SPECIAL SECTION: 2017* Human Factors *Prize Winner

Sentinels of Breach: Lexical Choice as a Measure of Urgency in Social Media

Andrew J. Hampton and **Valerie L. Shalin**, Wright State University, Dayton, Ohio

Objective: This paper identifies general properties of language style in social media to help identify areas of need in disasters.

Background: In the search for metrics of need in social media data, much of the existing literature ignores processes of language usage. Psychological concepts, such as narrative breach, Gricean maxims, and lexical marking in cognition, may assist the recovery of disaster-relevant metrics from altered patterns of word prevalence.

Method: We analyzed several hundred thousand location-specific microblogs from Twitter for Hurricane Sandy, Oklahoma tornadoes, and the Boston Marathon bombing along with a fantasy football control corpus, examining the relative frequency of words in 36 antonym pairs. We compared the ratio of words within these pairs to the corresponding ratios recovered from an online word norm database.

Results: Partial rank correlation values between observed antonym ratios demonstrate consistent patterns across disasters. For Hurricane Sandy data, 25 antonym pairs have moderate to large effect sizes for discrepancies between observed and normative ratios. Across disasters, 7 pairs are stable and meet effect size criteria. Sentiment analysis, supplementary word frequency counts with respect to disaster proximity, and examples support a "breach" account for the observed results.

Conclusion: Lexical choice between antonyms, only somewhat related to sentiment, suggests that social media capture wide-ranging breaches of normal functioning.

Application: Antonym selection contributes to screening tools based on language style for identifying relevant content and quantifying disruption using social media without the *a priori* specification of content keywords.

Keywords: psycholinguistics, disaster response

Address correspondence to Andrew Hampton, Department of Psychology, Ohio Center of Excellence in Knowledge-enabled Computing (Kno.e.sis), Wright State University, 3640 Colonel Glenn Hwy, Dayton, OH 45435, USA; e-mail: hampton.41@wright.edu.

HUMAN FACTORS
Vol. 59, No. 4, June 2017, pp. 505–519
DOI: 10.1177/0018720817691612
Copyright © 2017, Human Factors and Ergonomics Society.

Disaster disrupts the normal functioning of a society (Perry, 2007), creating what Quarantelli (2008) called the "problems of living" (p. 893) that need "solving" (p. 888). The evaluation of need during disaster response and recovery presents a daunting challenge. Metrics such as emergency call volume and hospital admissions are coarse. Satellite imagery, from a flood for example, may lag changes in need. Moreover, disaster conditions do not neatly correspond to unidimensional physical sensors. Flooding is more than rainfall or storm surge. Prior weather events, topography, demographics, and sociocultural factors such as construction practices and infrastructure all impact whether a given event constitutes a human disaster. Social media data from those situated in the environment promise to reflect the net urgency and experienced disruption. But identifying informative content in natural language poses substantial difficulty.

Many researchers (Palen & Liu, 2007; Sheth, 2009; Starbird, 2011) analyze social media message content, sentiment, organization, and dispersal (among other topics). However, difficulty remains in finding, interpreting, and scaling relevant, actionable signal in a virtual firehose of noise. We complement computationally inspired approaches to analysis by questioning the need for disaster-specific methods. We acknowledge the multiple functions of communication (Searle, 1976) in social media and suggest that these muddy reliance on sentiment as a metric of human experience. Consistent with researchers such as Vedula, Parthasarathy, and Shalin (2016) who consider source trustworthiness in the interpretation of social media data, we challenge the characterization of disruption based on the simple tally of signal counts. Initiated with Purohit et al. (2013), our human factors perspective concerns the psychological processes that generate the signal.

CITIZEN SENSORS IN SOCIAL MEDIA

Sheth (2009) conceptualizes the function of multimodal information broadcast from mobile

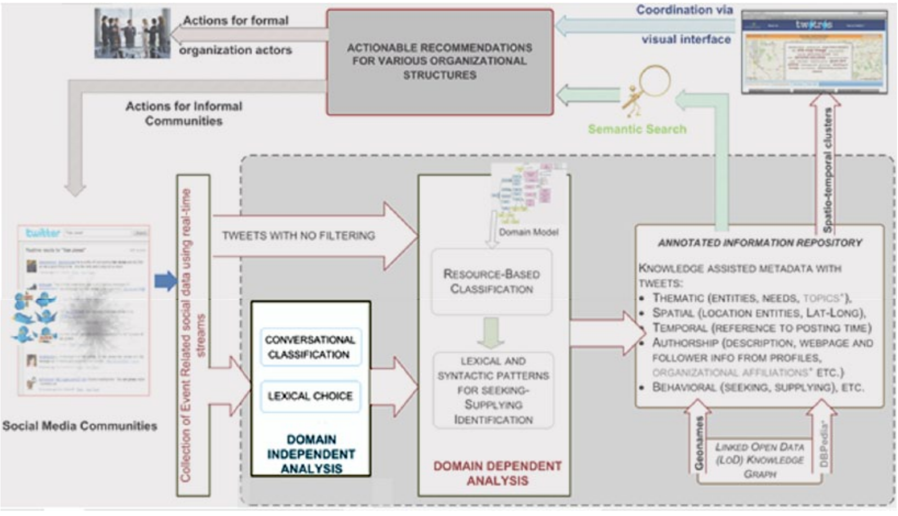

Figure 1. A process for filtering and formatting social media for emergency response. Event-related social media data pass through domain-independent analysis prior to domain-dependent analysis that supports automated annotation for a searchable database. Adapted from "Identifying Seekers and Suppliers in Social Media Communities to Support Crisis Coordination," by H. Purohit et al., 2014, *Computer Supported Cooperative Work, 23*(4–6), 513–545. Copyright 2014 by Springer. Adapted with permission.

computers as a *citizen sensor network*. One such network is Twitter, the world's largest micro-blogging service. Twitter's approximately 313 million active users produce 140-character messages, an artifact of Twitter's text message origins. Eighty-two percent of these users employ their mobile devices (Twitter, 2016a). The persistence and accessibility of Twitter permits a significant portion of the population to provide data reflecting their experience.

For example, when earthquakes struck Virginia in the summer of 2011, people immediately posted this event on their Twitter accounts or else sent similar text messages. Cellular data travel faster than seismic waves, warning others within the affected area as much as 40 seconds before they felt the actual waves. By comparison, the U.S. Geological survey's warning system in 2011 had a best alert time of around 2 minutes, and it required recipient enrollment (Hotz, 2011).

This incidental application supports continued effort to leverage social media for disaster response. To complement the significant existing research base on the function of citizen reports, we suggest psychologically inspired, computationally inexpensive heuristics to support the rapid identification and subsequent analysis of relevant social media messages. Our purpose is not to substitute social media for individual calls for specific assistance but rather, filter and mine the pattern of commentary to ascertain the degree of general distress and focus response. Figure 1 illustrates the role of our perspective in an envisioned tool for the broader emergency response system, initially presented in Purohit et al. (2014). From a collection of social media data, we first conduct domain-independent analyses (highlighted) to reduce the corpus to a more manageable size from which a more conventional domain-dependent analysis may find actionable information. Subsequently, we require an annotated information repository and visualization software to organize findings for the formal response community. This paper extends domain-independent, conversation-based screening presented in Purohit et al. (2013) to include lexical choice.

HUMANS AS INTERPRETIVE SENSORS

Viewing the *function* of humans as sensors within the broader system risks misunderstanding the *processes* whereby humans generate social media messages. Hong and Page (2008) distinguish between generated and interpreted signals. Generated signals map sensor values to world conditions. For example, an old-fashioned thermometer works because we understand the precise response of a set volume of mercury to an increase in temperature and graduate the side of the vial accordingly. Attempts to conceptualize humans as thermometer-like sensors trace to the earliest days of scientific psychology. Fechner (1860/1912) defined the relationship between perceptual experience and physical intensity as a logarithmic function. While psychologists have argued the specifics of this relationship (e.g., Stevens, 1961), they agree that human experience is not likely a linear function of physical intensity.

In agreement with Sheth and Thirunarayan (2012) as well as Hong and Page (2008), we argue that conceptualizing humans as sensors does not properly acknowledge their interpretive processes. These processes, reflecting multiple perspectives, provide the well-established benefit to accuracy from wisdom of the crowds (Parunak, Brueckner, Hong, Page, & Rohwer, 2013). Because disasters disrupt normal functionality (Perry, 2007), we seek linguistic metrics that indicate compromised daily life. Consistent with Norman (1988), we target human interpretation specifically with respect to wide-ranging affordances of and constraints on behavior. Drawing on Bruner's (2003) analysis of narrative, we suggest that social media in disaster convey a *breach* of normative conditions. *Breach* promises a generalizable construct, spanning different disasters, and potentially positive or negative experiences.

LEXICAL CHOICE

We operationalize breach in relation to normative conditions by exploiting the phenomenon of lexical choice. Most psycholinguists make a distinction between thought and language (Gleitman & Papafragou, 2013; Pinker, 1995). Numerous propositions have the same truth conditions (e.g., "This beef is 75% lean" vs. "This beef is 25% fat"), but the different styles have different functional meanings as evidenced by subsequent reasoning (Tversky & Kahneman, 1981). Rather than rely on raw frequency counts as a metric of need, we compare the observed ratio of lexical alternatives in a disaster social media corpus to normative ratios. This Bayesian-inspired approach shifts the focus from individual reports to the altered pattern of such reports. We describe in the following the various influences on lexical choice, including lexical marking (Gilpin, 1973), Gricean maxims (Grice, 1975), and conversational alignment (Levelt, 1993). All of these bear on the relationship between lexical choice observed in the disaster setting and normative values.

Lexical Marking

Clark (1969) explored lexical marking as a psychological phenomenon, defining foundational adjectives as those stored in memory in a more simple and accessible form than their "marked" antonyms. Response time methods revealed that marked adjectives in antonym pairs required longer processing time. Subsequently, Gilpin (1973) studied the use of bipolar rating scales such as *good–bad*. He compared ratings with bipolar scales to ratings with lexically unmarked unipolar scales (*good–not good*) and lexically marked unipolar scales (*bad–not bad*). He found that ratings with unmarked unipolar adjectives more closely resembled bipolar ratings than ratings with marked unipolar adjectives. From these findings, he concluded that bipolar antonyms are not semantically symmetric.

Gricean Maxim of Manner

Consistent with the Gricean maxim of manner (Grice, 1975), speakers should prefer the least obscure expression, namely, the unmarked option. The cognitive simplicity of an unmarked antonym suggests that it will be more frequent than its marked counterpart, and in many cases, this is true. Accordingly, the use of marked language suggests speaker emphasis and should increase in a disaster to capture recipient attention.

Gricean Maxim of Quantity

Message formulation should also respect informativeness (Grice, 1975). Normative English ratios reflect a prevalence of the adjective *big* over its opposite *small* (as measured by the GloWbe database; Davies & Fuchs, 2015), presumably because big things are informative. A particularly dangerous, transient storm should promote words pertaining to above average size and temporarily magnify the disparity with respect to norms. Similarly, "bad" experiences should promote the word *bad* over *good*. In fact, such intuitions support the use of sentiment analysis in assessing public response (Caragea, Squicciarini, Stehle, Neppalli, & Tapia, 2014; Lin & Margolin, 2014; Thelwall, Buckley, & Paltoglou, 2011). However, we suggest that the environment influences the patterns of lexical choice more broadly than sentiment, both in the long term as reflected in word norms that mirror daily life and in the short run via disaster-mediated departures in lexical choice.

Conversational Alignment

Lexical choice also reflects the phenomenon of conversational alignment or linguistic style matching (Levelt, 1993; Niederhoffer & Pennebaker, 2002; Pickering & Garrod, 2004). Accordingly, conversational partners converge on various message features. This phenomenon will prove relevant as we investigate the effect of geographic proximity on lexical choice.

OVERALL RATIONALE AND MOTIVATION

Limitations in the sufficiency and availability of metrics during and following a disaster highlight the need for correlated measures that provide readily accessible insight. The psychological processes that underlie language production suggest complementary approaches to the conventional computational focus on the tallying of specific message content. We suggest that message filtering and the general assessment of need is informed by comparing observed lexical choice between antonyms to word norms beyond expressions of sentiment. Consistent with the wisdom of the crowd literature, this approach relies on style to identify a variety of interpretive sensors that respond to different breaches of normal functioning in the disaster context. The focus on style allows us to employ a normative standard for evaluating observed patterns and supports comparison across varying population bases and disaster types. Geographic proximity to disaster provides a natural manipulation approximating the level of urgency.

In the following, we examine the discrepancy of lexical choice ratios relative to word norms. We first address the generality of our claims across three disaster and one non-disaster corpora and then examine the specific word pairings that contribute to our findings, with particular attention paid to the role of sentiment. We support our claim that deviation from lexical norms corresponds to personal narrative by supplementing our lexical choice analyses with a more fine-grained spatial analysis using aggregated word metrics from the Linguistic Inquiry Word Count software (LIWC) (Pennebaker, Boyd, Jordan, & Blackburn, 2015). Specific examples illustrate the kind of content that becomes identifiable using our heuristic.

METHOD

Data Sets

We collected several million unique tweets from three disasters in different regions and of different types across the United States. We eliminated retweets or forwarded messages because they are likely heavily influenced by organizational reports, which contaminate our interest in personal narrative (Starbird & Palen, 2010). Hurricane Sandy and the Oklahoma tornado represent natural hazards. The Boston Marathon bombing represents an intentional, manmade conflict (Quarantelli, 2005) that serves to test the generality of our findings. We used a social media analysis tool, Twitris (Purohit & Sheth, 2013), to identify the tweets within the target times (see Table 1 for inclusion criteria). As is typical of social media data, our location-specific corpus constitutes a subset of the full data stream. We further segmented the resulting corpora according to location tags specified in Tables 2 and 3. In choosing these bounding boxes, we balanced the size of the corpus, frequency of pairs, and region size.

For the purposes of comparison, we also randomly selected 50,000 tweets from a location-independent corpus, largely free of retweets, assembled from search terms related to fantasy football. Fantasy football is not immune to the

TABLE 1: Inclusion Criteria for Tweet Data Sets

Event	Start	End	Crawling Word Set
Hurricane Sandy n ~ 4.6 million	October 27, 2012	November 7, 2012	Hurricane Sandy, Frankenstorm, #Sandy
Boston Marathon Bombing n ~ 4.5 million	April 15, 2013	April 25, 2013	Boston explosion, Boston explosions, Boston blast, Boston blasts, Boston tragedies, Boston tragedy, PrayForBoston, Boston attack, Boston attacks, Boston terrorist, Boston terrorists, Boston tragic, Boston Marathon, Boston explosive, Boston bomb, Boston bombing
Oklahoma Tornado n ~ 2.8 million	May 20, 2013	May 30, 2013	Oklahoma tornado, Oklahoma storm, Oklahoma relief, Oklahoma volunteer, Oklahoma disaster, #Moore, Moore relief, Moore storm, Moore tornado, Moore flood, Moore disaster, Moore volunteer, #OKC relief, #OKC disaster, #OKC tornado, #OKC flood, #OKC volunteer, #OKC storm, #OKhaves, #OKwx, Shawnee, Norman, Pottawatomie, Mary Fallin, #OKC, #OKneeds, #OK, #OK tornado, #OK relief, #OK flood, #OK disaster, #OK volunteer, #OK storm
Fantasy Football n ~ 1.0 million	September 12, 2015	October 11, 2016	NFL, NFLFantasy, DawgPound, RiseUp, SieTheDay, GoPackGo, Skol, WhoDey, FlyEaglesFly, KeepPounding, Jaguars, Patriots, Broncos, Chargers, Chiefs, RaiderNation, ForTheShoe, GoNiners, WeAreTexans, OnePride, GiantsPride, GoBills, TitanUp, JetUp, MobSquad, RavensFlock, WeAre12, BeRedSeeRed, FinsUp, HereWeGo, FantasyFootball, Seahawks, Bengals, Falcons, CowBoys, Texans, 49ers, Titans, Redskins, Vikings, Buccaneers, MiamiDolphins, Eagles, Steelers, Cardinals

influence of marking, manner, quantity, and alignment. However, the limited importance of the events should damp the trends in lexical choice.

Antonym Pairs

Previous research into marked language (e.g., Gilpin, 1973) provided a starting point for compiling the set of antonym pairs. Additionally, we consulted a list of common positive adjectives from the Oxford English Dictionary online and paired each with its most common antonym using www.thesaurus.com. Finally, we read through 100 randomly selected tweets for each disaster corpus and manually identified the adjectives used, later adding the corresponding antonym as previously described. We eliminated redundancies to complete the list of 36 pairs. The selected pairs appear in Tables 5 and 6 in the Results section.

TABLE 2: Coordinates of the Bounding Boxes Used for Direct Comparison Between Events

	Southwest Corner		Northeast Corner		
Event	Latitude	Longitude	Latitude	Longitude	*N*
Hurricane Sandy	39.270	–74.612	41.327	–71.816	146,764
Boston Marathon bombing	42.022	–71.802	42.865	–70.572	54,348
Oklahoma tornado	34.551	–98.465	36.008	–96.597	45,788

TABLE 3: Coordinates of the Bounding Boxes Used for the "Doughnut" Analysis

		Southwest Corner		Northeast Corner		
Event		Latitude	Longitude	Latitude	Longitude	*N*
Hurricane Sandy	Small	39.270	–74.612	41.327	–71.816	146,764
	Doughnut	24.857	–87.931	47.469	–66.952	212,115
Boston Marathon	Small	42.332	–71.111	42.370	–71.053	33,977
bombing	Doughnut	42.022	–71.802	42.865	–70.572	20,371
Oklahoma	Small	35.250	–97.653	35.400	–97.319	4,383
tornado	Doughnut	34.551	–98.465	36.008	–96.597	41,405

Note. Tweets in our data set with distal geotags outside of these boxes include 366,604 from Hurricane Sandy, 974,314 from the Boston Marathon bombing, and 477,336 from the Oklahoma tornado.

TABLE 4: Partial Spearman Rank Correlation Values for Observed Proportion Values Controlling for Normative Influence

	Fantasy Football	Boston	Oklahoma
Sandy	.32	.50	.75
Oklahoma	.46	.64	
Boston	.19		

Note. n = 36 for all comparisons. Approximate critical r = .33 for a = .05 (Noether, 1976).

Word frequency norms. The GloWbe database of Internet language (Davies & Fuchs, 2015) estimated the normative frequency of words within the United States. We calculated a baseline ratio of use between each word and its matched alternate such that the less common word in the control corpus was represented as a fraction of the total use. For example, *all* appeared 1,306,886 times in the GloWbe corpus, whereas its marked alternate *some* appeared 724,227 times. Thus, the proportion of *some* to the pair total equals 0.36. This approach standardizes comparison across antonym pairs with different absolute frequencies.

Sentiment norms. We used SentiStrength (Thelwall, Buckley, Paltoglou, Cai, & Kappas, 2010) to obtain sentiment scale values for each member of the antonym pair. We determined any difference in sentiment value greater than zero to be an affectively asymmetric pair and confirmed this determination against the valence values in the 14,000-item Warriner, Kuperman, and Brysbaert (2013) database when possible.

Tabulation

We tallied how often each member of the antonym pair appeared in each disaster corpus using the NotePad++ "find string" function including a leading and trailing space. This yielded two scores (one for each member of the antonym pair) against which to compare the corresponding two scores in the normative database. We also applied the LIWC tool (Pennebaker et al., 2015) to characterize the overall content of specific subsets of the corpora.

Statistical Analysis

A number of statistical concerns challenge the comparison of observed proportions between the corpora, such as assumptions of linear relationships, the underlying distribution of proportion values, and the fundamental correlation of the observed proportions with a common normative base rate. We describe the relationship between two sets of observed proportions using a Spearman's rank correlation and recover partial correlations between disaster corpora controlling for the common normative value (National Institute of Standards and Technology, 2012).

To evaluate the proportions for a particular antonym pair, we employed effect size metrics. The Cox logit method (Lipsey & Wilson, 2001; see Figure 2) compares proportions using d values. We employed an effect size calculator (Wilson, 2000) to obtain both d values and the surrounding 95% confidence interval that takes into account the number of observations.

$$(\ln((\frac{observed}{(1-observed)})/(\frac{norm}{(1-norm)})))/1.65$$

Figure 2. Cox logit d formula.

However, although the source corpora are quite large, the number of instances of a particular antonym pair can be small, particularly as the geographic span of the corpus shrinks. As a result, a 95% confidence interval including zero may surround otherwise impressive d values. Moreover, d increases with asymmetry in the contributing binary proportions, reinforcing the standard caution regarding overemphasis on effect size (Cohen, 1988).

TABLE 5: Moderate to Large Effect Size Departures From Norms by Disaster

Word Pair	Hurricane Sandy *d*	Oklahoma Tornado *d*	Boston Marathon Bombing *d*	Fantasy Football *d*
Horrible/wonderful[a]	1.03	0.57	1.48	1.76
Stop/start	0.79	0.39	0.70	**–0.75**
Warm/cool	**0.60**	–0.87	–0.77	–0.50
Severe/minor	0.54	1.92	0.96	NS
Some/all	–0.39	–0.37	–0.65	BT
Sane/crazy[a,b]	–0.45	–1.58	–1.24	NS
Alone/together	–1.22	–1.62	–0.68	NS
Tiny/massive	–1.26	–1.54	–1.35	NS
Under/over	–0.57	–0.95	NS	–0.61
Soft/hard[b]	–0.83	–1.20	NS	**0.75**
Fake/real[a]	0.96	NS	1.12	NS
Whole/part	0.58	NS	0.90	BT
Terrible/great[a]	0.48	NS	0.73	0.68
Smart/stupid[a]	–0.89	NS	–0.58	–0.40
East/west	**–1.07**	NS	0.76	NS
Unsafe/safe[a]	–1.40	NS	–0.78	**0.64**
Low/high[a,b]	BT	–0.61	–0.56	NS
Global/local	BT	–0.71	–1.24	NS

Note. NS indicates that a *d* 95% confidence interval contained 0. BT indicates that the *d* value fell below our threshold of an absolute value of 0.37. Less frequent words according to GloWbe appear first in the pair description so that positive *d* indicates an observed increase in the less frequent word and negative *d* indicates an observed increase in the more frequent word. Bold entries indicate a reversal of direction across disasters. Pairings with results unique to Hurricane Sandy are not presented.
[a]Indicates a sentiment asymmetry per SentiStrength.
[b]Indicates disagreement between Warriner sentiment classification and SentiStrength.

TABLE 6: Hurricane Sandy *d* Analysis for Pairs Not Included in Table 5

Significant *d* Above Threshold	*d*	Significant *d* Below Threshold	*d*	Nonsignificant *d*
Worse/better[a]	0.63	**Global/local**	**0.15**	Stale/fresh[a]
Last/first	0.61	Big/little	0.13	Every/any
Out/in	0.46	Down/up	–0.15	Imperfect/perfect[a]
Slow/fast	0.43	Low/high[a,b]	–0.25	Few/much
Black/white	0.40	Dead/live[a]	–0.28	Dull/amusing[a]
Bad/good[a]	0.40	Boring/fun[a]	–0.32	Shorter/longer
Left/right	0.37			
Large/small	–0.72			

Note. Less frequent words according to GloWbe appear first in the pair description so that positive *d* indicates an observed increase in the less frequent word and negative *d* indicates an observed increase in the more frequent word. Bold pairings indicate a reversal of direction across disasters.
[a]Indicates a sentiment asymmetry per SentiStrength.
[b]Indicates disagreement between Warriner sentiment classification and SentiStrength.

RESULTS AND DISCUSSION

First, we correlate the observed proportions of lexical choice between pairs of disasters to demonstrate generality across disaster corpora. Next, we examine specific antonym pairs to determine which ones depart from normative ratios consistently across corpora, the directionality of this departure, and the dependence of these findings on sentiment. We then split the corpora by proximity to disaster epicenter and employ LIWC to support the claim that the corpora contain personal reflections concerning food, space, time, and motion with limited evidence of negative sentiment as the dominant construct. A sample of tweets determined by our lexical choice heuristic illustrates the notion of breach underlying the observed pattern of results.

Disaster Generality

To examine the relationship between the observed proportions across the three events, we report the observed partial Spearman r values for the relationship between two event corpora controlling for the underlying relationship of both to normative values (see Table 4). The resulting partial r values confirm positive relationships between the observed proportions despite the influence of a common normative base rate. These results suggest comparable patterns of lexical choice in disaster. Correlations with the fantasy football corpus are generally smaller, although not completely absent. Because Spearman values are a function of the entire data set, attributing the significant Oklahoma–fantasy football correlation is not straightforward. Some relatively unique word pairs appear to contribute to concordance, such as *east/west, up/down*, and *perfect/imperfect*.

Examination of Specific Antonym Pairs

Using an effect size metric to examine the discrepancy of observed ratios from the normative standard, we present our results regarding specific antonym pairs in two tables. The first (Table 5) focuses on those *d* values that exceed an absolute value of 0.37 with 95% confidence intervals that do not include zero for at least two disasters. The direction of the discrepancy is roughly evenly split between increases and decreases in the prevalence of the more rare term. Seven pairs are consistent across all three disaster events. Only two word pairs diverge from word norms in opposite directions between events. Of the seven pairs that show a consistent pattern across disaster events, only one is consistent with the fantasy football corpus: *horrible/wonderful. Stop/start* is contradictory.

Table 6 completes the list of 36 pairs, illustrating singleton effects for Hurricane Sandy, where the large corpus provides narrow confidence

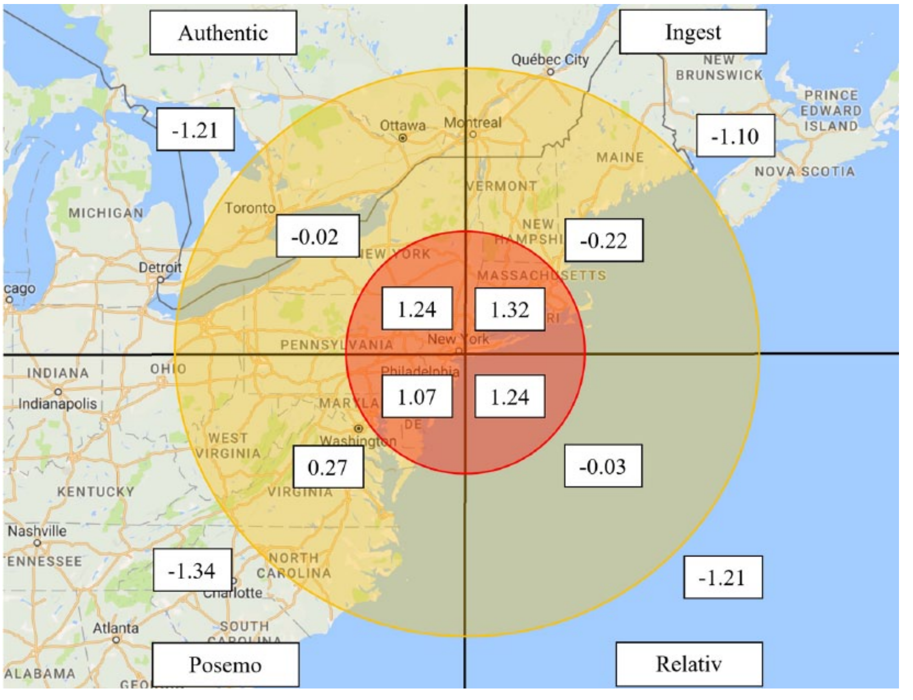

Figure 3. Spatial representation of Linguistic Inquiry Word Count software categories in *z* scores relative to the hardest hit areas of Hurricane Sandy. Image credit: Google Inc.

intervals. For Hurricane Sandy, most of the word pairs (24/36) meet our discrepancy criteria.

Annotations on the entries in Tables 5 and 6 indicate affective (sentiment) asymmetry in word pairings. Many pairs do not reflect affective asymmetry. Some pairs such as *minor*/*severe*, *massive*/*tiny*, and *in*/*out* have clear context relevance. But neither affective asymmetry nor context relevance explain pairs such as *some*/*all*, *stop*/*start*, and *alone*/*together*. The pair *some*/*all* is particularly interesting because it includes two high frequency "stop words" that are generally ignored in text mining.

Spatial Proximity

The *d* values should attenuate with distance from the disaster epicenter. However, spatial proximity analysis requires segmenting the corpora into small subsets that generally result in prohibitively low frequency values for our observed pairs, with a preponderance of missing data and nonsignificant *d* values. To examine trends with respect to spatial proximity, we change metrics from individual words to LIWC categories, each of which aggregates over dozens of words and returns a value for the observed frequency per thousand words.

We examine three subsets of data for each event as indicated in Table 3: a small bounding box, a doughnut consisting of the remainder of content in the large bounding box less the content in the small bounding box, and a distal subset consisting of content known to originate outside the large bounding box. LIWC categories are not orthogonal, and we do not report all of them. However, many of the trends across proximity are consistent for all three events. Most are consistent for Sandy and Oklahoma.

Figures 3 to 5 illustrate four of these metrics for the three events. To aid comparison, we converted observed words-per-thousand values to *z* scores based on our three samples (central, doughnut, and distal). High *authentic* scores indicate a more honest, personal, and disclosing text; lower scores suggest a more guarded, distanced text. *Ingest* refers to kinds of food and ingestion terminology such as *taste* and *dine. Posemo* refers to positively valenced emotional

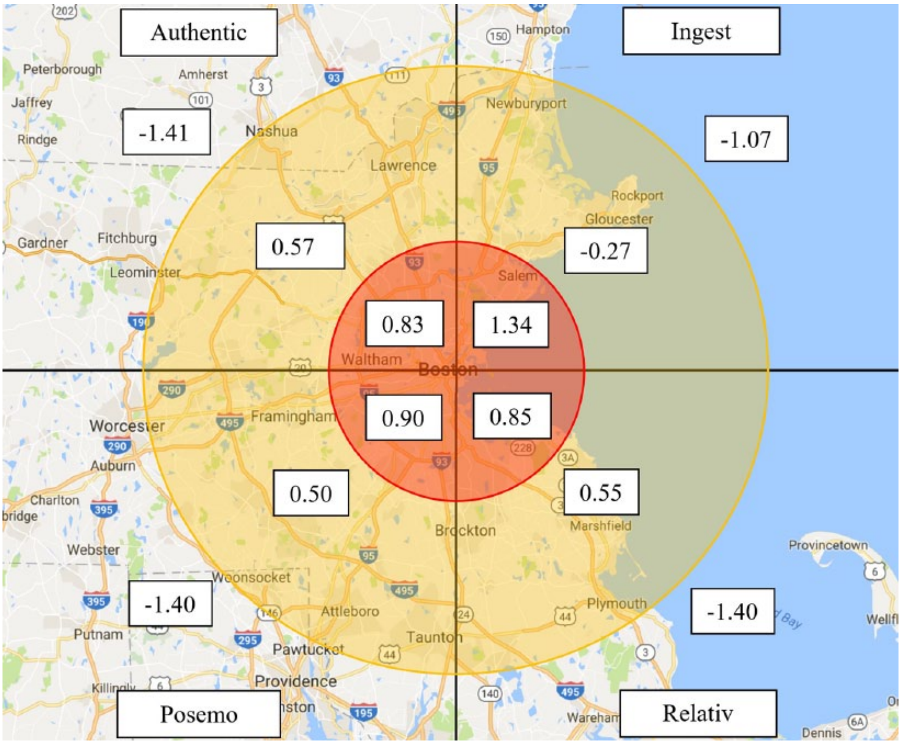

Figure 4. Spatial representation of Linguistic Inquiry Word Count software categories in *z* scores relative to the Boston Marathon finish line. Image credit: Google Inc.

words and word combinations. *Relativ* combines spatial, temporal, and motion references.

Across all three figures, authenticity decreases with distance from the epicenter, suggesting that our corpora are tapping personal comments and not organizational reports. Second, concern with food, space, and time, all critical to survival in a disaster, decrease with distance. Finally, positive emotion increases with proximity to the event. Negative emotion, though somewhat less compelling, is consistent with this trend. This reinforces concern for reliance on sentiment as a metric of need.

Tweet Content

We have established a consistent pattern of lexical choice across disasters, cast doubt on sentiment as the central factor in the set of significant antonym pairs, and suggested personal, disaster-relevant accounts as responsible for our results. Here, we use manually selected examples of tweet content to illustrate breach as an explanation for our findings. All of the examples in Table 7 illustrate the disruption in normal activity—notable but not uniformly highly negative. The mix of sentiment reflects the range of communicative function, including commissives, directives, and beliefs, along with factual assertion.

Certainly not all of the tweets meeting our criteria are actionable. The Boston police in Example 5 were surely aware of their presence at the train stations; the tweet does however indicate the public response. And while Example 7 does not require an organizational response, it does inform the response organizations of community activity, which can be highly influential in distributing resources. We note the wide-ranging idiosyncratic content and language apart from our antonym-pair heuristic, indicating power outages, downed trees, and disrupted traffic. Our stylistic heuristic indicators support the identification of numerous specific compromises, phrased in virtually unlimited fashion.

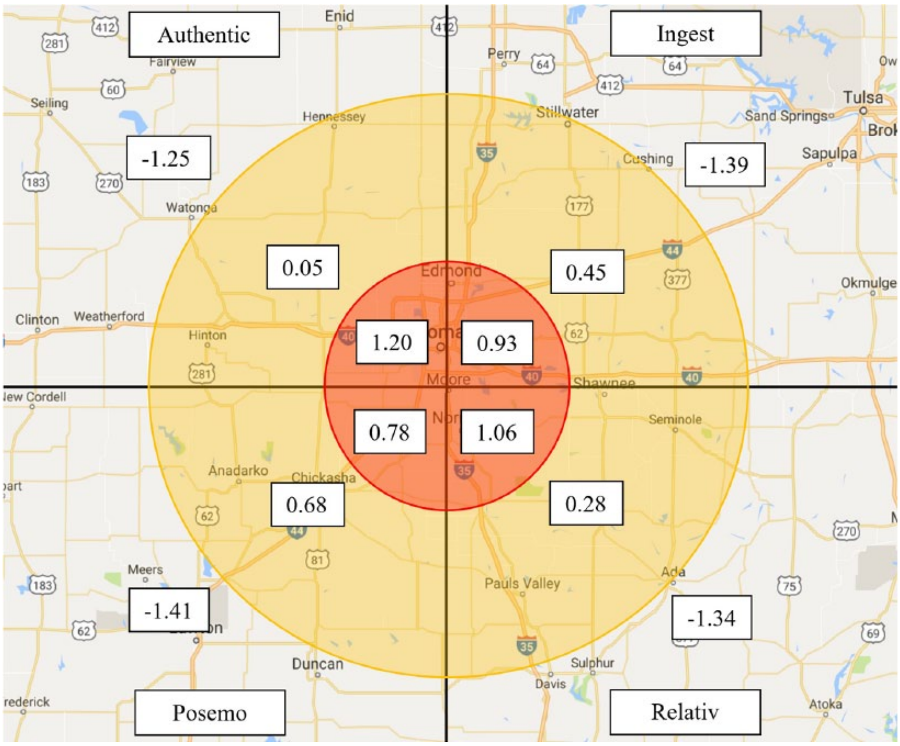

Figure 5. Spatial representation of Linguistic Inquiry Word Count software categories in z scores relative to hardest hit areas of the storms in central Oklahoma. Image credit: Google Inc.

CONCLUSIONS

By examining lexical choice across a variety of situations with reference to a normative distribution, we identify patterns suggestive of disruption to the patterns of normal living. Consistency across disaster types along with geographic and cultural diversity is encouraging. However, we note a number of limitations prior to identifying the contributions of this research.

Limitations

Antonym pairs. Although the psychological phenomenon of lexical marking led us to study a broad set of antonym pairs, such as *some*/*all, stop*/*start*, and *alone*/*together*, these are by no means an exhaustive set. The pairings themselves, though principled, remain subjective. A good example of the resulting problem is *big*/*little* and *large*/*small*. They diverge from norms in opposite directions. We could envision combining them or pairing them differently. Finally, other influences on lexical choice, like polysemy and idioms, potentially interfere with the exploitation of observed departures from word norms as indicative of departures from normative conditions. These influences may not be equally likely across disaster settings and hamper the effort to develop generalizable metrics. Specificity of these pairs to the disaster setting also requires consideration. While they provide a rather poor account of lexical choice in fantasy football, some pairs such as *wonderful*/*horrible* are consistent with breach in this far less consequential domain. However, the entire motivation for our approach is domain independence that does not require a model of the disaster in order to search for information. Some absence of specificity is to be expected. Our approach does not yet escape the empirical tradition of text mining, exploiting those pairs that happen to work for the corpus at hand. Replication and theory should guide the selection of diagnostic pairs and their direction. We do not require all pairs of antonyms

TABLE 7: Tweet Examples

	Word Pairs	Tweet Example **[SentiStrength Rating]**
1	Minor/severe	At work wanting to go home! **[2]** I've busted my butt all week and worked every day of the hurricane. **[1]** I'm tired and in severe pain! **[–5]**
2	Start/ stop	If you are driving through Moore on I-35, stop pausing to look at the wreckage. **[–2]** It's making traffic a problem **[0]**
3	Wonderful/horrible[a]	@Drew_Hampton horrible I still don't have any power from hurricane sandy & I'm freezing :(**[–4]**
4	All/some	Yes and all flights to Boston are totally full due to the bombing **[–2]**
5	Crazy/sane[b]	Wow there's a lot more security and police at the train stations now. **[0]** This is crazy ?? #BostonBombing **[–2]**
6	Massive/tiny	Storm knocked down one of the massive trees in front of my #house **[–2]** #rip #sandy #hurricane #HurricaneSandy **[NA]**
7	Together/alone	Getting a team together to go up near Moore to cut tree limbs. **[1]** Call me if you're interested #Oklahoma #Tornado #Relief **[2]**

Note. The SentiStrength scale ranges from −5 to 5. We have superficially altered tweet examples in compliance with Twitter's privacy policy. These examples illustrate a mix of assertions (3), directives (2, 7), and commissives (7) that influence overall sentiment ratings.
[a]Indicates affectively asymmetric pairs according to SentiStrength.
[b]Indicates a disagreement on affective asymmetry between Warriner and SentiStrength.

to be useful but rather that some subset proves consistently and ideally a priori diagnostic.

Computational issues. Modeling in the spirit of multiple regression is likely required to tease out the multiple additive, if not interactive, influences on lexical choice. The influence of linguistic style matching is a particular threat to the development of lexical choice metrics. Because a recipient is more likely to produce the adjective that someone else just used, we must attempt to estimate and adjust any metrics for this effect. Institutional reports (e.g., from the Red Cross or local hospitals) need to be screened out if a lexical choice metric is meant to identify public experience. Eliminating retweets, as we did in our analysis, and the LIWC authenticity metric somewhat mitigate this concern. We have not yet tackled the scaling issues, within or between disasters, in order to understand how our metric maps to true disruption. However, while our argument and analysis highlights this problem, it is not unique to our approach. Finally, a significant natural language processing problem persists in the automated interpretation of potentially informative messages.

Mediated communication. Most text-based communication interferes with monitoring recipient comprehension (Clark & Krych, 2004; Clark & Schaefer, 1989) and likely alters communication behavior. The absence of a normative corpus restricted to Twitter usage raises the possibility that the observed patterns reflect more general discrepancies between Twitter-mediated communication and other functions of Internet language. We partially addressed this concern by demonstrating attenuated effects with a non-disaster corpus.

However, we believe that the medium works to our advantage. Holtgraves and Paul (2013) studied text messages versus telephone conversations unobtrusively and found that people tended to speak more simply via text, using more words associated with the personal and affective than they did in recorded phone conversations. Simple, direct, and personally centered messages likely allow for easier automated processing (given the findings of Holtgraves & Paul, 2013) as well as greater likelihood of relevance with respect to the sender's immediate situation.

The standard Twitter API, while useful in capturing content, provides a limited sample of the full content stream. This reduced power in our *d* metric, requiring the exclusion of results embedded in large confidence intervals covering zero. This problem will yield to greater access, necessary not only for research purposes but also to obtain a real-time metric.

Contributions

We have demonstrated interpretable patterns of language behavior in social media during disasters using a novel, psychologically inspired metric of lexical choice relative to a normative standard. Comparison to an external standard constitutes an alternative approach to reliance on internal trend detection within a corpus (Bifet, Holmes, Pfahringer, & Gavaldà, 2011) or concern with the veracity of any individual report contributing to a tally. Methodological considerations such as concern for base rates and control corpora showcase the benefits of an experimentally oriented human factors approach to the analysis of big data. Our Bayesian-inspired analysis has exploited base rates to identify language patterns given a known breach, providing an important step in the identification of true social media alarms, that is, breach given observed language patterns.

General sentinels of breach enable analysis across different disasters and do not require an a priori set of content terms. Despite some superficial similarity with the skewness analysis of sentiment that Caragea et al. (2014) provide, we flag sentiment analysis itself as nuanced. Here, positive sentiment increases with proximity to the disaster, perhaps reflecting commissives associated with prosocial behavior (Quarantelli, 2008). Alternatively, our LIWC results may be revealing the amplified dispersal of tragic content or even misinformation (Lin & Margolin, 2014; Starbird, Maddock, Orand, Achterman, & Mason, 2014; Thelwall et al., 2011).

Sentinels naturally accommodate the diversity of perspective that supports wisdom of the crowds by revealing a variety of unanticipated, specific "problems of living." While others have noted the potential of social media for capturing the individual narrative (Anderson et al., 2016), our analysis points to the entire disaster corpus as exhibiting narrative properties in its own right.

Practical Considerations

The ability to identify relative levels of urgency at high resolution, as this research may facilitate, will allow disaster management professionals to deploy aid with more precise information in both spatial and temporal frames of reference. In addition to initial conditions, social media provide continuing status reports on the affected areas. Often with disasters that extend beyond days into weeks (or months, as with Hurricanes Sandy and Katrina), the challenges lie in understanding continuing unmet resource needs (Anderson et al., 2016; Purohit et al., 2014).

Twitter reached 100 million users in only five and a half years (Twitter, 2016b). Harnessing this medium for the exchange of information between the public and the authorities lacks the procedures we have for other media such as 9-1-1. Let this effort, grounded in the psychology of language production, be a step in that direction.

ACKNOWLEDGMENTS

The authors are grateful to Hussein Al-Olimat and Shreyansh Bhatt for assisting in the compilation of the Tweet corpora. John Flach, Art Graesser, Joseph Houpt, Ion Juvina, Brad Minnery, Hemant Purohit, Srinivasan Parthasarathy, Amit Sheth, Krishnaprasad Thirunarayan, and journal reviewers provided helpful feedback. We acknowledge partial support from the National Science Foundation (NSF) awards: SoCS IIS-1111182: Social Media Enhanced Organizational Sensemaking in Emergency Response and EAR 1520870: Hazard SEES: Social and Physical Sensing Enabled Decision Support for Disaster Management and Response. The opinions here reflect only those of the authors. Andrew J. Hampton is currently affiliated with the University of Memphis, Tennessee.

KEY POINTS

- Social media represent a powerful analysis tool in critical events such as disasters but requires principled reduction of the massive corpora.
- Psycholinguistic consideration of the mechanisms of language production add a domain-independent filter beyond that of traditional sentiment analysis or context-specific ontologies.
- Analysis of antonym pairs relative to normative use provides ratios that complement reliance on

frequency counts or trust-weighted individual reports.

- Breach in normative functioning reflected by language style may direct disaster response entities toward actionable information in the otherwise overwhelming social media stream.

REFERENCES

Anderson, J., Kogan, M., Bica, M., Palen, L., Anderson, K., Stowe, K., . . . Henderson, J. (2016). *Far, far away in far Rockaway: Responses to risks and impacts during Hurricane Sandy through first-person social media narratives*. Paper presented at the 13th Information Systems for Crisis Response and Management, Rio de Janeiro, Brazil.

Bifet, A., Holmes, G., Pfahringer, B., & Gavaldà, R. (2011). Detecting sentiment change in Twitter streaming data. *Journal of Machine Learning Workshop and Conference Proceedings*, *17*, 5–11.

Bruner, J. (2003). The narrative construction of reality. In M. Mateas and P. Sengers (Eds.), *Narrative Intelligence* (pp. 41–62). Philadelphia, PA: John Benjamins Publishing.

Caragea, C., Squicciarini, A., Stehle, S., Neppalli, K., & Tapia, A. (2014). *Mapping moods: Geo-mapped sentiment analysis during hurricane Sandy*. Paper presented at the 11th Information Systems for Crisis Response and Management, University Park, PA.

Clark, H. H. (1969). Linguistic processes in deductive reasoning. *Psychological Review*, *76*(4), 387–404.

Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, *50*, 62–81.

Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science*, *13*(2), 259–294.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Davies, M., & Fuchs, R. (2015). Expanding horizons in the study of world Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *English World-Wide*, *36*(1), 1–28.

Fechner, G. T. (1912). The measurement of sensation (Herbert Sidney Langfeld, Trans.). In B. Rand (Ed.), *The classic psychologists* (pp. 562–572). Boston, MA: Houghton Mifflin. (Original work published 1860)

Gilpin, A. R. (1973). Lexical marking effects in the semantic differential. *Journal of Psychology*, *85*(2), 277. doi:10.1080/00223980.1973.9915658

Gleitman, L., & Papafragou, A. (2013). Relations between language and thought. In D. Reisberg (Ed.), *The Oxford handbook of cognitive psychology* (pp. 504–523). Oxford, UK: Oxford University Press.

Grice, H. P. (1975). *Logic and conversation. Syntax and semantics* (Vol. *3*). San Diego, Ca: Academic Press.

Holtgraves, T., & Paul, K. (2013). Texting versus talking: An exploration in telecommunication language. *Telematics and Informatics*, *30*(4), 289–295.

Hong, L., & Page, S. E. (2008). *Some microfoundations of collective wisdom*. Retrieved from https://raison-publique.fr/IMG/pdf/Scott_Page-2.pdf

Hotz, R. L. (2011). Decoding our chatter. *The Wall Street Journal*. Retrieved from http://online.wsj.com/article/SB10001424052970204138204576598942105167646.html

Levelt, W. J. (1993). Lexical access in speech production. In *Knowledge and language* (pp. 241–251). New York, NY: Springer.

Lin, Y. R., & Margolin, D. (2014). The ripple of fear, sympathy and solidarity during the Boston bombings. *EPJ Data Science*, *3*(1), 1.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage Publications, Inc.

National Institute of Standards and Technology. (2012). *Partial rank correlation*. Retrieved from http://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/partraco.htm

Niederhoffer, K. G., & Pennebaker, J. W. (2002). Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, *21*(4), 337–360.

Noether, G. E. (1976). *Introduction to statistics: A nonparametric approach*. Boston, MA: Houghton Mifflin.

Norman, D. (1988). *The psychology of everyday things*. New York, NY: Basic Books.

Palen, L., & Liu, S. B. (2007, April). Citizen communications in crisis: Anticipating a future of ICT-supported public participation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 727–736). New York, NY: ACM.

Parunak, H., Brueckner, S. A., Hong, L., Page, S. E., & Rohwer, R. (2013). Characterizing and aggregating agent estimates. In *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-Agent Systems* (pp. 1021–1028). Liverpool: International Foundation for Autonomous Agents and Multi-agent Systems.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin.

Perry, R. W. (2007). What is a disaster? In H. Rodriguez, E. L. Quarantelli, & R. R. Dynes (Eds.), *Handbook of disaster research* (pp. 1–15). New York, NY: Springer Science and Business Media.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169–190.

Pinker, S. (1995). *The language instinct: The new science of language and mind* (Vol. *7529*). London: Penguin UK.

Purohit, H., Hampton, A., Bhatt, S., Shalin, V. L., Sheth, A. P., & Flach, J. M. (2014). Identifying seekers and suppliers in social media communities to support crisis coordination. *Computer Supported Cooperative Work (CSCW), 23*(4–6), 513–545.

Purohit, H., Hampton, A., Shalin, V. L., Sheth, A. P., Flach, J. M., & Bhatt, S. (2013). What kind of #conversation is Twitter? Mining #psycholinguistic cues for emergency coordination. *Computers in Human Behavior*, *29*(6), 2438–2447.

Purohit, H., & Sheth, A. P. (2013). Twitris v3: From citizen sensing to analysis, coordination and action. In *Proceedings of the Seventh International Conference on Weblogs and Social Media, 8-11 July 2013, Cambridge, MA* (pp. 746–747). Palo Alto, CA: The AAAI Press.

Quarantelli, E. (Ed.). (2005). *What is a disaster? A dozen perspectives on the question*. London: Routledge.

Quarantelli, E. (2008). Conventional beliefs and counterintuitive realities. *Social Research: An International Quarterly of the Social Sciences*, *75*(3), 873–904.

Searle, J. R. (1976). *A taxonomy of illocutionary acts*. Trier, Germany: Linguistic Agency University of Trier.

Sheth, A. P. (2009). Citizen sensing, social signals, and enriching human experience. *IEEE Internet Computing*, *13*(4), 87.

Sheth, A. P., & Thirunarayan, K. (2012). Semantics empowered web 3.0: Managing enterprise, social, sensor, and cloud-based data and services for advanced applications. *Synthesis Lectures* *on Data Management*, *4*(6), 1–175 doi:10.2200/S00455ED-1V01Y201211DTM031

Starbird, K. (2011). Digital volunteerism during disaster: Crowdsourcing information processing. In *Conference on Human Factors in Computing Systems* (pp. 7–12). New York, NY: ACM.

Starbird, K., Maddock, J., Orand, M., Achterman, P., & Mason, R. M. (2014). *Rumors, false flags and digital vigilantes: Misinformation on Twitter after the 2013 Boston Marathon bombing*. Retrieved from https://www.researchgate.net/publication/266629432_Rumors_False_Flags_and_Digital_Vigilantes_Misinformation_on_Twitter_After_the_2013_Boston_Marathon_Bombing

Starbird, K., & Palen, L. (2010, May). *Pass it on?: Retweeting in mass emergencies*. Paper presented at the 7th International Information Systems for Crisis Response and Management Conference, Seattle, WA.

Stevens, S. S. (1961). To honor Fechner and repeal his law. *Science*, *133*, 80–86.

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology*, *62*(2), 406–418.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, *61*(12), 2544–2558.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458.

Twitter. (2016a). *Company*. Retrieved from https://about.twitter.com/company

Twitter. (2016b). *Milestones*. Retrieved from https://about.twitter.com/milestones

Vedula, N., Parthasarathy, S., & Shalin, V. L. (2016). Predicting trust relations among users in a social network. On the roles of influence, cohesion and valence. Retrieved from http://sentic.net/wisdom2016vedula.pdf

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207.

Wilson, D. B. (2000). Practical meta-analysis effect size calculator [Software]. Retrieved from http://www.campbellcollaboration.org/escalc/html/EffectSizeCalculator-SMD10.php

Andrew J. Hampton is a doctoral student at Wright State University, having obtained his MS in human factors and industrial/organizational psychology from that institute in 2012. He currently works as a researcher at the Institute for Intelligent Systems at the University of Memphis.

Valerie L. Shalin is an associate professor of psychology at Wright State University, having earned her PhD in learning, developmental, and cognitive psychology from the University of Pittsburgh in 1987.

Date received: November 21, 2016
Date accepted: December 28, 2016