TWITTER VS. PRINTED ENGLISH: AN INFORMATION-THEORETIC COMPARISON

Emma Glennon, Lalitha Sankar, and H. Vincent Poor

Department of Electrical Engineering,
Princeton University, Princeton, NJ 08544
{eglennon,lalitha,poor}@princeton.edu

ABSTRACT

The popular social networking and microblogging service Twitter contains language that is very different from what is considered proper. This paper quantifies those linguistic differences between printed English and *Tweetspeak* using information-theoretic concepts. Letter-based n-gram entropies are calculated and compared to analagous data from two corpora of printed English to demonstrate that 1) Twitter's entropy is overall higher than that of printed English, and 2) individual users' entropies are on average higher the less conventional their language use is. The implications for digitally-mediated communication in general are also discussed.

Index Terms— Twitter, computer mediated communication, information theory, information entropy, redundancy

1. INTRODUCTION

Twitter began in 2006 as a microblogging service [1], a way to send messages of no more than 140 characters, known as *tweets*, to all of one's *followers*, generally about the minutiae of one's day. Since then, its use has expanded to broadcasting and following news, communicating with friends, sharing links, and even, as with the Dalai Lama's account, sharing wisdom—all while adhering to the same format. Such a diverse group of Twitter users includes 13 percent of all Internet-using American adults [2].

Twitter is the subject of much concern, including around the question of what it does to language. Criticism of the effects of Twitter on language echo those pertaining to all methods of *digitally-mediated communication (DMC)*. Some consider *Tweetspeak*, the language of Twitter, to be "the voluntary cannibalism of standard English," among other charges (see, for example [3]). However, Twitter also receives praise for encouraging efficiency, creativity, and linguistic evolution [4,5]. Regardless of this debate, Twitter and DMC as a whole undoubtedly have a strong effect on language that also extends offline.

Studying Twitter may be valuable both to examine its individual impact and to better quantify that of DMC in general. Because of its ubiquity and sheer size, Twitter may have a long-lasting impact on both language and the public consciousness. In addition, the broadcast nature of tweets makes it easy to study. Furthermore, its diversity of users and writing styles makes it one of the closest parallels to a comprehensive corpus of printed English, which itself consists of everything from personal letters to bestselling books.

Entropy is a useful quantity for quantifying and comparing languages. This fact is first demonstrated by Shannon in his seminal work [6] in which he quantifies the single (*unigram*) and multi-letter (*n-gram*) entropies of written English and thereby precisely calculates the redundancy of the language at nearly 50%. His methods have proven useful to others: in [7] the entropy of Old English reveals that English once existed in a less redundant form, and in [8] the high entropy of American Sign Language provides a possible solution to the puzzle of why, though signing a word takes longer than speaking it, sentences can be completed just as quickly in both signed and spoken languages.

Likewise, the entropy of Twitter may reveal some of the qualities of the language of DMC in general and Twitter in particular. We observe that the language of Twitter is less redundant and has a higher entropy than printed English and that its entropy is especially high for the users with the least standard English. The paper is organized as follows. In Section 2, we outline our methods of data collection and analysis. We present our results in Section 3 and conclude in Section 4.

2. METHODS

We compute letter-level n-gram entropies up to trigrams. Letter-level entropies relate to the most basic information-theoretic properties of a source: its redundancy, letter-by-letter predictability, susceptibility to noise, and ease of transmission. In contrast, word-level n-gram entropies would provide a more semantic understanding of the qualities of the source, taking into account size and variability of vocabulary. However, we restrict ourselves to letter entropies given

The research was supported by the National Science Foundation under Grants CNS-09-05086 and CCF-10-16671

Twitter's shorthand and creative spelling. Furthermore, we focus on 26-character entropies because the brevity of tweets results in a disproportionate amount of spaces compared to other characters and compared to the printed corpora.

Entropy approaches an asymptotic limit when the size of the source is increased, but the amount of text needed to approach that limit to a statistically acceptable degree increases dramatically when the lengths of n-grams used to measure entropy increase. Unigram entropy for most sources, for example, approaches its limit after only a few hundred characters, while octogram entropy takes about two million to do the same [7]. Therefore we estimate that any entropy up to bigrams will be satisfactorily accurate for the files of individual users, and any entropy up to trigrams will be satisfactorily accurate for larger, concatenated files of tweets.

For a random source X, whose realization $x \epsilon \mathcal{X}$ has probability $p(x)$, the entropy $H(X)$ and redundancy $R(X)$ are defined as follows:

$$H(X) = \sum p(x) \log_2 p(x);\; R(X) = 1 - \frac{H(X)}{\log_2 |\mathcal{X}|}.$$

2.1. Twitter corpora

Because of Twitter's diversity as both a one-to-many and many-to-one channel, we collect two sets of tweets for our Twitter corpus: those of the accounts with the most followers (the most read tweets) and those of average users (the most written tweets), henceforth referred to as *popular tweets* and *random tweets*, respectively. We use an existing interface to collect tweets [9]. We gather popular tweets by harvesting all tweets, or the approximately 3200 most recent tweets if a user had posted more than 3200 at the time of collection, of 150 of the users with the most followers. We use Twitter rankings to select users from most to least popular and only exclude those who frequently tweet in languages other than English or who frequently post unmarked retweets.[1]

We collect the corpus of random tweets by accessing Twitter's public timeline eighteen times, each no closer together than an hour and a half, between a Friday and the following Wednesday. Each time we access the timeline, we select the tweets of the first ten users that fit the same criteria as the popular users (i.e. no foreign languages and few unmarked retweets), again either all or about 3200 of them.

2.2. Standard corpora

We use two complementary corpora of printed English for comparison. The first, Project Gutenberg [10], is a collection of tens of thousands of books and other literary works. Most, but not all, of the works included were written before 1923 due to copyright laws; because of this, any changes in language we detect between the Gutenberg and Twitter samples may not be entirely because of the qualities of DMC but may take into account other changes in modern English that have occurred in the past century. Despite this, Project Gutenberg is a representative body of written English from diverse types of sources, and as such is a satisfactory corpus of printed English up until the advent of DMC. We collect all of Project Gutenberg's English-language text files for a total of 47,575 works.

The second set of non-Twitter data comes from the Corpus of Contemporary American English (COCA), from which we obtain the frequencies of the 500,000 most common words (well over 99% of words in the corpus and any word that appears at least four times out of 422 million words) [11]. This corpus is more modern than Project Gutenberg, including various American English media from 1990 to the present day. One limitation of COCA is that we cannot divide it into individual sources to make it more analagous to the Twitter corpora. Nonetheless, as a huge and balanced modern corpus, it provides a good complement to the Project Gutenberg data.

2.3. Data processing and analysis

2.3.1. *Twitter*

Because DMC's colloquial nature and its frequent unconventional language use are some of its defining features, we assess each user's amounts of non-standard English (NSE)—consisting of abbreviations, acronyms, misspellings (purposeful or otherwise), and improper slang—and divide them into four categories. Users' tweets labeled "*no NSE*" contain only standard spelling with no more than one or two exceptions out of about 30 tweets scanned, while those labeled "*little NSE*" contain one instance of NSE use per four or five tweets, those labeled "*some NSE*" contain such an instance every second or third tweet, and those labeled "*much NSE*" contain NSE in every tweet.

We eliminate URLs, which we view as a DMC-specific punctuation mark, and replace them with a delimiter to preserve the previous and following words' relationships with spaces. To avoid skewing statistics, we also delete retweets.

2.3.2. *Project Gutenberg*

We remove the repetitive header information prefixed to each file to avoid skewing the analysis. We generate 300 datasets from Project Gutenberg. Each dataset is comprised of 159 unique sources (single-author works) such that it is comparable in both size and number of sources to the 180 sets of random tweets and 150 sets of popular tweets.

2.3.3. *Analysis*

We categorize tweets into the popular and random corpora and further categorize them according to the four NSE groups,

[1] *Retweets* are tweets copied from another user and are either prefixed with RT or accompanied by the original poster's photo. The latter are unmarked by text and thus impossible to pick out.

as mentioned earlier. We process the individual and concatenated files to determine each file's unigram, bigram, and trigram frequencies as well as the corresponding entropies.

3. RESULTS

3.1. Corpus observations

After removing retweets and URLs, our corpora collectively contain 677,419 tweets, 48% from the popular and 52% from the random samples. Accounts included in the popular sample are overwhelmingly (slightly over 72%) personal accounts of celebrities, ranging from pop stars and rappers to the Dalai Lama and Barack Obama.[2] The popular corpus also includes the accounts of many companies and news organizations. On the other hand, 85% of accounts in the random corpus discuss personal details. The remaining accounts in both corpora are those of companies and those dedicated to sharing news or links. Random accounts have an average of 783 followers while each of the popular accounts has millions. On average, random accounts also contain much more NSE than do popular ones.

However, both the random and the popular Twitter groups contain highly unconventional language. English words or even whole tweets frequently are written not with Latin letters, but with numbers, punctuation, and other symbols. While instances like these are not taken into account in the entropies given below, this variety of punctuation likely increases the entropy of Twitter compared to standard English. In addition, users who used no NSE are surprisingly rare: even the Dalai Lama frequently used the acronym HHDL for "His Holiness the Dalai Lama."

3.2. Unigrams

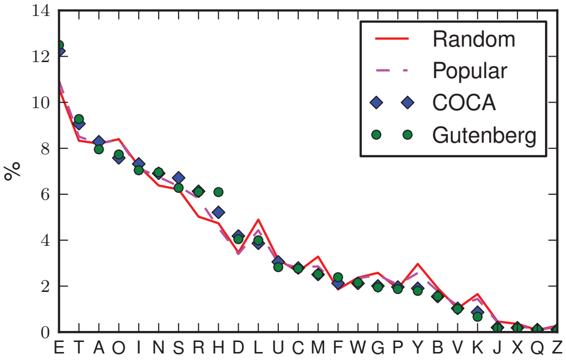

Fig. 1. Unigram frequencies vs. unigrams

In Fig. 1 we plot the unigram frequencies for all four corpora in descending order of the letter probabilities of a randomly-selected Gutenberg sample. The orders and frequencies for the concatenated random and popular corpora are remarkably similar, as are those for the two printed English corpora. Notably, relative to the COCA and Gutenberg plots, both the random and popular curves have distinct peaks, suggesting a consistent difference in usage patterns between tweeted language and written English. Furthermore, the Twitter data demonstrate increased use of letter with low-probabilities in written English and vice versa. The data thereby suggest that Tweetspeak has a higher entropy than standard English, as we show below.

3.3. Non-standard English (NSE)

	Ran. H_1	Ran. H_2	Pop. H_1	Pop. H_2
No NSE	4.2405	3.1403	4.2327	3.3724
Little NSE	4.2441	3.2636	4.2412	3.3818
Some NSE	4.2592	3.2948	4.2545	3.3595
Much NSE	4.2691	3.3654	4.2683	3.4074

Table 1. Average entropies (in bits) for different NSE categories.

In Table 1, we list the average unigram and bigram entropies, denoted by H_1 and H_2, respectively, for both the random (denoted Ran.) and popular (denoted Pop.) groups for each category of NSE. We observe that the number of users in different NSE categories varies, and the largest and smallest groups are different for the two Twitter corpora. In general, because mixing increases entropy [12, p. 36], concatenated sources have a higher entropy than individual files. Thus, to make a fair comparison, we average unigram and bigram entropies of individual users in each NSE category instead of using concatenated files of variable size. With the exception of bigram entropies of the "some NSE" group in the popular sample, the trend is that entropy increases, if only very slightly, with an increasing amount of NSE use.

3.4. Bigram and trigram entropies

n	Random	Popular	COCA	Gutenberg
1	4.2803	4.2600	4.1868	4.1765
2	3.5442	3.4779	3.3049	3.2643
3	2.9500	2.9158	2.7952	2.7103

Table 2. Concatenated n-gram entropies in bits.

In Fig. 2 we plot bigram frequencies for the same concatenated files shown in Fig. 1, again in descending order of the bigram frequencies of the chosen Gutenberg sample. As with unigram frequencies, the curves of the two printed English corpora look remarkably similar, while the plots for the Twitter corpora are less smooth in comparision. In Table 2,

[2] Barack Obama is notable as the only person whose work is included in both the Twitter and Gutenberg corpora. Project Gutenberg contains his inaugural speech.

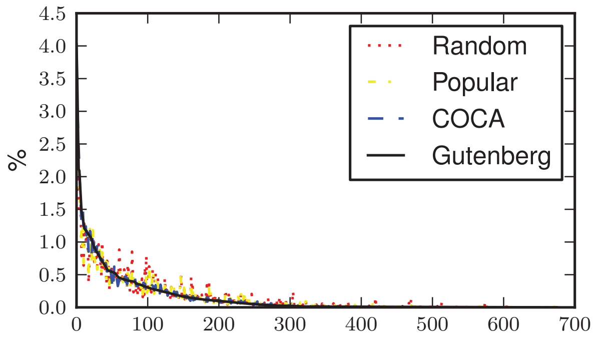

Fig. 2. Bigram frequencies vs. bigram index

n	Random	Popular	COCA	Gutenberg
1	8.9298%	9.3617%	10.9191%	11.1373%
2	24.5915%	26.0021%	29.6830%	30.5473%
3	37.2340%	37.9617%	40.5277%	42.3337%

Table 3. 26-character n-gram redundancy for concatenated sources.

we list all calculated n-gram entropies ($n \leq 3$) for the different corpora. For Project Gutenberg, rather than give the entropy of a single 159-source file, we average the entropies of all 300 such files. We give the same results as percentages of redundancy in Table 3. The entropies for Gutenberg and COCA are close together, as are the entropies of the two Twitter corpora, but the entropies for Twitter are significantly higher than those of standard English. Specifically, COCA's entropy lies between those of Gutenberg and Twitter, which may be accounted for by the sheer number and diversity of sources in COCA as compared to 159-source Gutenberg files. While COCA and Twitter are both contemporary sources, the increased entropy and lower redundancy of Twitter suggests the influence of DMC is stronger than that of modernity alone. Finally, because popular tweets tend to have fewer instances of NSE, their lower entropy compared to random tweets also validates the contribution of this unique aspect (NSE) of DMC to Twitter's higher entropy.

4. CONCLUSION

The results of letter frequency analysis and entropy calculations confirm that language on Twitter is indeed less redundant than other forms of the English language. The differences we found consistently point to this conclusion: the entropy of Twitter is higher than that of standard English, and its higher entropy is related to the frequency of its unique forms of improper language. Increasing efficiency might be in some ways a logical change for users to make to their language when presented with Twitter's immediacy and brevity. That such a change occurred is most likely a sign not of the corruption of language, but of adaptation. While less redundant language is more subject to human interpretation errors, such errors are less likely for short messages.

5. ACKNOWLEDGMENT

The authors would like to thank Daniel Bolan at the University of New Hampshire for his help throughout the stages of data processing and analysis.

6. REFERENCES

[1] "Twitter." [Online]. Available: `http://www.twitter.com`.

[2] A. Smith, "13% of online adults use Twitter," *Pew Internet*, June 2011. [Online]. Available: `http://pewinternet.org/~/media//Files/Reports/2011/Twitter\%20Update\%202011.pdf`. [Accessed June 27, 2011].

[3] D. Ruth, "Twitter versus thought," *New Orleans*, vol. 44, pp. 44–45, May 2010.

[4] D. Craig, "Instant messaging: The language of youth literacy," in *Essays from the Program in Writing and Rhetoric at Stanford University*, pp. 116–133, 2003.

[5] D. Crystal, *Language and the Internet*. Cambridge: Cambridge University Press, 2 ed., 2006.

[6] C. E. Shannon, "Prediction and entropy of printed English," *Bell System Technical Journal*, vol. 30, pp. 50–64, Jan. 1951.

[7] K. O'Brien O'Keefe and W. Rundell, "An information-theoretic approach to the written transmission of Old English," *Computers and the Humanities*, vol. 23, pp. 459–467, Dec. 1989.

[8] A. Chong, L. Sankar, and H. V. Poor, "Frequency of occurrence and information entropy of American sign language," *Computing Research Repository*, Dec. 2009. [Online]. Available: `http://arxiv.org/abs/0912.1768v1`. [Accessed June 8, 2011].

[9] "The tweet collector." *This is Cave*. [Online]. Available: `http://thisiscave.co.uk/tweet/tweetcollector.php`. [Accessed June 10, 2011].

[10] "Project Gutenberg." [Online]. Available: `http://www.gutenberg.org/`. [Accessed June 27, 2011].

[11] M. Davies, "Word frequency data from the Corpus of Contemporary American English (COCA)." [Online]. Available: `http://www.wordfrequency.info`. [Accessed July 14, 2011].

[12] T. M. Cover and J. A. Thomas, *Elements of information theory*. New York: Wiley-Interscience, 1991.