status) and motivations for

This article was published Online First June 16, 2014.

Cynthia C. Johnson, Department of English, Missouri State University; Erin M. Buchanan and Kayla N. Jordan, Department of Psychology, Missouri State University.

We thank Scott Handley and Tracy Dalton for their guidance and support during the creation of this article.

Correspondence concerning this article should be addressed to Erin M. Buchanan, Department of Psychology, Missouri State University, 901 S National Ave, Springfield, MO 65897. E-mail: ErinBuchanan@MissouriState.edu

blogging. The bloggers described their writings as a way of sharing expertise, experiences, specific interests, and business marketing.

Additionally, classifying blogs interacted with characteristics of the blogger. Technorati.com (2012b) further reported that approximately three-fifths of bloggers were male, and although most bloggers were between the ages of 25 and 43 years, a third was over 44 years. Therefore, understanding blogger characteristics impacts the ability to classify blogs by content. Fullwood, Sheehan, and Nicholls's (2009) analysis of 120 public domain MySpace blogs showed that bloggers between the ages of 18 and 29 years were more likely to use blogs to gather information and use semiformal language, the age-group of 30–49 years was more likely to have diary blogs, and the 50+ years age-group was most likely to use blogs as an emotional outlet and use negative tones. Furthermore, although this study found no gender differences in purpose, tone, or symbol use, Haferkamp and Krämer (2008) conducted a survey with 79 bloggers focused on motivations for blogging. They found that male bloggers tend to present more informative material, and female bloggers are more interested in writing about personal experiences.

Yet, differences in blogs expand even beyond the demographics of the authors and include characteristics of the community. The blogging medium overall has developed many unique attributes. For example, in its current state, the "Giant blogging terms glossary" (2006) on the blog *Quick Online Tips* contains more than 150 terms and states that it is constantly updated. It then, in the true nature of a blog, invites readers to propose additional definitions and revisions in their comments. Beyond this, however, blogging communities develop norms separate from the medium as a whole. For example, Wei (2004) compared the written and actual norms within a blogging community devoted to knitting, which included a sample of 33 blogs, and discussed the tendencies for blogs of similar topics to form around each other in specific communities. Within these communities, writing norms and guidelines begin to form according to the desires of the bloggers and their readers. Wei then found that as blogs within a community become more diverse, they break off into specialty groups, as was shown by the formation of the communities *Men Who Knit*, *Knitting Kitty*, and *QueerKnit*. Similar analyses have been performed on goth community blogs (Hodkinson, 2004), and community blogs have been described as virtual cities (Efimova & Hendrick, 2005). As can be gathered from these studies, writing in blogs is greatly affected by the readers themselves. Thus, an attribute of an effective blog, especially within specified communities, is careful audience analysis, as reader's perceptions can change the activity of a blog (Baumer, Sueyoshi, & Tomlinson, 2008; Markel, 2010). Therefore, writing in blogs may largely differ based on the bloggers' effectiveness at communicating to their desired audiences and adhering to community expectations; however, it remains to be seen whether blogs with similar purposes and high popularity differ in word use based primarily on their differing topics.

Thus, the current study analyzed basic word use across 24 blogs with the purpose to inform: political, technological, entertainment, and business. We selected blogs that were considered popular (by the number of other Web sites linking to their pieces) and blogs lower in popularity to examine differences in blog status. Although machine models are used to classify blogs, we sought to explore blogging from a more traditional psycholinguistic perspective through word frequency and usage—a piece that is missing from the literature on blogging. First, we examined the frequency of different parts of speech among these blogs to determine if a part of speech is used more or less often for a specific blog topic by blog status. For instance, did the technology blogs have a greatly different vocabulary base than the political blogs? Then, we investigated whether there was a pattern of specific word usage that matches a particular blog type by status. Finally, this investigation determined whether certain blog types use low-frequency words more often than others. Specific predictions are described below.

## Hypotheses

### Hypothesis 1

Though the primary purpose of all four blog types is to inform, different parts of speech will appear more commonly in each one based on its topic. We will use an independence chi-square

test to analyze this hypothesis on each part of speech type by blog status. This analysis will allow us to examine how blog status affects word usage: if chi-square values are significant, standardized residuals will indicate how high- and low-status blogs differ in their usage. For example, the high-status entertainment blog may be more likely to have more names than others because it focuses closely on people, productions, and events, and likewise, the technology blog will focus more on specific products, Web sites, and companies. The political and business blogs, however, will likely have more verbs because, though they focus on people and companies, they place more emphasis on their actions.

### Hypothesis 2

Each blog will also have its own vocabulary of words particular to its topic because they are informing different reader bases. Therefore, we will examine the amount of unique words for each blog type by first controlling for overlapping words with an independence chi-square test.

### Hypothesis 3

Some blogs will use infrequent (low frequency) words more often than others based on the topic, and therefore, the audience. For example, the technology blog is aimed at an audience with a certain level of technological knowledge and, therefore, can use more rare, elevated, technical words than an entertainment blog will use. A mixed-model ANOVA will be used to assess word frequency differences across blogs and status.

## Method

### Materials

Four blog topics with the primary purpose to inform were selected from the Technorati.com (2012c) Blog Directory: politics, technology, business, and entertainment. From each category, three top-ranked blogs and three lower ranked blogs were chosen on June 3, 2013, and November 3, 2013. A blog's rank is determined by its "authority," which is calculated by its linking behavior and, in turn, its influence within its topic. The highest authority a blog can be rated is 1,000, and 0 is the lowest (Technorati.com, 2013b). An effort was made to match high and lower status blogs on face validity topic (i.e., blog posts appeared to cover the same topics, as many blogs cover several themes) with approximately the same authority for each lower status blog. Please note that both ranking and authority are dynamic, changing as post linkage changes from day to day and included numbers are based on time of data collection. Table 1 includes all blogs for each status and topic combination, along with their Technorati ranking and authority. For political blogs, the *Huffington Post Politics* were selected using the specific politics page from *Huffington Post* (i.e., *http://www.huffingtonpost*

Table 1
*Blogs Used for Data Collection Arranged by Topic and Status*

| Status | Entertainment | Politics | Business | Technology |
|---|---|---|---|---|
| High | Deadline Hollywood (1, 952) | Huffington Post Politics (1, 844) | Zero Hedge (1, 933) | The Verge (1, 902) |
| | Hollywood Life (3, 896) | CNN Political Ticker (2, 931) | Money, Life, and More (1, 902) | Engadget (1, 898) |
| | Bleeding Cool (5, 890) | Daily Kos (6, 851) | Felix Salmon (2, 931) | Tech Crunch (2, 889) |
| Low | The Edge (Boston Herald Entertainment) (510, 519) | Jeff Weintraub (507, 428) | The Big Picture (312, 570) | A Bugged Life (373, 459) |
| | Gossip Juice (519, 521) | Jon Caldara (561, 461) | Financial Sense (407, 480) | Computer World (393, 499) |
| | Heckler Spray (546, 498) | Washington Wire (562, 444) | Tim Harford (519, 479) | Techware Labs (402, 453) |

*Note*. Technorati information is listed as (ranking, authority).

*.com/politics/*). The rest of the blog posts were taken from the blog homepage.

Posts present on the front pages of the blogs from no earlier than 2011 were then copied for analyses, excluding the comments. Because comments tend to take a more informal conversational tone, their inclusion may have considerably altered the results. Posts were copied until the unedited word count was >50,000 words, which would equal approximately 50,000 words after single letters and uncategorizable words were removed (see below). A list of blog word frequencies was then developed by counting the number of times each word appeared in each blog using PHP code developed by the corresponding author, which has been used for other linguistic analyses involving frequency counts (Buchanan, Holmes, Teasley, & Hutchison, 2013).

### Data Processing Procedure

The blog entries were spell checked, and several misspelled words were corrected. After word frequencies were recorded, each word was reviewed, and those appearing in several forms were combined. For example, "walk," "walked," "walking," and "walks" were recorded as four instances of only one word as long as their most common part of speech was the same (i.e., all verbs), as were "fast," "faster," and "fastest." Furthermore, uncategorizable word/letter combinations were removed (words with no Google definitions or hits), and single letters other than "a" and "I" were removed. Acronyms were kept and treated as single words with a special coding for acronym. The remaining words were then categorized according to their most common parts of speech using database norms (The English Lexicon Project, Balota et al., 2007) and the Google define feature for nonoverlapping words. Specific names of companies, celebrities, and so forth were coded as special noun types described as names below. The first and third author coded words not present in the English Lexicon Project (usually names and acronyms), and unclear items were discussed among authors. The second author combined words with the same parts of speech that was used for hypothesis 1's analysis. Furthermore, individual words were compared across the categories to determine what concepts were unique to each type of blog for hypothesis 2. Finally, the frequency of individual words was analyzed using the norms from the Hyperspace Analogue to Language (HAL; Burgess & Lund, 1997) for hypothesis 3. Other norms, such as SUBTLEX (Brysbaert & New, 2009) and Kucera and Francis' Brown Corpus (1967) were considered for frequency analysis, and the HAL norms were selected, as they included the most overlap with our dataset.

## Results

### Data Statistics

After data processing, 30,342 unique word-to-part of speech combinations were present across all four blogs, which totaled 1,265,659 words to examine. Table 2 includes the total number of words for each blog, as well as the word breakdown by part of speech (with percentages of words calculated by type). The majority of words found were the expected nouns, verbs, adjectives, and preposition words. Many names and acronyms were used, which was predicted for certain blog types because of their expected content (i.e., the entertainment blogs). As indicated in Table 2, total frequencies were approximately 50,000 words per blog (therefore, 150,000 across all three per cell), with a low of 154,157 and high of 162,433. This difference was due to the removal of spelled out hyperlinks (i.e., *www.example.com* was removed), uncategorizeable letter combinations, or letter–number combinations.

### Hypothesis 1

For this hypothesis, we analyzed acronyms, adjectives, adverbs, names, nouns, pronouns, and verbs individually in a 2 (status: high, low) by 4 (blog type: technology, entertainment, politics, business) independence chi-square test to assess if word count differed from expected word counts. Interjection word counts were not analyzed owing to small sample size. Conjunction and preposition word counts were considered filler words and not of interest for this analysis. Chi-square analyses values and effect sizes are listed in Table 3. A Bonferroni correction was used to correct for Type I error, using $\chi^2(3)_{p < .001} = 16.27$ as our critical value. All analyses on parts of speech showed an interac-

Table 2
*Part of Speech Percentages by Blog Type and Status*

| Word type | Technology | Entertainment | Business | Politics | Total |
|---|---|---|---|---|---|
| High status | | | | | |
| Acronym | 0.68 | 0.58 | 0.44 | 0.30 | 0.50 |
| Adjective | 17.58 | 16.50 | 17.66 | 17.83 | 17.40 |
| Adverb | 7.16 | 7.24 | 8.12 | 5.92 | 7.12 |
| Conjunction | 4.14 | 4.08 | 4.17 | 3.52 | 3.98 |
| Interjection | 0.03 | 0.08 | 0.04 | 0.04 | 0.05 |
| Name | 4.30 | 6.20 | 2.08 | 3.86 | 4.09 |
| Noun | 29.47 | 27.56 | 28.86 | 31.34 | 29.31 |
| Preposition | 12.81 | 11.99 | 12.37 | 13.22 | 12.60 |
| Pronoun | 5.70 | 7.51 | 7.29 | 6.42 | 6.73 |
| Verb | 18.12 | 18.27 | 18.95 | 17.55 | 18.22 |
| Total | 159,313 | 158,175 | 162,433 | 159,751 | 639,672 |
| Low status | | | | | |
| Acronym | 1.26 | 0.33 | 0.44 | 0.34 | 0.59 |
| Adjective | 17.73 | 17.38 | 19.69 | 19.05 | 18.46 |
| Adverb | 6.53 | 7.11 | 6.88 | 6.32 | 6.71 |
| Conjunction | 3.96 | 3.98 | 4.05 | 3.95 | 3.99 |
| Interjection | 0.07 | 0.10 | 0.05 | 0.04 | 0.06 |
| Name | 3.40 | 4.48 | 1.90 | 3.10 | 3.22 |
| Noun | 30.11 | 28.34 | 30.06 | 30.80 | 29.82 |
| Preposition | 12.79 | 12.05 | 12.64 | 13.44 | 12.73 |
| Pronoun | 5.99 | 7.90 | 5.92 | 6.08 | 6.48 |
| Verb | 18.16 | 18.34 | 18.37 | 16.87 | 17.94 |
| Total | 154,157 | 159,623 | 156,977 | 155,230 | 625,987 |

*Note.* Percentages are calculated by part of speech (i.e., down each column).

tion between blog status and type. Therefore, we examined each cell's standardized residuals to show which blogs were contributing to the goodness of fit difference. Table 3 contains a visual of the results from the analysis of standardized residuals. Positive residuals over two are indicated by a ▲ symbol that signifies a larger number of words found than expected, while ▼ indicates a smaller number of words found than expected (negative residuals over −2). The equal sign shows cell frequencies that were equal to expected values.

The high-status technology blogs were more likely to use adjectives and adverbs, whereas the low-status blogs were more likely to use acronyms and pronouns. This finding may indi-

Table 3
*Chi-Square Analyses of Parts of Speech Across Blogs*

| Blog type | Blog status | Acronyms | Adjectives | Adverbs | Names | Nouns | Pronouns | Verbs |
|---|---|---|---|---|---|---|---|---|
| Technology | High | ▼ | ▲ | ▲ | = | = | ▼ | = |
| | Low | ▲ | ▼ | ▼ | = | = | ▲ | = |
| Entertainment | High | ▲ | = | ▼ | ▲ | ▼ | ▼ | ▼ |
| | Low | ▼ | = | ▲ | ▼ | ▲ | ▲ | ▲ |
| Business | High | ▼ | ▼ | ▲ | ▼ | = | ▲ | ▲ |
| | Low | ▲ | ▲ | ▼ | ▲ | = | ▼ | ▼ |
| Politics | High | = | = | ▼ | = | ▲ | = | ▲ |
| | Low | = | = | ▲ | = | ▼ | = | ▼ |
| | $\chi^2(3)$ | 326.07 | 81.17 | 191.77 | 41.60 | 84.49 | 291.67 | 61.78 |
| | Cramer's V | 0.22 | 0.02 | 0.05 | 0.03 | 0.02 | 0.06 | 0.02 |

*Note.* ▲ Indicates a larger number of words than expected, ▼ indicates a smaller number of words than expected, and = indicates an expected number of words. All chi-square values are significantly greater than a Bonferroni-corrected critical value.

cate that the high-status blogs were focused on product descriptions, whereas the low-status blogs had a more informal approach of discussing nonbrand name descriptions (i.e., CPU, HD, USB; acronyms) of products for the user (pronouns). However, contrary to hypotheses, there was not a significant residual for names, showing that both status blogs used company or personal names about what was expected given other blog topics. As predicted, the high-status entertainment blogs used more names, as well as more acronyms, which was predominantly the usage of slang terms (OMG), place abbreviations (U.K., LA), and common terminology (DVD, TV). In contrast, low-status entertainment blogs used more adverbs, nouns, pronouns, and verbs than expected when compared with other blog topics and high-status blogs. This finding may be tied to the low-status blogs' tendency to be more of a celebrity gossip site while the high-status blogs more often reported information on upcoming movies/TV shows and award ceremonies.

The high-status business blogs showed more adverbs, pronouns, and verbs, whereas the low-status blogs used more adjectives and names. Potentially, this difference could be based on the high-status blogs' tendency to write from a first person, informal style discussing how a reader might save money or pay down debts, whereas the low-status blogs are more formal and descriptive of market trends. Finally, the political blogs showed only a few differences; mainly, that the high-status blogs were more likely to use nouns and verbs, and the low-status blogs were more likely to use adverbs. The hypothesis for verbs was supported with this analysis showing that high-status blogs of both types used more verbs. This finding may indicate that high-status political blogs, although equally likely to use names of people/companies as low-status blogs, were more likely to focus on actions.

## Hypothesis 2

The words collected were then condensed to only unique words. These words were items that appeared on only one blog, although they could appear multiple times on that particular blog. For example, the word "repped" appeared only on the entertainment high-status blogs, but was listed 45 times across our sample, while "SEGA" appeared only on the technology high-status blogs 17 times. Therefore, 14,307 words were examined with a total frequency of 27,634 mentions. The purpose of this analysis was to assess blog language, as each blog should have separate word uses due to their differing topics. Results are displayed graphically in Table 4, as with the last hypothesis. An analysis of standardized residuals indicated only entertainment high-status blogs were more likely to use unique words ($p < .001$) when compared across status and topic, while business high-status blogs were actually likely to use less unique words. However, this word list contained many ($N = 5{,}152$) words that were specific names or acronyms, which might have biased the analyses. We then removed these words and reanalyzed the number of unique words by blog, which was still significant ($p < .001$), but only the business effect remained wherein low-status business blogs were more likely to use unique words.

## Hypothesis 3

This hypothesis evaluated the normed word frequency of the words each blog used. First, logHAL values were matched to our collected word set. Words without normed frequencies were excluded, leaving 32,768 for high-status blogs (8,519 business, 8,729 entertainment, 8,130 politics, 7,390 technology) and 34,757 for low-status blogs (8,516 business, 10,423 entertainment, 8,650 politics, 7,168 technology). We then weighted each word frequency by the number of times it was mentioned in each blog (i.e., logHAL * word count for blog). Larger values in this analysis would indicate more use of more

Table 4
*Chi-Square Analyses for Unique Words Across Blogs*

| Blog type | Blog name | With names | Without names |
|---|---|---|---|
| Technology | High | = | = |
| | Low | = | = |
| Entertainment | High | ▲ | = |
| | Low | ▼ | = |
| Business | High | ▼ | ▼ |
| | Low | ▲ | ▲ |
| Politics | High | = | = |
| | Low | = | = |
| | $\chi^2(3)$ | 176.72 | 19.50 |
| | Cramer's V | 0.08 | 0.04 |

frequent words, while smaller values would indicate smaller use of more frequent words and use of more low-frequency words. A 2 (status: high, low) by 4 (blog type: technology, business, politics, entertainment) ANOVA showed no significant differences in weighted word frequency for status ($F < 1$, $p = .42$), type, $F(3, 67{,}517) = 1.55$, $p = .20$, or the interaction ($F < 1$, $p = .95$). However, using normed frequencies eliminated many individualized words that are probably infrequent or newer terms. Therefore, we tested only the words used in all eight blogs to examine differences in word frequency and usage. This analysis included 2,531 words that were common to all blogs. As seen in Figure 1, the weighted word frequency varied across blog types; however, this difference was not significant, $F(3, 7590) = 1.95$, $p = .12$, $n_p^2 < .01$. However, blog status was significantly different overall, $F(1, 7590) = 8.49$, $p = .004$, $n_p^2 < .01$. High-status blogs ($M = 698.00$, $SE = 88.45$) used more frequent words than the low-status blogs ($M = 678.41$, $SE = 90.21$). A marginal interaction effect was found between blog type and status, $F(3, 7590) = 2.40$, $p = .066$, $n_p^2 < .01$. Planned comparisons were used to compare high and low status blogs for each blog type as shown in Table 5. Only technology and business blogs showed significant differences, while politics showed a marginal trend in the same direction. For these high-status blogs, more frequent words were used than their low-status counterparts, which, coupled with earlier results, could indicate a less formal tone to connect with a wider audience.

## Discussion

### Hypothesis 1

We hypothesized different parts of speech will appear more commonly in each type of blog based on its topic and audience. Specifically, we expected the entertainment and technology blogs to contain more names because they focus on people and products, and we expected the political and business blogs to contain more verbs because they focus on events, with support for these hypotheses focusing on high-status blogs (minus technology name support). While, more than 1.2 million words were used for our corpus, future data analyses could examine this finding over time because, as mentioned earlier, blog status rankings change based on other sources citing their posts. Furthermore, we did not compare our results for specific types of informative blogs to a more overarching type, such as news. It is likely these results would have echoed those found by Qu, Pietra, and Poon (2006) because, as they explain, news blogs often contain a variety of topics. Additionally,

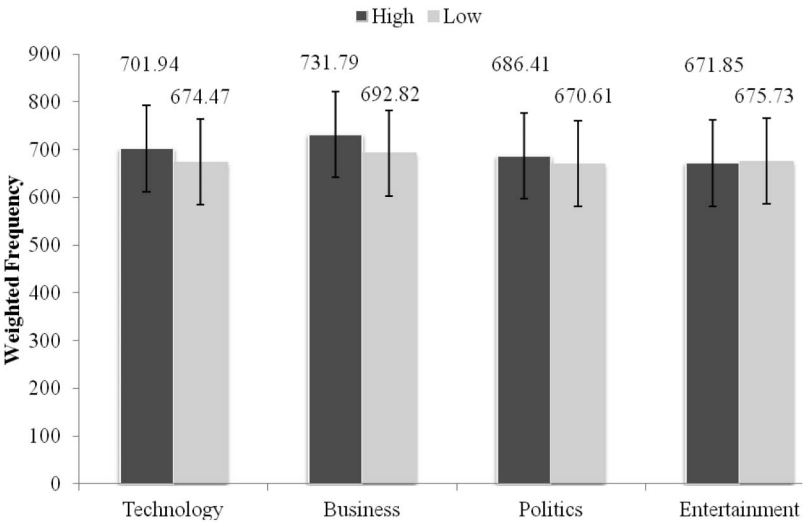


*Figure 1.* Weighted word frequency by blog. Error bars represent standard error. Means are listed above each column.

Table 5
*Post Hoc Examination of Hypothesis 3 Differences in Blog Weighted Word Frequency*

| Comparison | *M* difference | *t* | *p* | *d* |
|---|---|---|---|---|
| Technology | | | | |
| High–Low | 27.47 | 2.19 | 0.03 | 0.04 |
| Entertainment | | | | |
| High–Low | −3.88 | −0.47 | 0.65 | −0.01 |
| Business High–Low | 38.97 | 2.29 | 0.02 | 0.05 |
| Politics High–Low | 15.80 | 1.77 | 0.08 | 0.04 |

*Note.* *df* for all tests = 2,530. Effect size values are Cohen's *d* for dependent *t* tests using standard deviation of the differences as the denominator.

because blogs are not standardized, any blog may consist of multiple topics. That said, since Technorati.com (2012c) listed each of the examined blogs as the most popular within their category, their styles may be used as models for effective writing within each subject, though other influencing factors on popularity must be considered.

**Entertainment blogs.** The high-status blogs did in fact contain significantly more names than their comparison blogs. However, because our data were collected across several award seasons, the blogs may have listed more names than usual, albeit both types of blogs would be expected to list names and winners. Furthermore, despite containing more names than expected, the blogs also contained more acronyms, with fewer nouns, adverbs, pronouns, and verbs. These results imply that, as we hypothesized, the blogs have a high focus on people (names), and potentially this focus leads to presentation of basic factual information of upcoming entertainment, rather than the low status focus on what the celebrities have been doing (verbs) and descriptions (adverbs).

**Technology blogs.** Our high-status technology blogs did not support the name hypothesis, but instead showed that high-status blogs are more likely to use adjectives and adverbs, while low-status blogs are more likely to used acronyms and pronouns. The high-status blogs may be focused on presenting product descriptions and performance tests (adjectives/adverbs), but it is not necessarily as likely to use more common abbreviations like CPU (acronyms). The high-status blogs may expect a more informed reader, which allows them to discuss products and brands in a different way than the lesser known blogs.

**Business blogs.** The high-status business blogs did contain more verbs than the low-status blogs, as well as more pronouns and adverbs. The blogs appear to discuss more how readers (pronouns) could use suggested tips (verbs) to improve finances, while the low-status blogs spent more time describing (adjectives) various stocks and financial people (names).

**Political blogs.** Finally, the high-status political blogs showed few differences across parts of speech, but supported our hypothesis that high-status blogs would use more verbs. The high-status blogs were more likely to use nouns as well, and the low-status blogs were more likely to use adverbs. As described earlier, both high- and low-status blogs used the names of political figures/companies equally, but these slight differences in parts of speech may show differences in style. The high-status blogs focus on actions (verbs) and events (nouns), while the low status blogs may only be describing the actions (adverbs).

## Hypothesis 2

We further hypothesized that each blog has its own vocabulary of unique words particular to its topic. We examined words that only appeared in one of the blogs. Then, we assessed the same data a second time excluding all names and acronyms, which are overwhelmingly unique. Generally, high- and low-status blogs used the same amount of unique words, except for the high-status business blogs that used less specialty words. This finding was unexpected because of Wei's (2004) previous examination of blogging norms wherein bloggers use communities to access specific information. The high-status entertainment blogs were found to use more unique words, but this effect was primarily driven by the use of names and acronyms, since this effect disappeared after removing those items. The high status business blogs used more common words to connect with the reader. This finding was echoed in Hypothesis 1, as we found that the high status blogs appeared to take a more informal, helpful tone, which would be more attractive to readers with a vocabulary they were already familiar with.

## Hypothesis 3

Our third hypothesis stated that, based largely on differing audiences, some blogs will use either low- or high-frequency words more often than others. After eliminating extremely infrequent or new terms by testing only the 2,531 words common to all four topics and statuses of

blogs, we found that-high status blogs use more frequent words more often than the low-status blogs. This style may make them more attractive to a common reader by using nonspecific terminology. Technology and business blogs specifically showed this difference between high- and low-status blogs, while political blogs showed a marginal effect and no effect for entertainment blogs. It appears that the high-status business blogs not only used less unique words but also adopted a style that used more frequent words to help the reader wade through information. We might point to audience (wide range political news) as the likely reason political blogs would show these effects. However, technology blogs would be expected to use more infrequent words because of the specific knowledge required for the intended topic. In this case, many of those infrequent words were not normed across available frequency databases, and we suspect there may be different findings with updated frequency norms to include technology words.

## Limitations

While this data collection was large and varied, there are many other blogs within each of these topics to consider. Blogs were selected with consideration of targeting similar audiences with their posts and discussions, but many blogs often have overlapping themes. For example, a technology blog may also discuss business practices of the companies that they profile (i.e., Apple). Lastly, the Internet blogosphere is clearly dynamic, and the language they use may change over time. The high-impact blogs are likely to continue to post articles with the same word usage because their past articles are clearly drawing readers, while the lower impact blogs may slowly change their style to attract more readers.

## Future Research

This article presented an exploration in using traditional quantitative psycholinguistic analyses on the understanding of topic and status for different types of blogs. Future research may expand on several fronts. Though Twitter (microblogging) audiences have been analyzed (Marwick & Boyd, 2010), more research could be conducted to understand blog audiences' expectations and specialized knowledge within specific blogging communities. Further research could examine if these word-frequency patterns could also be predictors of future blog effectiveness or success. Potentially, these characteristics can be used to aid classification, and language could be used to distinguish informative blogs of narrow topics, such as politics, from an overarching topic, such as news. As Qu et al. (2006) discovered, it is most difficult to classify blogs of similar purposes or overlapping topics; however, our research uncovered differences in word use that suggest characteristics that may be useful to this goal. Practically, blogs may be able to use this information to help shape the style of posts to attract readers, similar to Wei's (2004) research on blog community norms. For example, for certain topics, it appeared to be beneficiary to use more frequent words so readers would not get lost in the material, but it was also important for high-status blogs to use reader-specific knowledge, such as political names or technology products. Therefore, a blog may gain status by posting updates that are both accessible and on topic for the expected material for that blog.

## References

Allen, R. (2008). Overview: The impact of blogging. In S. Engdahl (Ed.), *Blogs* (pp. 114–117). New York, NY: Greenhaven.

Ayyasamy, R., Alhashmi, S., Eu-Gene, S., & Tahayna, B. (2012). Enhancing automatic blog classification using concept-category vectorization. *Knowledge Engineering and Management,* 487–497.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., . . . Treiman, R. (2007). The English Lexicon project. *Behavior Research Methods, 39,* 445–459. doi:10.3758/BF03193014

Baumer, E., Sueyoshi, M., & Tomlinson, B. (2008). Exploring the role of the reader in the activity of blogging. In *CHI '08: Proceeding of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems (2008),* pp. 1111–1120, doi:10.1145/1357054.1357228

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41,* 977–990. doi:10.3758/BRM.41.4.977

Buchanan, E. M., Holmes, J. L., Teasley, M. L., & Hutchison, K. A. (2013). English semantic word-pair norms and a searchable Web portal for experimental stimulus creation. *Behavior Research Methods, 45,* 746–757, doi:10.3758/s13428-012-0284-z

Burgess, C., & Lund, K. (1997). Modeling parsing constraints with high-dimensional context space. *Language and Cognitive Processes, 12,* 177–210. doi:10.1080/016909697386844

Efimova, L., & Hendrick, S. (2005). In search for a virtual settlement: An exploration of weblog community boundaries. In P. van den Besselaar, G. De Michelis, J. Preece, & C. Simone (Eds.), *Communities and Technologies* (Vol. 5). Proceedings of the Second Communities and Technologies Conference. Milano: Springer.

Elgersma, E., & de Rijke, M. (2008). Personal vs non-personal blogs: Initial classification experiments. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 723–724). Singapore: ACM.

Fullwood, C., Sheehan, N., & Nicholls, W. (2009). Blog function revisited: A content analysis of MySpace blogs. *CyberPsychology and Behavior, 12,* 685–689. doi:10.1089/cpb.2009.0138

Giant blogging terms glossary: Need a blog dictionary? (2006, May 6). Retrieved from http://www.quickonlinetips.com/archives/2006/06/the-giant-blogging-terms-glossary/. Retrieved May 1, 2012.

Haferkamp, N., & Krämer, N. (2008). Entering the blogosphere: Motives for reading, writing, and commenting. *Conference Papers—International Communication Association* (2008 Annual Meeting), pp. 1–33.

Herring, S. C., Scheidt, L. A., Bonus, S., & Wright, E. (2004). Bridging the gap: A genre analysis of Weblogs. In *Proceedings of the 37th Hawai'i International Conference on System Sciences (HICSS-37)* (pp. 1–11). Los Alamitos, CA: IEEE Computer Society Press.

Hodkinson, P. (2004). Subcultural blogging? Individual, community and communication. In *Association of Internet Researchers Annual Conference: IR 5.0.* Sussex, England: Ubiquity, September 2004.

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Liu, J., Birnbaum, L., & Pardo, B. (2009). *Spectrum: Retrieving different points of view from the blogosphere*. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM*; Vol. 118). San Jose, CA.

Markel, M. (2010). *Technical communication*. (9th ed.). Boston, MA: Beford/St. Martin's Press.

Marwick, A. E., & Boyd, D. (2010). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media and Society, 14,* 299–315. doi:10.1177/1461444810365313

Qu, H., Pietra, A. L., & Poon, S. (2006). Automated blog classification: Challenges and pitfalls. In N. Nicolov, F. Salvetti, M. Liberman, & J. H. Martin (Eds.), *Computational approaches to analyzing Weblogs: Papers from the 2006 Spring Symposium* (pp. 184–186) (Technical Report No. SS-06–03). Menlo Park, CA: AAAI Press.

Technorati.com. (2012a, November 4). State of the blogosphere 2011: Pt. 2. Retrieved from http://technorati.com/social-media/article/state-of-the-blogosphere-2011-part2/. Retrieved May 2, 2012.

Technorati.com. (2012b, November 4). State of the blogosphere 2011: Pt. 1. Retrieved from http://technorati.com/social-media/article/state-of-the-blogosphere-2011-part1/. Retrieved May 2, 2012.

Technorati.com. (2012c). Blog directory. Retrieved from http://technorati.com/blogs/directory/. Retrieved February 4, 2012.

Technorati.com. (2013a). Blog quality guidelines. Retrieved from http://technorati.com/blog-quality-guidelines-faq/. Retrieved August 12, 2013.

Technorati.com. (2013b). Technorati authority FAQ. Retrieved from http://technorati.com/what-is-technorati-authority/. Retrieved August 7, 2013.

Wei, C. (2004). Formation of norms in a blog community. In L. Gurak, S. Antonijevic, L. Johnson, C. Ratliff, & J. Reyman (Eds.), *Into the blogosphere; rhetoric, community and culture of Weblogs*. Retrieved March 15, 2012, from http://blog.lib.umn.edu/blogosphere/formation_of_norms.html

Received March 29, 2013
Revision received March 1, 2014
Accepted March 4, 2014 ■