Shared common ground influences information density in microblog texts

Gabriel Doyle
Dept. of Psychology
Stanford University
Stanford, CA, USA, 94305
gdoyle@stanford.edu

Michael C. Frank
Dept. of Psychology
Stanford University
Stanford, CA, USA, 94305
mcfrank@stanford.edu

Abstract

If speakers use language rationally, they should structure their messages to achieve approximately uniform information density (UID), in order to optimize transmission via a noisy channel. Previous work identified a consistent increase in linguistic information across sentences in text as a signature of the UID hypothesis. This increase was derived from a predicted increase in context, but the context itself was not quantified. We use microblog texts from Twitter, tied to a single shared event (the baseball World Series), to quantify both linguistic and non-linguistic context. By tracking changes in contextual information, we predict and identify gradual and rapid changes in information content in response to in-game events. These findings lend further support to the UID hypothesis and highlights the importance of non-linguistic common ground for language production and processing.

1 Introduction

There are many ways express a given message in natural language, so how do speakers decide between potential structures? One prominent hypothesis is that they aim for structures that best convey the intendeed message in the context of the communication. On this view, the use of natural languages is assumed to follow optimal information transmission results from information theory (Shannon, 1948). In particular, speakers should structure their messages to approximate *uniform information density* across symbols (words and phonemes), which is optimal for transmission of information through a noisy channel.

At least three lines of evidence suggest that speakers do make choices to increase the uniformity of information density across their utterances. First, speakers phonologically reduce more predictable material (Aylett and Turk, 2004; Aylett and Turk, 2006; Bell et al., 2003). Second, they omit or reduce optional lexical material in cases where the subsequent syntactic information is relatively more predictable (Levy and Jaeger, 2007; Frank and Jaeger, 2008; Jaeger, 2010). Third, and most relevant to our current hypothesis, speakers appear to increase the complexity of their utterances as a discourse develops (Genzel and Charniak, 2002; Genzel and Charniak, 2003; Qian and Jaeger, 2012). We expand on this finding below.

Following the UID hypothesis, Genzel and Charniak 2002 proposed that $H(Y_i)$, the total entropy of part i of a message (e.g., a word) is constant. They compute this expression by considering X_i, the random variable representing the precise word that will appear at position i, conditioned on all the previously observed words. They then further factor this expression into two terms:

$$\begin{aligned} H(Y_i) &= H(X_i|C_i, L_i) \\ &= H(X_i|L_i) - I(X_i; C_i|L_i) \end{aligned} \quad (1)$$

where the first term $H(X_i|L_i)$ is the dependence of the current word on only the local linguistic context (e.g. within the rest of the sentence L_i) and the second is the mutual information between the current word and the broader linguistic context C_i, given the rest of the current sentence. On their logic, with

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 1587–1596,
Denver, Colorado, May 31 – June 5, 2015.

greater amounts of contextual information, the predictability of linguistic material based on context, $I(X_i|C_i, L_i)$, must go up. Therefore, they predicted that $H(X_i|L_i)$ should also increase, so as to maintain a constant total amount of information.

Genzel and Charniak then approximated $H(X_i|L_i)$ using a number of methods and showed that it did increase systematically in documents. Later work showed that this increase was strongest within paragraphs and was general across document types (Genzel and Charniak, 2003) and languages (Qian and Jaeger, 2012). This work, however, did not attempt to measure shared context (and its influence on message expectations) directly. This challenge is the focus of our current work.

1.1 Contextual effects on complexity

In psycholinguistics, the notion of shared *common ground* is a more precise replacement for the general notion of "context" (Clark, 1996). Common ground is defined as the knowledge that participants in a discourse have and that participants know other participants have, including the current conversational context. A large literature supports the idea that speakers consider referential context and other linguistic common ground in selecting the appropriate expression to refer to a particular physical object (Brennan and Clark, 1996; Metzing and Brennan, 2003; Dale and Reiter, 1995; Sedivy et al., 1999). In principle, Genzel and Charniak's formulation can be considered as capturing the relationship between all of the shared common ground—both linguistic and non-linguistic—and the predictability of language, even though in the previous work only linguistic information was considered.

When there is both linguistic and non-linguistic information passing through the noisy channel, the relevant quantity is not the marginal entropy of only the linguistic stream but the joint entropy of both streams. Let T_j be the linguistic information in part j of the discourse, and E_j be the non-linguistic information in part j. If C_j is the built-up context from the preceding parts $\{1, \cdots, j-1\}$ of the discourse, then we can break down the joint entropy as:

$$\begin{aligned} & H(T_j, E_j|C_j) \\ &= H(T_j|E_j, C_j) + H(E_j|C_j) \\ &= H(T_j|C_j) - I(T_j; E_j|C_j) + H(E_j|C_j) \\ &= H(T_j) - I(T_j; C_j) \\ &\quad - I(T_j; E_j|C_j) + H(E_j|C_j) \\ &= H(T_j) - I(T_j; E_j, C_j) + H(E_j|C_j) \quad (2) \end{aligned}$$

By the UID hypothesis, we expect the left-hand side of this equation, the information content of each part of the discourse, to be constant. The first term of the right-hand side is the out-of-context entropy of the linguistic information. The second term is the mutual information of the linguistic information and the union of the preceding context plus the current non-linguistic information (the events occurring at the time). The third term is the entropy of the non-linguistic information, given the preceding context.

This breakdown suggests that rational participants in a discourse will exhibit both slow and fast adaptation to context in order to maintain overall constant entropy. As context slowly builds, the mutual information term grows (and the non-linguistic entropy likely shrinks), resulting in the time-based increase in $H(T_j)$ that previous work has found. In addition, an individual event can have high or low information content given the context, without having a large effect on the mutual information term. To maintain constant entropy, high-information events should be accompanied by low-information linguistic responses, and vice versa. With an operationalization of shared context, we should be able to observe these two types of adaptation directly, not just via the increasing trend shown in previous work (Genzel and Charniak, 2002; Qian and Jaeger, 2012).

To test this prediction, we leverage Twitter, a popular microblogging service, to operationalize common ground. Because of its structure, Twitter is an ideal platform for this investigation. One common method of using Twitter is to mark messages with hashtags, which serve as ad-hoc categories, allowing anyone interested in a topic to find the messages relevant to that topic. This strategy is especially used when users are commenting on an external event (e.g. a sporting, media, or political event). We focus here on the World Series of baseball, an annual sporting event with large viewership and a single broadcast stream; in this case, the hashtag is `#worldseries`. Hashtagged messages are part of a discourse with extremely limited prior linguistic context, as no two tweeters will have seen the same set of tweets. The total shared context with the au-

dience that can be assumed by the writer of a tweet is the non-linguistic content of the event being hashtagged.

We begin by describing our corpus and our method of calculating linguistic content (by computing entropy within a simple n-gram model). We then investigate gradual changes in word-by-word information content as the event goes on (testing adaptation driven by contextual mutual information in Equation 2, replicating Genzel and Charniak 2002) and rapid changes in the total information content of tweets in response to important in-game events (testing adaptation driven by non-linguistic information in Equation 2). We end by considering control analyses that provide evidence against alternative accounts of our results.

2 Corpus and Methods

2.1 #Worldseries Corpus

Our current analysis looked at tweets during the 2014 World Series, a series of seven baseball games in late October 2014. We obtained these tweets by searching publicly-available tweets through the Twitter API, using an adaptation of SeeTweet (Doyle, 2014) to compile tweets containing the hashtag `#WorldSeries`. To synchronize tweets with game events, we used the Major League Baseball Advance Media XML repository,[1] which contains pitch-by-pitch data including the ongoing state of the game and timestamps at the start of each at-bat. Using this timestamp information, we binned tweets by at-bats so that they could be co-registered with other in-game statistics. These bins extend from the time of the first pitch in an at-bat to the beginning of the next at-bat, and thus provide time for reactions to the events of the at-bat.[2] The mean at-bat length was 2.76 minutes, and there were 512 total at-bats. We limited our analysis to tweets timestamped during one of these at-bats, resulting in a total corpus of 109,207 tweets. Each game had its first pitch at approximately 0008 UTC, and lasted between three and four hours.

Our tweet corpus was compiled from the "gardenhose" Twitter search API, which returns a subset of all relevant tweets. Our searches captured approximately 4% of all relevant tweets; Twitter reported 420,329 relevant tweets during Game 1 of the World Series[3], and our dataset contained 17,538 tweets during the same time period. We address potential confounds from this sampling process in Section 5.2.

2.2 Entropy Computation

Estimating the linguistic information content of each tweet is a key task in this work. Social media text has been described as "bad language" (Eisenstein, 2013): It can be difficult to model due to its idiosyncratic abbreviations, typographic errors, and other non-standard forms. Relevant to our goal of assessing information content, it can also be difficult to create an appropriate training corpus for language models, since the vocabulary and composition of tweets of change rapidly (Eisenstein, 2013).

We attempted to minimize these difficulties in two ways. First, we estimated language models with domain-specific corpora. In particular, for tweets from each game we used a training corpus consisting of the tweets from all the other games. This training set provided a vocabulary and structure that was similar in topic and style to the test set. We removed all punctuation and emoji except word-initial @ and #, which refer to users and hashtags, respectively. Usernames were replaced with *[MENTION]* to reduce sparsity; hashtags were not altered, as these often function as words or phrases within the tweet's syntax. Words with fewer than 5 occurrences in the training corpus were marked as out-of-vocabulary items. We estimated trigram models using a modification of NLTK (Bird, 2006)[4] with Witten-Bell smoothing, and estimated per-word and total entropy for each tweet from these models.

Second, we included tweet length (in characters) as an alternative metric of information content (see Section 5.2). Unless information rate varies sys-

[1] `http://gd2.mlb.com/components/game/mlb/`

[2] We tested a series of potential offset times in case Twitter and MLB used different clocks or at-bats were not long enough to capture reactions. We did not adjust the times as there was no significant increase in the correlation between Leverage Index (Sect. 5.1) and tweet rate for these offsets.

[3] `http://Twitter.com/TwitterData/status/524972545930301440`

[4] Smoothing on n-gram models in NLTK can be inaccurate (see `http://github.com/nltk/nltk/issues/367`), so we used a modified version courtesy of B. C. Roy (personal communication).

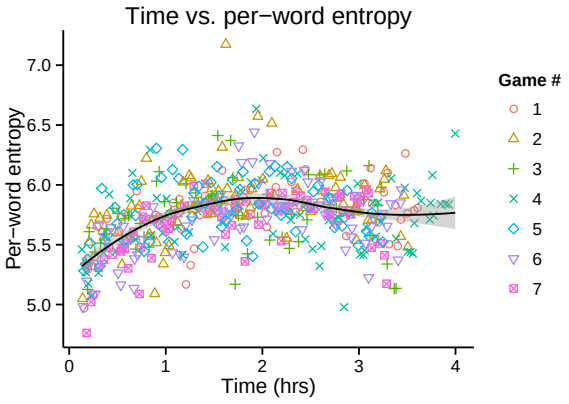

Figure 1: Per-word entropy increases with time for the first two hours of the games, then levels off and slightly declines. Color reflects in-game time; line shows loess fit with 95% confidence intervals.

tematically and substantially across tweets of different lengths—counter to existing results suggesting uniform information density operates at multiple structural levels (e.g., Qian and Jaeger 2009)—longer tweets will generally carry more information.

3 Gradual Changes in Information Rate

Our first analytic goal was to examine changes in the information content of tweets due to the long-term build-up of context in a shared event. We predicted that we would see similar developments in information structure as in more traditional conversational settings, even though there was no formal conversation or explicit linguistic history to develop common ground. Specifically, we predicted that the build-up of contextual information would cause the context-independent per-word entropy to rise over time, replicating the effect that has been observed across languages and genres (Genzel and Charniak, 2003; Qian and Jaeger, 2012).

Figure 1 shows evidence for changes in per-word entropy over the course of games. Per-word entropy rises throughout in the first two hours of each game, slowly levels off and finally declines slightly over time. This pattern is consistent with the constant entropy rate proposal of Genzel and Charniak 2002, and more specifically with the context decay model of Qian and Jaeger 2012.[5]

[5]A late decline in per-word entropy also appeared in Qian and Jaeger 2012's analysis of Swedish.

We used mixed-effects linear regression to quantify this relationship, using the time of an at-bat to predict both per-word and per-tweet entropy during the at-bat. Specifically, we used the logarithm of time as our fixed-effect predictor, per the context-decay models of Qian and Jaeger 2012. We added game-specific random intercepts and slopes of log-time to capture cross-game variation. This model showed significant positive effects of time on entropy, using likelihood-ratio tests for both models (per-word entropy: $.348 \pm .045; p < .001, \chi^2(3) = 104.6$, per-tweet entropy: $10.31 \pm 2.08; p = .001, \chi^2(3) = 74.65$).

We hypothesize that this finding—greater linguistic entropy for later tweets—is due to the accrual of common ground across users from shared non-linguistic information. As they watch more of the game, they share more referents and have stronger expectations about what aspects of the game will be discussed. This shared common ground licenses more complex language and more sophisticated linguistic references. Table 1 gives example tweets at different time points; as a game progresses, references can expand from generic references to the teams or series, to specific individuals and events, and eventually to sequences of events.

While this finding is consistent with previous work on the effect of context, it expands the definition of context. In previous work, the context came from explicit linguistic information built up through paragraphs in a formally-structured, written document. In the Twitter dataset, the context comes from real-world events during the games, as there is no canonical shared sequence of tweets that the tweeters can refer back to (indeed, two random users of the #Worldseries hashtag probably have relatively little Twitter context in common). In sum, contextual influences on entropy need not be explicitly linguistic, so long as discourse participants have reason to believe that the other participants share their knowledge.

4 Fast Changes In Information Content

Intuitively, after an exciting, game-changing event, tweets will be shorter and make more reference to the shared knowledge that this event has just happened. Such events should also generate more re-

Minute	Tweet	Per-word entropy
0	It's finally here! #WorldSeries	4.74
0	#WorldSeries Play Ball	4.96
0	IDEA: @mayoredlee, #SanFrancisco can pledge to throw our @SFGiants an #OrangeOctober parade regardless of #WorldSeries outcome! #SFGiants	8.20
12	The guy with the Marlins sweater is behind home plate again. #worldseries	4.26
12	The Giants 3-0! #WorldSeries	5.43
12	Something about Hunter Pence really, really bothers me. Don't ask me what, cause I havent figured it out, but I don't like him. #WorldSeries	6.64
73	Three HORRIBLE at-bats (mixed in with Cain's walk) prevent Royals from breaking through in the third. #WorldSeries	9.39
130	As Hardy Boy #2, Joe Panik just pulled the mask off of Vargas and discovered it's Old Man Withers from down the street. #WorldSeries	8.12
178	#WorldSeries it's funny the non body names have a great hits. Frm now n on consider the Postseson as Cinderla run. No names needed, #MLB	10.04

Table 1: Example tweets, grouped by minutes since the first pitch.

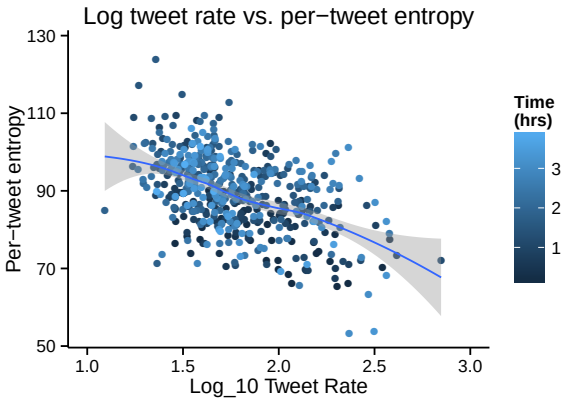

Figure 2: Total tweet entropy plotted against log tweet rate. Color reflects in-game time; line shows loess fit with 95% confidence intervals.

sponses, suggesting that the number of tweets per unit time can serve as a proxy for the information content of an event. This relationship is captured by Equation 2, in which unexpected events have large information content, so linguistic information content should be reduced correspondingly to maintain constant entropy. Our next set of analyses test this relationship.

The examples shown in Table 2 provide anecdotal evidence for the hypothesized relationship between in-game events and linguistic complexity, with examples of consecutive tweets from high-rate and low-rate at-bats, along with their information content. The top triplet comes from one of the highest-rate at-bats, in which Gregor Blanco committed a crucial error in the last inning of the last game. The bottom triplet comes from a low-rate at-bat, mid-game, with one team well ahead of the other; in this case, tweets all refer to different events as there is no single salient shared event.

We quantified the predicted relationship by again fitting a mixed-effect linear regression model, in this case using the logarithm of per-minute tweet rate as a predictor of tweet entropy. Given its significance in the previous model, we included log(time) as a control factor in this analysis, and added by-game random intercepts and slopes for log(rate) and log(time). The log of the tweet rate had a significant negative effect on per-word and per-tweet entropy by likelihood-ratio tests (per-word-entropy: $-.333 \pm .073; p < .001, \chi^2(4) = 59.37$, per-tweet-entropy: $-21.82 \pm 2.43; p < .001, \chi^2(4) = 194.6$).

Log(time) retained significance ($p < .001$) as a predictor for both entropy measures even when rate was accounted for, showing evidence for both

Log rate	Tweet	Per-word entropy
2.49	Holy shitballs, @Royals! #WorldSeries #Game7	3.99
2.49	Just when you thought the #WorldSeries was over.... #E8	4.76
2.49	Fuck you, Blanco. #Giants #WorldSeries	5.54
1.66	Lets Go Giants!!! 5-0 #SFGiants #WorldSeries	3.26
1.66	The guy in Marlins gear behind home plate needs to escorted off property for annoying everybody. #WorldSeries #WhoDoesThat	4.85
1.66	I suppose I appreciate Bochy's "ASG" approach with Bumgarner. Of course, who are any of us to question him in late October? #WorldSeries	7.42

Table 2: Example tweets, grouped by the per-minute tweet rate during each at-bat.

slow and fast adaptation occurring in the discourse. The effects are both in the predicted directions: Entropy increases with time as more informative context builds up, but decreases with tweet rate as more exciting events encourage less information-laden tweets.

5 Control Analyses

5.1 Non-Rate Metrics of Context

Since tweet rate is an organic reflection of the interest accrued by in-game events, it is an important metric for examining fast adaptation. Nevertheless, it could be confounded with other factors influencing tweet production. For instance, there is evidence that online interactions exhibit rational responses to information overload, the state where the amount of incoming information exceeds a user's ability to process it (Miller, 1956; Schoberth et al., 2003). Previous investigations into forum posting behavior have shown that users adapt to overload by posting shorter messages (Jones et al., 2001b; Jones et al., 2001a; Whittaker et al., 2003; Schoberth et al., 2003), and a similar result was found for the more explicitly conversational setting of IRC chat channels (Jones et al., 2008).

To show that the changes in information content are not merely reactions to increased tweet competition—that they have independent informational motivations—we need metrics of event importance and predictability that are not dependent on social media behavior. Luckily, baseball has a long history of statistical analysis, and as a result, there

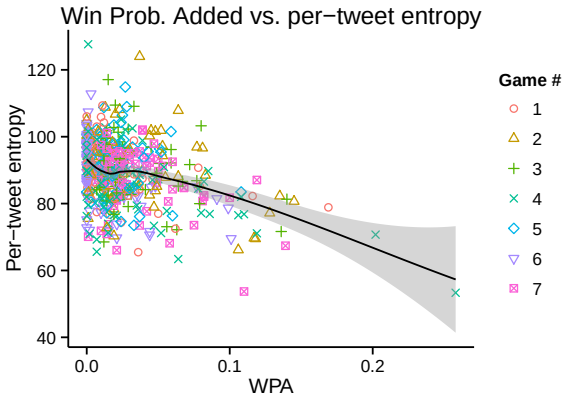

Figure 3: Total entropy decreases for at-bats with greater win probability changes. Loess curve fitting with 95% confidence intervals.

are independently-derived metrics that fit this bill. Two that are appropriate for this purpose are *Leverage Index* (LI)[6] and *Win Probability Added* (WPA) (Tango et al., 2007).

LI is an estimate of how critical an at-bat is to the outcome of the game. It is based on the difference in resultant win probability if the current batter gets a hit or an out, normalized by the mean change in win probability over all at-bats. 1 is the average LI, and greater LI indicates greater importance. LI, as a measure of the expected change in win probability, is similar to non-linguistic entropy term in Equation 2.

WPA depends on the result of an at-bat, and es-

[6] http://www.hardballtimes.com/crucial-situations/

timates how much the win probability changed as a result of what happened during the at-bat. WPA thus provides an estimate of how much information about the game outcome this at-bat has provided, conditioned on the current game context. These measures are well-correlated (Kendall's $\tau = .77$), since a high-LI at-bat's value comes from its ability to affect win probability.

As high LI or WPA values indicate an at-bat whose result has a large effect on the game, these metrics provide an estimate for non-linguistic informativity that is independent of medium-specific influences on tweet production. To assess their effects, we constructed four mixed-effects linear regression models, using LI and WPA to predict per-word and per-tweet entropy in all pairwise combinations (we built separate models for LI and WPA due to their high collinearity). Fixed- and by-game random-effects of log(time) and log(rate) were included as controls in all models; if there is an effect of LI or WPA beyond the effect of rate, this effect can be interpreted as evidence of speaker adaptation to non-linguistic information content.

Both LI and WPA had significant negative effects on per-tweet entropy (LI: $-1.52 \pm .43; p = .001, \chi^2(5) = 20.1$, WPA: $-2.27 \pm .40; p < .001, \chi^2(5) = 44.18$), over and above the effect of tweet rate. Per-word entropy did not show a significant effect of LI or WPA when rate was included as a control factor. Each was a significant factor on per-word entropy ($p = .008, p = .005$) when rate was not included as a control, though, suggesting that the explanatory power of these independent metrics may be subsumed in the more complex factor of tweet rate.

5.2 Speaker Normalization

A second alternative hypothesis for the observed behavioral changes with tweet rate is that they arise not from changes in the behavior of individuals but rather from a change in demographics. It is plausible that rising tweet rates come from an influx of new tweeters using the hashtag, and that these new tweeters simply produce shorter, less informative tweets in general. For instance, spambots often include trending hashtags in their spam tweets (Martinez-Romo and Araujo, 2013). To account for this, we treated the users whose tweets are in our corpus as

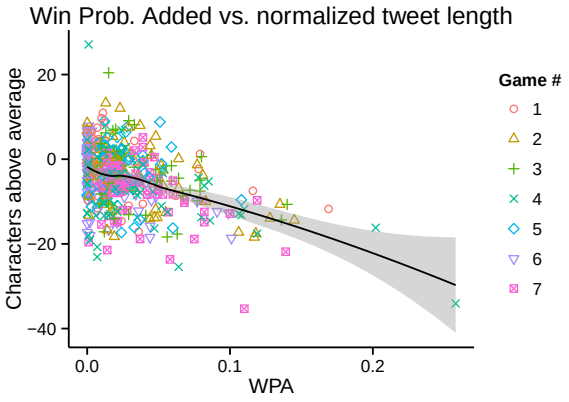

Figure 4: Speaker-normalized tweet length also decreases for at-bats with greater win probability changes. Loess curve fitting with 95% confidence intervals.

a "computational focus group" (Lin et al., 2013; Lin et al., 2014), and used the Twitter API to collected a further 100 tweets from each user outside the timeframe of the games. We used these tweets to estimate an average tweet length for each user, and subtracted this value from the length of their #worldseries tweets during the games.[7] If this baselined metric displays the same effects as shown above, we have reason to believe that users are in fact changing their individual behaviors in response to information factors, rather than that a demographic shift is mimicking a behavioral shift.

For this analysis, we created a mixed-effects model with WPA, log(rate) and log(time) as predictors of tweet length. All three factors were significant (WPA: $-1.64 \pm .36; p < .001, \chi^2(5) = 72.3$; log(rate): $-6.15 \pm .47; p < .001, \chi^2(5) = 303.6$; log(time): $.82 \pm .40; p = .001, \chi^2(5) = 20.6$). We then created a second model using the same factors to predict the mean change in tweet length from the baseline length. Again, all three factors were significant (WPA: $-2.01 \pm .29; p < .001, \chi^2(5) = 70.2$; log(rate): $-5.10 \pm .49; p < .001, \chi^2(5) = 252.6$; log(time): $.61 \pm .35; p = .016, \chi^2(5) = 14.0$). By ruling out demographic shifts (e.g., an influx of terser tweeters), this analysis provides additional support for the idea that tweeters indeed shift their behavior in response to in-game information.

[7] Note that these analyses are conducted over tweet length, rather than total entropy, as there was no obvious way of normalizing entropy by speaker.

6 Discussion

We investigated the hypothesis that speakers optimize their language production so as to approximate *uniform information density*, a signature of efficient communication through a noisy channel (Shannon, 1948; Levy and Jaeger, 2007). Previous work had observed indirect evidence for UID via increases in linguistic complexity (which were hypothesized to reflect increasing discourse/contextual knowledge), but this work neither measured contextual information directly nor included non-linguistic measures of context (Genzel and Charniak, 2002; Genzel and Charniak, 2003; Qian and Jaeger, 2012). Our current work takes a first step towards addressing these issues by using microblog texts around shared events (baseball games) as a case study in which a known context can be characterized more precisely. With this approach, we find systematic differences in information rate and total information content as a function of nonlinguistic factors.

We successfully replicated the effect found in previous work: a gradual increase in entropy rate over the course of individual baseball games. But in addition to this effect, we found a striking pattern of short-timescale changes in total message entropy (reflected in the changing lengths of messages). When in-game events were exciting, unpredictable, and outcome-relevant (hence, highly informative), message length and total entropy went down. This regularity suggests that Twitter users were regulating the information content of their messages relative to the total communicative content of the context more broadly, a prediction that can be derived directly from the UID model.

Our work highlights the importance of non-linguistic context for the informational content of language. This relationship is widely acknowledged in theories of pragmatic communication (Grice, 1975; Sperber and Wilson, 1986; Clark, 1996; Frank and Goodman, 2012), but has been largely absent in information-theoretic treatments of linguistic complexity. The omission of this information has largely been for pragmatic, rather than theoretical, reasons: As Genzel and Charniak 2002 note, it is typically very difficult to compute semantic—let alone non-linguistic—information content. Our work suggests that internet communications surrounding shared media events may be a promising source of grounded language use where context can be quantified more effectively due to the existence of substantial metadata.

A growing literature suggests that the information content of language is the critical variable for understanding processing difficulty in language comprehension (Levy, 2008; Demberg and Keller, 2008; Boston et al., 2008; Smith and Levy, 2013). Under surprisal theory (Hale, 2001; Levy, 2008), the overall predictability of individual elements of language is assumed to be due to a predictive model of its likelihood in the current context. Given this model of processing difficulty, our work here makes a strong prediction: that the information processing difficulty of a word or sentence should track with its total information content (including its relationship to the non-linguistic context), rather than its linguistic information content alone. Some preliminary evidence supports this idea. In a study of the processing complexity of negative utterances, Nordmeyer and Frank 2014 found that the processing cost of negation was predicted by the surprisal of encountering the negation in a particular pragmatic context. But future work should test this hypothesis across a wider variety of structures and contexts.

In sum, our work contributes to the growing body of evidence in favor of the UID hypothesis. The mechanisms underlying the tendency to regulate information content are still unknown, however. While UID would follow from a strong form of *audience design*, in which speakers explicitly consider the processing difficulty of different content (Clark, 1996), the UID hypothesis could also emerge from simpler production processes. Untangling these possibilities will not be trivial. Regardless of the resolution of this issue, however, UID appears to be an important descriptive tool in capturing how speakers make production choices.

Acknowledgments

We gratefully acknowledge the support of ONR Grant N00014-13-1-0287.

References

Matthew Aylett and Alice Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1):31–56.

Matthew Aylett and Alice Turk. 2006. Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *The Journal of the Acoustical Society of America*, 119(5):3048–3058.

Alan Bell, Daniel Jurafsky, Eric Fosler-Lussier, Cynthia Girand, Michelle Gregory, and Daniel Gildea. 2003. Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustical Society of America*, 113(2):1001–1024.

Steven Bird. 2006. NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics.

Marisa Boston, John Hale, Reinhold Kliegl, Umesh Patil, and Shravan Vasishth. 2008. Parsing costs as predictors of reading difficulty: An evaluation using the potsdam sentence corpus. *Journal of Eye Movement Research*, 2(1):1–12.

Susan E Brennan and Herbert H Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482.

Herbert H Clark. 1996. *Using language*, volume 1996. Cambridge University Press Cambridge.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Gabriel Doyle. 2014. Mapping dialectal variation by querying social media. In *Proceedings of the European Chapter of the Association for Computational Linguistics*.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369.

Michael C Frank and Noah D Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Austin Frank and T Florian Jaeger. 2008. Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, pages 933–938. Cognitive Science Society Washington, DC.

Dmitriy Genzel and Eugene Charniak. 2002. Entropy rate constancy in text. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 199–206. Association for Computational Linguistics.

Dmitriy Genzel and Eugene Charniak. 2003. Variation of entropy and parse trees of sentences as a function of the sentence number. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 65–72. Association for Computational Linguistics.

H Paul Grice. 1975. Logic and conversation. *Syntax and Semantics*, 3:41–58.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, pages 1–8. Association for Computational Linguistics.

T Florian Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1):23–62.

Quentin Jones, Gilad Ravid, and Sheizaf Rafaeli. 2001a. Empirical evidence for information overload in mass interaction. In *CHI'01 Extended Abstracts on Human Factors in Computing Systems*, pages 177–178. ACM.

Quentin Jones, Gilad Ravid, and Sheizaf Rafaeli. 2001b. Information overload and virtual public discourse boundaries. In *INTERACT'01: 13th International Conference on Human-Computer Interaction*, page 43. IOS Press.

Quentin Jones, Mihai Moldovan, Daphne Raban, and Brian Butler. 2008. Empirical evidence of information overload constraining chat channel community interactions. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, pages 323–332. ACM.

Roger Levy and T Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems*, pages 849–856.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Yu-Ru Lin, Drew Margolin, Brian Keegan, and David Lazer. 2013. Voices of victory: A computational focus group framework for tracking opinion shift in real time. In *Proceedings of the 22nd international conference on World Wide Web*, pages 737–748. International World Wide Web Conferences Steering Committee.

Yu-Ru Lin, Brian Keegan, Drew Margolin, and David Lazer. 2014. Rising tides or rising stars?: Dynamics of shared attention on twitter during media events. *PLoS One*, 9(5):e94093.

Juan Martinez-Romo and Lourdes Araujo. 2013. Detecting malicious tweets in trending topics using a statistical analysis of language. *Expert Systems with Applications*, 40(8):2992–3000.

Charles Metzing and Susan E Brennan. 2003. When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, 49(2):201–213.

George A Miller. 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63(2):81.

Ann E Nordmeyer and Michael C Frank. 2014. A pragmatic account of the processing of negative sentences. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*.

Ting Qian and T Florian Jaeger. 2009. Evidence for efficient language production in Chinese. In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*.

Ting Qian and T Florian Jaeger. 2012. Cue effectiveness in communicatively efficient discourse production. *Cognitive Science*, 36(7):1312–1336.

Thomas Schoberth, Jennifer Preece, and Armin Heinzl. 2003. Online communities: A longitudinal analysis of communication activities. In *Proceedings of the 36th Annual Hawaii International Conference on System Sciences*, pages 10–18. IEEE.

Julie C Sedivy, Michael K Tanenhaus, Craig G Chambers, and Gregory N Carlson. 1999. Achieving incremental semantic interpretation through contextual representation. *Cognition*, 71(2):109–147.

Claude E Shannon. 1948. Bell system tech. j. 27 (1948) 379; ce shannon. *Bell System Tech. J*, 27:623.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and Cognition*. Harvard University Press, Cambridge, MA.

Tom M Tango, Mitchel G Lichtman, and Andrew E Dolphin. 2007. *The book: Playing the percentages in baseball*. Potomac Books, Inc.

Steve Whittaker, Loen Terveen, Will Hill, and Lynn Cherny. 2003. The dynamics of mass interaction. In *From Usenet to CoWebs*, pages 79–91. Springer.