Listening. Learning. Leading.®

Research Report
ETS RR-12-19

Topical Trends in a Corpus of Persuasive Writing

Michael Heilman

Nitin Madnani

October 2012

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Topical Trends in a Corpus of Persuasive Writing

Michael Heilman and Nitin Madnani
ETS, Princeton, New Jersey

October 2012

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Associate Editor: Joel Tetreault

Reviewers: Christopher Brew and Paul Deane

.

Abstract

Many writing assessments use generic prompts about social issues. However, we currently lack an understanding of how test takers respond to such prompts. In the absence of such an understanding, automated scoring systems may not be as reliable as they could be and may worsen over time. To move toward a deeper understanding of responses to generic issue prompts, we analyzed topical trends in test takers' responses and correlated these trends with those found in the news. We found evidence that many trends are similar across essays and the news. However, we also observed some interesting differences. Based on these analyses, we make recommendations in this paper for developers of writing assessments and automated scoring systems.

Key words: trends, time series, writing assessment, automated scoring, computational linguistics

Acknowledgments

We would like to thank Dan Blanchard, Chris Brew, Paul Deane, Diane Napolitano, and Joel Tetreault for their comments on this work.

1 Overview

Many educational assessments aim to measure a test taker's ability to write, and to separate writing ability from content knowledge, it is common to use prompts that ask test takers to discuss a general social issue (e.g., the effects of technology on society). Rubrics for scoring test takers' responses often focus on characteristics such as spelling and grammatical usage, thereby avoiding particular expectations with respect to content; that is, tests of writing ability typically do not care what test takers write about.

However, many important writing skills are closely linked to content. Coherence and argumentation, which are areas of growing interest in automated scoring research (Burstein, Marcu, & Knight, 2003; Burstein, Tetreault, & Andreyev, 2010; Burstein, Tetreault, Chodorow, Blanchard, & Andreyev, in press), depend heavily on content. In fact, it is arguable that such deeper writing constructs cannot be properly analyzed without semantic information.

Thus, even though content is not an important part of generic writing assessments, we believe that it is interesting and productive to study what test takers write about. Understanding how test takers respond to generic issue prompts may help researchers create better automated scoring technologies and may help test developers create fairer, more reliable assessments.

In this report, we present an analysis of topical trends in a large corpus of essays written over a period of approximately 4 years. We are interested in trends related to very specific topics, and so we focused our analyses on the frequencies of individual words (e.g., *president-elect*). We used time series analysis methods to identify topics that exhibit various trends, and we correlated trends in essays with trends in a corpus of news articles over the same time period.

1.1 Issue Writing Prompts

In this section, we provide further details about and examples of the types of writing prompts we study in this report. Issue prompts are generic in that they are not about

specific events or topics. This feature is in contrast to other types of prompts such as those that ask test takers to explain a particular scientific principle (e.g., to demonstrate content knowledge) or to evaluate a given argument (e.g., to demonstrate logical analysis skills).

Consider the issue prompt given in Example 1:

1. In this age of intensive media coverage, it is no longer possible for a society to regard any woman or man as a hero. The reputation of anyone who is subjected to media scrutiny will eventually be diminished.

While the test taker cannot simply write about whatever he or she wishes, the prompt provides considerable freedom with respect to specific examples (unlike, e.g., a prompt such as *Compare and contrast the characters Konstantin and Anna in Leo Tolstoy's* Anna Karenina). For example, a test taker responding to the preceding prompt could discuss any of a number of important historical or contemporary figures.

In this work, as we discuss in more detail later, we found that a major influence on responses and sources of examples is the news. For example, while historical examples such as Example 2 occur, current events also lead to examples such as Examples 3 and 4:

2. Jefferson is considered to be a great hero since he wrote Constitution but today he would be excommunicated because he kept slaves.

3. In essence, people will remember Tiger Woods as the man who had numerous affairs, instead as a great golf legend because of media scrutiny.

4. The way the media protrayed Sarah Palin helped Obama win in a landslide in 2008.

An important point to stress is that examples are not just a superficial aspect of writing. Relevant, specific details and examples are a key element of good argumentative discourse (Perelman & Olbrechts-Tyteca, 1991) and writing (Spandel & Stiggins, 1990). A better understanding of how examples are used could thus provide a better understanding of the writing process.

1.2 Research Questions

In this work, we analyzed a large corpus of essays and investigated the following research questions:

1. Can we use time series analysis methods, such as cross-correlations, to detect different types of topical trends in essays? (section 4.1)
2. Do test takers respond to news events? (section 4.2)
3. On average, how much time elapses between specific news events and their being mentioned frequently in essays? (section 4.2)
4. Are there trends in the news that are not mentioned in essays or trends mentioned in essays but not in the news? (section 4.3)

Specifically, we made the following contributions:

- A large-scale application of time series analysis to a previously unexplored genre
- Evidence that issue essays are strongly affected by news trends
- Exploratory analyses of the differences between the trends found in news stories and essays
- Recommendations for assessment development and automated scoring

Although we leave unanswered the question of *why* test takers respond to some news trends but not others, we believe that these contributions have the potential to lead to improved writing assessments and automated scoring techniques.

2 Data

We used two corpora in our analyses. Because our main objective is to gain knowledge about topical trends in persuasive essay writing, the first corpus was a set of short essays written for a high-stakes assessment aimed at college students. The essay writing task was designed to assess test takers' critical thinking skills. It asked them to

respond in writing to a brief prompt pertaining to a topic of general interest. For example, one prompt might ask test takers to discuss the value of secrecy in politics, another whether technological change has improved the conditions of humanity. The task was timed, and test takers were not given access to the Internet or any other outside materials. The corpus consisted of approximately 2 million essays written in response to 112 different prompts. It also contained time stamps for when each essay was written; the essays spanned the period from August 25, 2006, to September 25, 2010.

As a source of news trends, we used a second corpus of approximately 310,000 *New York Times* (NYT) news stories from the fifth edition of the Gigaword corpus (Parker, Graff, Kong, Chen, & Maeda, 2011). Gigaword includes publication dates for each story, and we included only stories that were from the same time period as the essays described previously. We compared essay trends to these news trends, both for the purpose of externally validating the essay trends and for understanding similarities and differences between the two types of trends. Table 1 provides additional descriptive statistics for each corpus.

Table 1

Descriptive Statistics for the Two Corpora Used in This Report

	Documents	Words	Mean words per document
Essay corpus	1,969,799	939,144,706	476.8
NYT corpus	308,105	230,379,095	747.7

Note. NYT = *New York Times*.

3 Methods for Detecting and Comparing Trends

This section presents the statistical methods we used to study the topical trends in our corpora of essays and news stories. To understand these trends, we measured the prevalence of various topics at different points in time. Because we were interested in the sort of fine-grained topics that relate to specific people and events, we focused on case-normalized word unigrams rather than on the sort of general topic clusters provided

by LDA (Blei, Ng, & Jordan, 2003) or K-means clustering (Shaparenko, Caruana, Gehrke, & Joachims, 2005).

To measure the prevalence over time of a topic represented by the word unigram w, we computed a smoothed version of the document relative frequency for the word (i.e., the proportion of documents containing the word). As a preprocessing step, we converted all characters to lowercase, but we did not perform stemming. Let $d_{w,t}$ be the number of documents (i.e., either essays or news stories) at time t that include the word w, and let d_t be the total number of documents at time t. The raw proportion $d_{w,t}/d_t$ is likely to be bumpy, making informative patterns difficult to identify. Therefore, using weighted moving averages, we smoothed the frequencies at each time step by incorporating information from previous time steps. Specifically, we defined the smoothed proportion of documents at t containing word w as a ratio of weighted moving averages over k time steps, as follows:

$$a_{w,t} = \frac{\sum_{i=0}^{k-1}(k-i)d_{w,t-i}}{\sum_{i=0}^{k-1}(k-i)d_{t-i}}. \tag{1}$$

In our analyses, the time steps were in days, and we chose $k = 30$. We use $\mathbf{a_w}$ to indicate the vector of weighted moving averages over all time steps.[1]

We adapted existing statistical techniques to detect different types of trends in how these smoothed document frequencies change over time. Specifically, we looked for two types of trends: spikes and gradual increases. Later sections provide examples of these two types.

The first measure we used is cross-correlation (Box, Jenkins, & Reinsel, 2008, section 12.1), which provides a measure of the similarity between two time series that is a function of the lag, or shift, z between them. It returns values from -1 for a perfect negative correlation to 1 for a perfect positive correlation. If $z = 0$, then the cross-correlation is equal to the Pearson correlation coefficient. It is computed as follows:

$$\underset{z}{\text{ccf}}(\mathbf{x}, \mathbf{y}) = \frac{\frac{1}{T}\sum_{i=\max(1,-z+1)}^{\min(T-z,T)}(x_{i+z}-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{T}\sum_{i=1}^{T}(x_i-\bar{x})^2}\sqrt{\frac{1}{T}\sum_{i=1}^{T}(y_i-\bar{y})^2}}, \tag{2}$$

where s_x and s_y are estimates of the standard deviations of x and y. A strongly positive $\text{ccf}_z(\mathbf{x}, \mathbf{y})$ indicates that x lags behind y (i.e., that the values of x tend to be similar to

the values of y from previous time steps). One can compute the cross-correlation of a single time series with itself, with different time shifts—this is commonly referred to as the auto-correlation at lag z. We can detect spikes in an individual time series of document frequencies by finding words with high values for autocorrelation at lag $z = 1$. We use acf_z to refer to the autocorrelation function.

Autocorrelation provides a means to detect spikes that occur a few times or at constant intervals, but it is less useful for detecting gradual increases or decreases over time. To detect such gradual changes, we employed the Mann–Kendall τ statistic (Kendall, 1938; Mann, 1945), which is frequently used in environmental science (Yue, Pilon, & Cavadias, 2002), for example, to detect gradual changes in streamflow or in concentrations of pollutants. The Mann–Kendall statistic is based on the Kendall's τ rank correlation between time steps $t = 1, 2, \ldots, T$ and corresponding values x_t at those time steps. A positive τ indicates an increase over time, and a negative τ indicates a decrease over time.

4 Results

This section presents the results of our analyses. First, in section 4.1, we review the methods we used from section 3 to discover spikes and gradually changing trends in our essay data set. In section 4.2, we discuss how we used cross-correlations to measure the delay before news events are mentioned in essays. Finally, in section 4.3, we discuss the words we found with trends in essays and news stories that are very similar and very different.

4.1 Types of Trends

We identified essay vocabulary trends using acf_1, the autocorrelation at 1. We filtered out words that occurred 1,000 times or fewer, computed autocorrelations over moving averages (section 3), and then sorted to find the strongest trends. The top results are shown in Table 2. (Note that the lowest autocorrelations are mostly uninteresting and very close to zero; few were lower than -0.05.)

Table 2

Top Trends in Essays: Spiking Trends in Essays, as Identified by Having the Highest Autocorrelations at 1

Word	Frequency[a]	acf_1
president-elect	1,574	0.8439
haiti	3,389	0.7697
obama	36,908	0.7388
barack	14,706	0.7357
bp	1,169	0.7289
spill	3,247	0.7273
gulf	5,464	0.6652
facebook	9,565	0.6241
china	80,253	0.5854
swine	2,280	0.5813
earthquake	6,615	0.5624
chinese	43,807	0.5604
crisis	48,486	0.5554
even	870,807	0.5504
admittedly	56,121	0.5425
beijing	5,386	0.5402
insofar	13,394	0.5371
twitter	3,196	0.5305
however	971,417	0.5280
sum	147,507	0.5270
...	...	...

Note. acf_1 = autocorrelation function.

[a] The Frequency column shows the number of documents in which a word occurred.

The words with the strongest acf_1 were closely connected to important events that occurred during the time period during which the essays were written (2006–2010). For example, the word with the most spiking document frequency was *president-elect*, which was used frequently for the relatively short period of time between the election and inauguration of Barack Obama. Multiple words related to the 2010 earthquake in Haiti and the 2010 Gulf of Mexico oil spill also appear in the list. Note that function words such as *however* and *insofar* appear because of a particular trend in the test administration schedule itself,

which is not the focus of this report.

Next, we repeated the process using the Mann–Kendall τ statistic instead of the autocorrelation statistic. Table 3 shows the words with the strongest increasing trends, which indicated gradually growing interest in specific topics, as well as decreasing trends, which indicated waning interest. Although there was some overlap with Table 3 (e.g., *obama*, *twitter*), the Mann–Kendall statistic showed more general, longer term trends such as the growth in importance of social networking sites (*facebook*) and smartphones (*iphone*). The recent economic troubles also showed up as a long-term trend (*recession*). We also observed gradually decreasing trends related to Iraq, terrorism, and George W. Bush.

4.2 Lead–Lag Cross-Correlation Analysis

Next, we addressed the question of how long it typically takes for essay writers to respond to news events. To do this, we computed the cross-correlation function between smoothed document frequencies in our essay corpus and the NYT corpus for all words at various values for lead–lag shift z. We then filtered out words that occur 1,000 or fewer times in the NYT corpus or have auto-correlations at 1 in the NYT corpus equal to or less than 0.5. For each value of z, we then averaged over the cross-correlation values for all remaining words. The results are plotted in Figure 1.

Note that even the values for large leads and lags are positive, due to the smoothing created by using moving averages. The key feature of Figure 1, however, is that its peak is at $z = 7$, indicating that, in general, test takers are most likely to use a word related to a news event 7 days after the event is mentioned in the NYT corpus.

4.3 Comparisons of Trends

Next, we identified words with frequency trends that are similar and different between the essay corpus and the NYT corpus. We first filtered out words that occurred 1,000 times or fewer in either corpus.

To identify words that exhibit similar trends, we first filtered out words that have τ values with different signs. Then, for each word, we summed the values for the

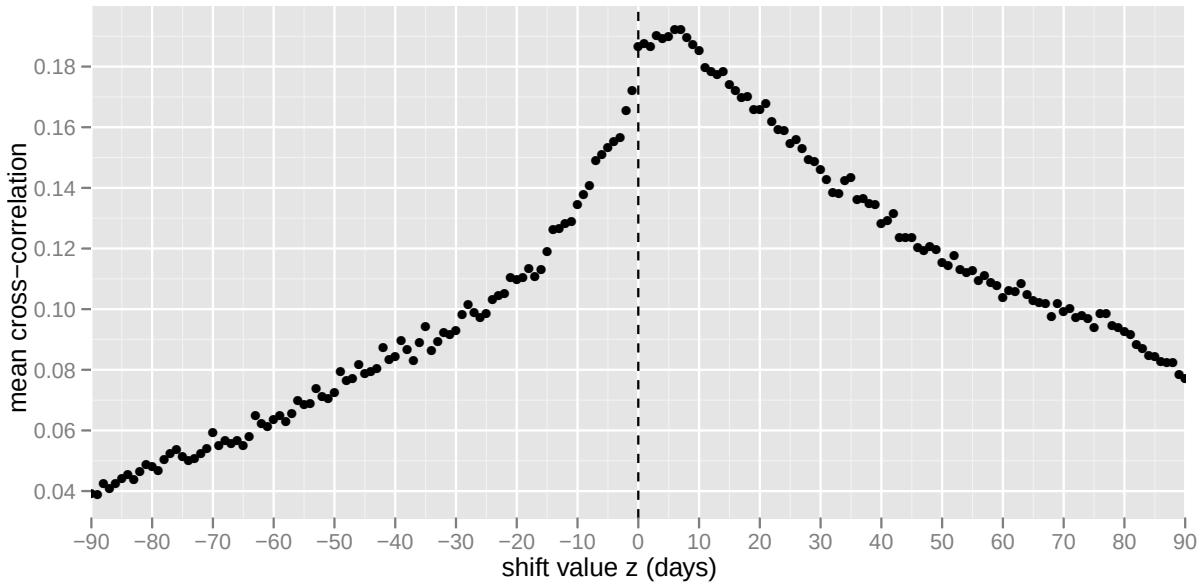

Figure 1. **The cross-correlation function comparing smoothed document frequencies in essays and** *New York Times* **articles. High values on the right side of the plot indicate that essays lag behind news stories.**

Table 3

Top Trends in Essays: Gradually Increasing and Decreasing Trends in Essays, as Identified by Mann–Kendall τ

Word	Frequency[a]	τ
twitter	3,196	0.6266
facebook	9,565	0.6024
obama	36,908	0.5526
texting	4,733	0.4798
barack	14,706	0.4777
recession	14,586	0.4748
iphone	3,672	0.4699
downturn	3,884	0.4069
swine	2,280	0.4057
iphones	1,173	0.3856
...	...	...
preservation	22,273	−0.2283
foley	474	−0.2300
terrorism	21,552	−0.2320
nicole	991	−0.2386
bush	33,878	−0.2396
disciplines	24,251	−0.2398
saddam	5,176	−0.2459
iraqi	5,413	−0.2555
aids	37,930	−0.3002
iraq	50,050	−0.4388

[a]The Frequency column shows the number of documents in which a word occurred.

Mann–Kendall τ statistic from the two corpora. We then sorted by the absolute value of the sum:[2]

$$\text{score}_1(w) = \left|\tau\left(\mathbf{a}_w^{(\text{NYT})}\right) + \tau\left(\mathbf{a}_w^{(\text{essays})}\right)\right|_1 . \tag{3}$$

Recall that $\mathbf{a_w}$ is the vector of moving averages of document frequencies for word w. The results are shown in Table 4.

Many of the trends in essays also appeared in NYT articles: The list of words with

Table 4

Words With Similar Trends in Both Essays and NYT Articles

Word	τ_{NYT}	τ_{essays}
twitter	0.6113	0.6266
obama	0.5756	0.5526
facebook	0.4936	0.6024
recession	0.4874	0.4748
iraq	−0.5120	−0.4388
barack	0.4678	0.4777
iphone	0.3359	0.4699
bush	−0.4978	−0.2396
crisis	0.3888	0.3397
palin	0.3713	0.3570
bailout	0.3942	0.3298
downturn	0.3023	0.4069
stimulus	0.5144	0.1897
iraqi	−0.3946	−0.2555
banks	0.3473	0.2665
unemployment	0.4809	0.1281
recovery	0.4121	0.1771
economy	0.3869	0.1844
reform	0.2575	0.2844
afghanistan	0.3191	0.2110
debt	0.3260	0.2006
said	0.2734	0.2453
...	...	...

Note. Words were sorted by the absolute value of the sum of Mann–Kendall's τ values for each corpus and filtered to include only words with the same sign. NYT = *New York Times*.

similar trends is comparable to Table 3. Social networking, Barack Obama, the financial crisis, and foreign wars were common trending topics. For example, the trends for *facebook* were quite similar, as illustrated by Figure 2.

It is worth pointing out that *facebook* went from only rarely occurring in essays in

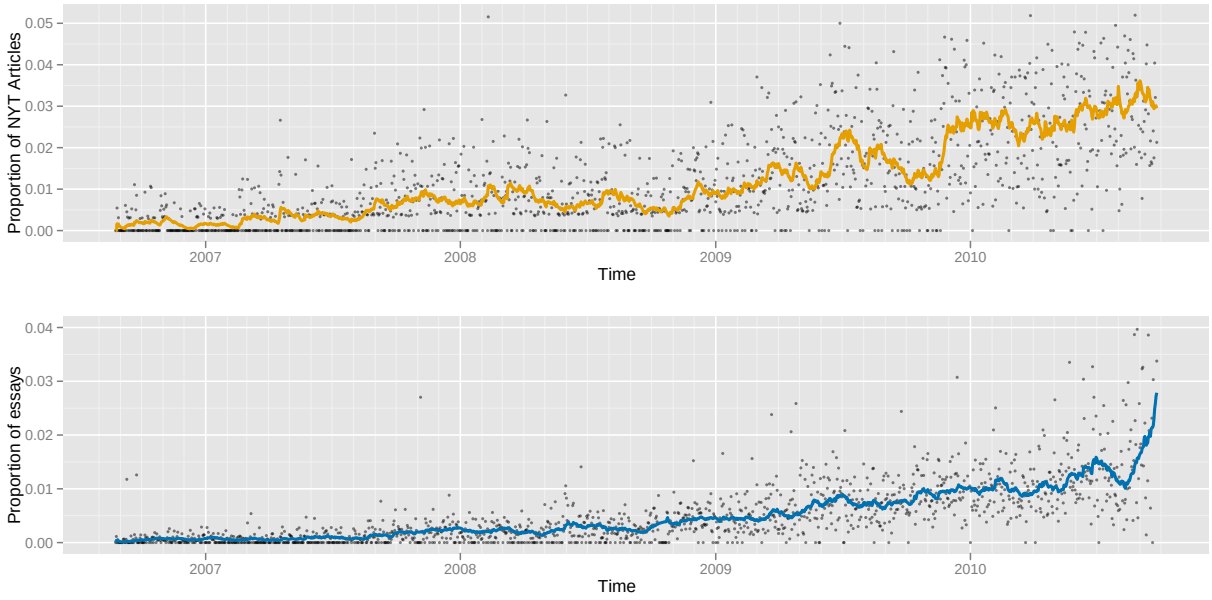

Figure 2. **The smoothed relative document frequency of the term *facebook* over time in essays and the news.**

2006 to occurring in more than 1% of essays in 2010. Specifically, it appeared in only 34 of the 56,793 essays in December 2006 but then in 492 of the 35,044 essays in August 2010.

Next, we found words that have strong trends (either positive or negative) in one corpus but not in the other. To do this, we computed the absolute value of Mann–Kendall τ for each corpus and then subtracted one absolute value from the other. For example, to find words that are trending in essays but not the NYT, we scored each word by subtracting the absolute value of τ for the NYT corpus from the absolute value of τ for the essay corpus:

$$\text{score}_2(w) = \left|\tau\left(\mathbf{a}_w^{(\text{essays})}\right)\right|_1 - \left|\tau\left(\mathbf{a}_w^{(\text{NYT})}\right)\right|_1 \tag{4}$$

$$\text{score}_3(w) = \left|\tau\left(\mathbf{a}_w^{(\text{NYT})}\right)\right|_1 - \left|\tau\left(\mathbf{a}_w^{(\text{essays})}\right)\right|_1 . \tag{5}$$

As shown in Table 5, many of the strong trends in the essay corpus but not in the news corpus were related to technology (*apple*, *youtube*, *online*). Also, we observed some words that were related to specific issue prompts used in the data set (e.g., a prompt about raising children).

Table 5

Words With Trends in Essays but Not NYT Articles

Word	τ_{NYT}	τ_{essays}
networking	0.0598	0.3647
forward	0.0041	0.2812
bring	−0.0650	0.3092
raise	0.0264	0.2621
raised	0.0623	0.2974
raising	0.0431	0.2714
youtube	0.0175	0.2457
instead	−0.0114	0.2368
skepticism	−0.0049	0.2231
online	−0.0218	0.2378
discipline	0.0018	−0.2175
grow	−0.0149	0.2286
tough	0.0010	0.2111
generation	−0.0126	0.2214
accepting	−0.0011	0.2091
preservation	−0.0218	−0.2283
destiny	−0.1076	0.3109
apple	0.0530	0.2519
academic	0.0240	−0.2195
cities	−0.0053	−0.2002
...	...	...

Note. Words were sorted by the difference in absolute values of Mann–Kendall's τ.
NYT = *New York Times*.

Many of the strong trends that appeared in the news but not in essays (Table 6) were related to finance and government (e.g., *stimulus*, *federal*). For example, Figure 3 shows that the word *unemployment* had a somewhat stronger positive trend in NYT articles than in essays.[3]

Table 6

Words With Trends in NYT Articles but Not Essays

Word	τ_{NYT}	τ_{essays}
service	−0.4667	−0.0028
atlanta	−0.4617	−0.0415
government	0.4561	0.0407
unemployment	0.4809	0.1281
com	−0.3414	0.0035
e-mail	−0.3378	−0.0031
federal	0.3337	0.0074
stimulus	0.5144	0.1897
news	−0.3449	0.0364
writes	−0.3140	0.0107
material	−0.3452	0.0440
jobless	0.3776	0.0925
et	−0.3026	−0.0198
officials	0.2756	0.0002
seattle	−0.3312	−0.0596
taliban	0.3489	0.0782
weekends	−0.2800	−0.0094
jobs	0.3132	0.0447
graphics	−0.2948	−0.0302
devil	−0.2646	0.0037
...	...	...

Note. Words were sorted by the difference in absolute values of Mann–Kendall's τ.
NYT = *New York Times*.

Finally, we found words that have trends in opposite directions. We filtered out words whose τ values have the same sign and then took the absolute value of the difference between Mann–Kendall's τ for each corpus:

$$\text{score}_4(w) = \left|\tau\left(\mathbf{a}_w^{(\text{NYT})}\right) - \tau\left(\mathbf{a}_w^{(\text{essays})}\right)\right|_1 . \tag{6}$$

Table 7 presents the results.

It is particularly difficult to draw conclusions about the opposite trends shown in

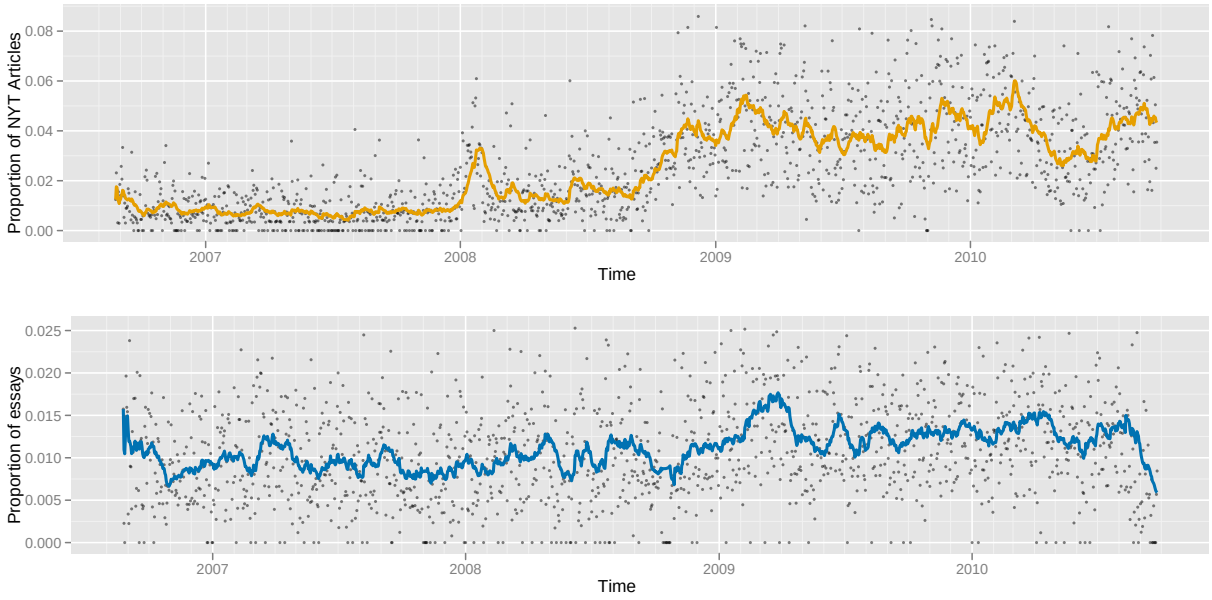

Figure 3. **The smoothed relative document frequency of the term *unemployment* over time in essays and the news.**

Table 7. Some of the differences appear to be related to politics (e.g., *mccain*, *healthcare*), but most appear to be idiosyncratic. For example, the social networking site MySpace was a slightly increasing trend in the essay corpus but a slightly decreasing one in the news corpus, as shown in Figure 4. This result was perhaps because of the growing popularity of social networking sites in general and the use of historical references to MySpace by essay writers. In contrast, MySpace was unlikely to show up in news articles in the time period under investigation.

5 Related Work

Various fields analyze trends in large data sets. Here we provide a brief review of some of the literature.

In computational linguistics, there has been a long line of work on topic detection and tracking (Allan, 2002) as well as analyses of scientific literature (Blei & Lafferty, 2006).

Table 7

Words With Opposite Trends in Essays and NYT Articles

Word	τ_{NYT}	τ_{essays}
edwards	−0.2809	0.2302
story	−0.3411	0.1601
myspace	−0.2598	0.2254
spitzer	−0.2088	0.2416
carolina	−0.3030	0.1269
additionally	−0.2264	0.2020
destiny	−0.1076	0.3109
steve	−0.1252	0.2735
healthcare	−0.1190	0.2733
material	−0.3452	0.0440
news	−0.3449	0.0364
bring	−0.0650	0.3092
mccain	−0.1351	0.2162
illinois	−0.1989	0.1505
kids	−0.1756	0.1706
never	−0.1268	0.2193
com	−0.3414	0.0035
always	−0.1471	0.1907
key	−0.1680	0.1692
learned	−0.0816	0.2493
...	...	...

Note. Sorted by the absolute value of the difference between Mann–Kendall's τ values for each corpus and filtered to include only words with the opposite sign. NYT = *New York Times*.

Recently, there has been considerable interest in other areas, such as political science, particularly related to social media (Hopkins & King, 2009; Leskovec, Backstrom, & Kleinber, 2009; O'Connor, Balasubramanyan, Routledge, & Smith, 2010; Quinn, Monroe, Colaresi, Crespin, & Radev, 2010).

The methods we used in this study for tracking trends related to very specific topics, represented by individual words, could be complemented by models of trends related to

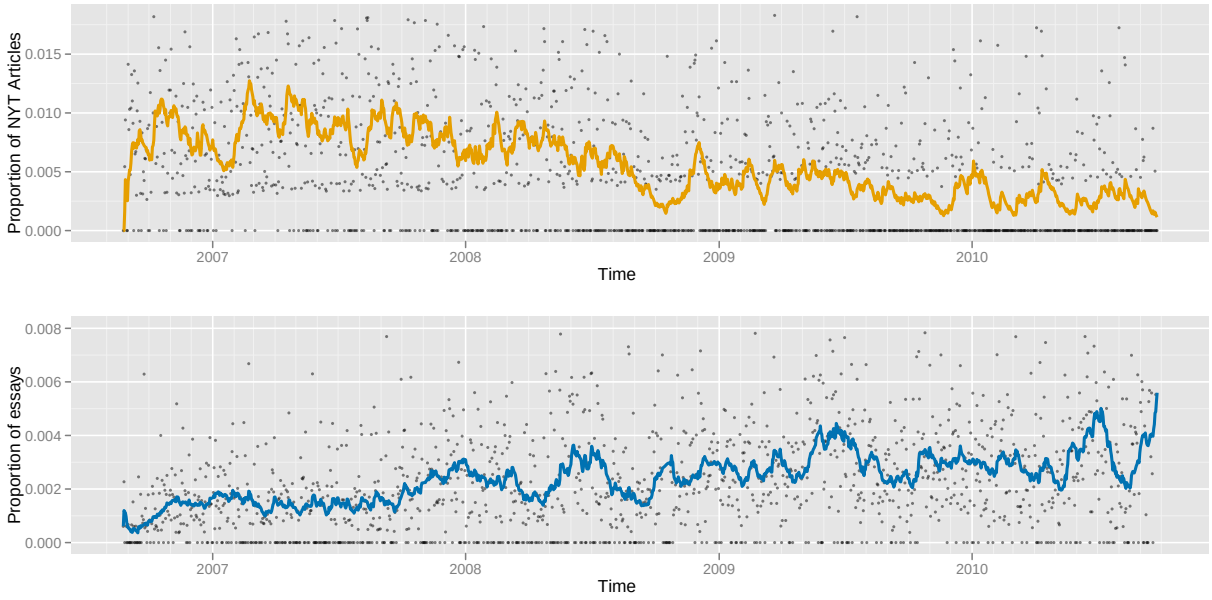

Figure 4. **The smoothed relative document frequency of the term *myspace* over time in essays and the news.**

broader topics, such as the work by Blei and Lafferty (2006). For example, our approach might capture a trend related to the specific word *obama*, whereas the model of Blei and Lafferty might capture a trend related to the broader topic of politics. We leave such analyses to future work.

In economics and finance, trends are a central issue. Identifying patterns in market indicators is a key step in understanding economic change. Moving averages and autocorrelations are widely employed for such analyses (Hamilton, 1994).

In ecological science, tracking trends is also important because scientists need to identify meaningful changes in environmental variables such as the concentration of pollutants in water sources. Methods for detecting gradual changes, such as the Mann–Kendall test, are frequently used (Yue et al., 2002).

Other researchers have studied corpora of educational texts. For example, Boyer et al. (2009) studied transcripts of human–tutor dialogues, and McLaren et al. (2009)

developed machine learning techniques to classify discussion board posts. Researchers have begun to explore large corpora of texts produced by language learners to understand how they make grammatical errors (Granger, 1993; Leacock, Chodorow, Gamon, & Tetreault, 2010). To our knowledge, however, this is the only work that analyzes topical trends in a large corpus of educational texts.

6 Conclusion

In this report, we explored whether time series analyses could detect topical trends in generic issue essays. Rather than producing noisy sets of incoherent trending words, these methods found patterns that were both internally consistent (e.g., there were many trending words related to social networking) and often in agreement with trends in a separate corpus of news events.

We found evidence that essay writers often refer to world events shortly after they are reported in the news. On average, we found a lag between news stories and essays of 7 days.

Comparing trends in essays and the news, we observed many similarities in trends (e.g., related to the 2008 election and to social networking sites). However, we also observed a number of interesting differences. For example, trends related to technology often appeared in essays but not as strongly in the news (e.g., *myspace*, *youtube*, *apple*), whereas trends related to politics and business often appeared in the news but not as strongly in essays (e.g., *unemployment*, *stimulus*).

For developers of automated scoring systems, our findings suggest not only that examples are an important aspect of writing but that recent examples are particularly important. For example, in our corpus, *facebook* appeared only very rarely in essays before 2007 but then showed up in more than 1% of essays by mid-2010. For current automated scoring systems, addressing the changes over time seems important. If trained on old data, an essay scoring model with features for prompt-specific vocabulary usage might expect *clinton* or *newspaper* to be mentioned but not expect *obama* or *blog* and not produce a valid score as a result. Addressing changes over time might be as simple as recreating

word lists to keep vocabularies up to date or retraining natural language processing (NLP) preprocessing components (e.g., named entity recognizers). However, future systems that more deeply analyze coherence and argumentation may need to maintain up-to-date information about the entities that test takers might discuss. Such systems might need to automatically crawl and extract information from news sites, for example, to anticipate test takers' responses.

For creators of writing assessments, if it can be assumed that an essay prompt will be used only in the short term, then it may be desirable to have the prompt address recent events because such prompts would be more likely to elicit better informed and more detailed responses. For general assessments to be used over the long term, it appears that prompts that depend heavily on current events or technologies (e.g., the financial crisis or hybrid cars) should be avoided because those topics may or may not be relevant for extended periods of time. In future work, the methods discussed in this report might also be used to detect shifts in trends for particular prompts. Such shifts might warrant revisions of a rubric or a prompt or even the discontinuation of use of a prompt.

From our analyses, it appears that test takers altered the content of their essays according to news trends. It would be interesting to explore whether human scoring is also affected in some way. For example, a human scorer might not be aware of a recent event mentioned in an essay and thus might not understand its relevance to the essay prompt. This suggests various possible future studies. For example, one could test whether human scores vary according to the difference in time between writing and scoring. One could also explore whether showing a summary of news trends to human scorers would affect their scores.

Another potential area for future work is studying whether geographical factors affect the topics discussed in essays, either independently of or in conjunction with temporal factors. For example, we could study how test takers' nationalities or countries of residence affect the news trends discussed in their essays. Strong geographical effects on essay topics would indicate a need for human scorers to be aware of global news trends, not just local ones.

References

Allan, J. (2002). *Topic detection and tracking: Event-based information organization.* Norwell, MA: Kluwer Academic.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 113–120). New York, NY: ACM Press.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2008). *Time series analysis: Forecasting and control* (4th ed.). Hoboken, NJ: John Wiley.

Boyer, K. E., Ha, E. Y., Wallis, M. D., Phillips, R., Vouk, M. A., & Lester, J. C. (2009). Discovering tutorial dialogue strategies with hidden Markov models. In V. Dimitrova, R. Mizoguchi, B. Du Boulay, & A. Graesser (Eds.), *Frontiers in artificial intelligence and applications: Volume 20. Artificial intelligence in Education—Building learning systems that care: From knowledge representation to affective modelling* (pp. 141–148). Amsterdam, The Netherlands: IOS Press.

Burstein, J., Marcu, D., & Knight, K. (2003). Finding the WRITE stuff: Automatic identification of discourse structure in students essays. *IEEE Intelligent Systems, 18*(1), 32–39.

Burstein, J., Tetreault, J., & Andreyev, S. (2010). Using entity-based features to model coherence in student essays. In *Human language technologies: The 2010 annual conference of the North American chapter of the Association for Computational Linguistics* (pp. 681–684). Stroudsburg, PA: Association for Computation Linguistics.

Burstein, J., Tetreault, J., Chodorow, M., Blanchard, D., & Andreyev, S. (in press). Automated evaluation of discourse coherence quality in essay writing. In M. D. Shermis & J. Burstein (Eds.), *Handbook for automated essay scoring.* New York, NY: Taylor and Francis.

Granger, S. (1993). International corpus of learner English. In J. Aarts, P. de Haan, & N. Oostdijk (Eds.), *English language corpora: Design, analysis and exploitation* (pp.

57–71). Amsterdam, The Netherlands: Rodopi.

Hamilton, J. D. (1994). *Time series analysis*. Princeton, NJ: Princeton University Press.

Hopkins, D. J., & King, G. (2009). A method of automated nonparametric content analysis for social science. *American Journal of Political Science, 54*, 229–247.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika, 30*(1–2), 81–93.

Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2010). Automated grammatical error detection for language learners. *Synthesis lectures on Human Language Technologies, 3*(1), 1–134.

Leskovec, J. K., Backstrom, L., & Kleinber, J. (2009). Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 487–506). New York, NY: ACM.

Mann, H. B. (1945). Nonparametric tests against trend. *Econometrica, 13*(3), 245–259.

McLaren, B. M., Wegerif, R., Mikšátko, J., Scheuer, O., Chamrada, M., & Mansour, N. (2009). Are your students working creatively together? Automatically recognizing creative turns in student e-discussions. In V. Dimitrova, R. Mizoguchi, B. Du Boulay, & A. Graesser (Eds.), *Frontiers in artificial intelligence and applications: Volume 20. Artificial intelligence in Education—Building learning systems that care: From knowledge representation to affective modelling* (pp. 317–324). Amsterdam, The Netherlands: IOS Press.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media* (pp. 122–129). Menlo Park, CA: AAAI Press.

Parker, R., Graff, D., Kong, J., Chen, K., & Maeda, K. (2011). *English Gigaword* (5th ed.). Philadelphia, PA: Linguistic Data Consortium.

Perelman, C., & Olbrechts-Tyteca, L. (1991). *The new rhetoric* (New ed.). South Bend, IN: University of Notre Dame Press.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal*

of Political Science, 54(1), 209–228.

Shaparenko, B., Caruana, R., Gehrke, J., & Joachims, T. (2005). Identifying temporal patterns and key players in document collections. In S. Ma, T. Li, & C. Perng (Eds.), *Proceeedings of the IEEE ICDM Workshop on Temporal Data Mining* (pp. 165–174). Halifax, Nova Scotia, Canada: Saint Mary's University.

Spandel, V., & Stiggins, R. J. (1990). *Creating writers: Linking assessment and writing instruction.* New York, NY: Longman.

Yue, S., Pilon, P., & Cavadias, G. (2002). Power of the Mann–Kendall and Spearman's rho tests for detecting monotonic trends in hydrological series. *Journal of Hydrology, 264*, 254–271.

Notes

[1] This ratio is similar to the volume-weighted moving average from the field of finance, except for the substitution of document proportions for stock prices and total document counts for trading volumes.

[2] We also explored other options for identifying similar words, including a direct comparison using Kendall's τ of the frequencies over time in each corpus, or cross-correlations. However, those approaches did not seem to work quite as well with our data as the approach we used.

[3] The token *et* shows a decreasing trend in the NYT corpus. We are not sure why, but from looking at some of the texts, possible explanations include decreased use of the long form of *et cetera* or fewer mentions of French proper nouns such as *Societe Nationale Industrielle et Miniere*.