Audience size and contextual effects on information density in Twitter conversations

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

The "uniform information density" (UID) hypothesis proposes that language producers aim for a constant rate of information flow within a message, and research on monologue-like written texts has found evidence for UID in production. We consider conversational messages, using a large corpus of tweets, and look for UID behavior. We do not find evidence of UID behavior, and even find context effects that are opposite that of previous, monologue-based research. We propose that a more collaborative conception of information density and careful consideration of channel noise may be needed in the information-theoretic framework for conversation.

1 Introduction

Linguistic communication can be viewed from an information theoretic standpoint as communication via a noisy channel. If humans are approximately rational in their communications and the noisy channel model is appropriate, then we expect to see communication follow an approximately constant rate of information flow. This is the Uniform Information Density (UID) hypothesis.

Evidence in favor of UID has been found in many levels of language production. At the level of within-sentence context, there is clear evidence from phonology that speakers reduce more predictable sounds (Aylett and Turk, 2004; Aylett and Turk, 2006; Bell et al., 2003; Demberg et al., 2012), suggesting that they are giving more "air time" to less predictable material to equalize information density. And in syntax, speakers tend to drop optional materials (like the word "that" as a sentence-complementizer) in more predictable scenarios (Levy and Jaeger, 2007; Frank and Jaeger, 2008; Jaeger, 2010), again implying a process of allocating communication time relative to predictability. These effects appear in both monologues and dialogues, suggesting that local linguistic context shapes message complexity.

There is also some evidence for UID based on broader, discourse-level context. Genzel and Charniak (2002) showed that word-by-word complexity (measured by a standard n-gram language model) increases across sequences of sentences. They hypothesized that this increase was due to a corresponding increase in non-linguistic information that would make even more complex linguistic structures easier to predict. Follow-ups have shown that this same complexity increase effect is attested in different document types and across languages (Genzel and Charniak, 2003; Qian and Jaeger, 2012). However, these studies draw almost exclusively from long, well-structured written texts that function as monologues from writer to reader.

This leaves an important gap in these tests of the UID hypothesis: little work has looked at the influence of discourse-level context on information structure in interpersonal dialogue, the archetype of human linguistic communication. With the exception of one preliminary study that provided a partial replication of the original complexity increase effect using the Switchboard corpus (Vega and Ward, 2009), to our knowledge no work has explored how the broader dynamics of conversation interact with UID.

Proceedings of CMCL 2015, pages 19–28,
Denver, Colorado, June 4, 2015.

The present study applies information-theoretic analysis to a corpus of social media microblog posts that include a large number of natural dialogues. Surprisingly, we do not see clear evidence of the UID hypothesis in these dialogues. Instead, we propose that differences in the discourse-level structure of conversation compared to monologues, such as the desire to establish that mutual understanding has been reached, may interfere with attaining UID in the standard formulation. A more collaborative view of UID, encompassing content generation and grounding (Clark and Schaefer, 1987), may be needed to fully represent conversational structure.

1.1 Conversations, context, and content

One common motivation for the UID hypothesis is a rational analysis based on a noisy-channel model of communication (Levy and Jaeger, 2007).[1] In the noisy-channel analysis, the amount of noise in the channel sets an optimal value for information density to obtain fast, error-free transmission. For a noise level α, we will refer to the optimal information content per discourse unit Y_i as $H_\alpha(Y_i)$. Discourse units, depending on the analysis, range from syllables to whole documents; in our analyses, we focus on words and tweets as our discourse units.

In the course of a message, as argued by Genzel and Charniak (2002), the actual information content per discourse unit is predicted by the entropy of the random variable X_i representing the precise word choice or choices within the discourse unit, conditioned on the available context. The precise extent of this context is difficult to pin down.

We estimate context for our studies by thinking in terms of the *common ground* that a rational speaker believes to exist, given the expected audience of their message. 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 (Clark, 1996). This common ground can be built from a combination of linguistic and non-linguistic context, including previous messages within the discourse, preceding interactions between the conversation participants, and world knowledge.

[1]The other common motivation is a surprisal-based argument (Levy, 2008): maintaining UID also minimizes the listener's comprehension effort.

To formalize this relationship, let C_i be the common ground that exists prior to the production of discourse unit Y_i, and let α be the expected noise level in the channel that Y_i is transmitted through. Then optimality within a noisy channel model predicts that the noise-dependent optimal information rate $H_\alpha(Y_i)$ is related to the actual information rate as follows:

$$H_\alpha(Y_i|C_i) = H(X_i) - I(X_i; C_i) \qquad (1)$$

Here, $H(X_i)$ is the apparent decontextualized entropy of the discourse unit independent of the common ground. This quantity is often estimated from a language model that uses only local context, not higher-level discourse context or common ground. We use a trigram Markov model in this study.

$I(X_i; C_i)$ is the mutual information of the discourse unit random variable X_i and the common ground C_i—essentially how much more predictable the next discourse unit becomes from knowing the common ground. Common ground is difficult to quantify—both in the particular datasets we consider and more generally—so we rely on the assumption that more common ground is correlated with greater mutual information, as in Genzel and Charniak (2002).

Then, based on this assumption, Eq. 1 allows us to make two UID-based predictions. First, as channel noise increases, transmission error should increase, which in turn should cause the optimal information transfer rate $H_\alpha(Y_i)$ to decrease. Thus, to maintain equality with rising noise, the apparent entropy $H(X_i)$ should decrease. This prediction translates into communicators "slowing down" their speech (albeit in terms of information per word, rather than per unit time) to account for increased errors.

Second, as common ground increases, $I(X_i; C_i)$ should increase. To maintain equality with rising common ground, $H(X_i)$ should thus also increase, so as not to convey information slower than necessary. This prediction translates into communicators "going faster" (e.g., packing more information into each word) because of an assumption that listeners share more common ground with them.

1.2 The current study

We take advantage of the conversational structure of the popular social media microblogging platform Twitter (`http://twitter.com`) to test these predictions. Twitter allows users to post 140 character "tweets" in a number of different conversational contexts. In particular, because some tweets are replies to previous tweets, we can use this reply structure to build conversational trees, and to track the number of participants. In addition, specific choices in tweet production can affect what audience is likely to see the tweet. These variables are discussed in depth in Section 2.2.

To test the entropy effects predicted by Eq. 1, we first examine different types of tweets that reach different audience sizes. We then restrict our analysis to reply tweets with varying audience sizes to analyze audience size independently of noise. Finally, we look at the effects of common ground (by way of conversation structure) on tweet entropy. Contrary to previous UID findings, we do not see a clear increase in apparent entropy estimates due to more extensive common ground, as had been found in previous non-conversational work (Genzel and Charniak, 2002; Qian and Jaeger, 2012; Doyle and Frank, 2015).

We propose two factors that may be influencing conversational content in addition to UID factors. First, achieving conversational goals may be more dependent on certain discourse units that carry low linguistic informativity but substantial social/conversational importance. Second, considering and adapting to two different types of noise—message loss and message corruption—may cause tweeters to make large-scale decisions that overwhelm UID effects.

2 Corpus

Randomly sampling conversations on a medium like Twitter is a difficult problem. Twitter users routinely use the medium to converse in smaller groups via the mention functionality (described in more detail below). Yet such conversations are not uniformly distributed: A random sample of tweets—perhaps chosen because they contain the word "the" or a similarly common token (Doyle, 2014)—yields mostly isolated tweets rather than complete dialogues. Dialogues depend on users interacting back and forth within communities.

Seed Users	Category
`@camerondallas` `@rickypdillon`	Youtube stars
`@edsheeran` `@yelyahwilliams`	Musicians
`@felixsalmon` `@tanehisicoates`	Journalists
`@jahimes` `@jaredpolis` `@leezeldin`	Politicians
`@larrymishel` `@paulnvandewater`	Economists
`@neiltyson` `@profbriancox` `@richardwiseman`	Scientists

Table 1: Seed users for our dataset.

2.1 Seed strategy

To sample such interactions, we developed a "seed" strategy where we identified popular Twitter accounts and then downloaded a large sample of their tweets, then downloaded a sample of the tweets of all the users they mentioned. This strategy allowed us to reconstruct a relatively dense sample of dialogues (reply chains).

We began by choosing a set of 14 seed Twitter accounts (Table 1) that spanned a variety of genres, were popular enough to elicit replies, and interacted with other users often enough to build up a community.

To build conversations, we needed to obtain tweets directed to and from these seed users. For each seed user, we downloaded their last 1500 tweets, extracted all users mentioned within those tweets, and downloaded each of their last 1500 tweets. To capture tweets that failed to start conversations with the seed users, we also added the last 1000 tweets mentioning each seed user's handle. Tweets that appeared in multiple communities were removed. Each reply contains the ID of the tweet it replies to, so we could rebuild conversation trees back to their roots, so long as all of the preceding tweets were made by users in our communities.

2.2 Conversation structure and visibility

Twitter conversations follow a basic tree structure with a unique root node. Each tweet is marked as a reply or not; for replies, the user and tweet IDs of the tweet it replies to is stored. Each tweet can be a reply to at most one other tweet, so a long conversation resembles a linked list with a unique root node. "Mentions," the inclusion of a username in a tweet, are included in tweets by default throughout a conversation unless a tweeter chooses to remove some of them, so tweets deep in a conversation may be primarily composed of mentions rather than new information.

After some processing described below, our sampling process resulted in 5.5 million tweets, of which 3.3 million were not part of a conversation (not a reply, and received no replies). Within this data, we found 63,673 conversations that could be traced back to a root tweet, spanning 228,923 total tweets. Unfortunately, Twitter only tracks replies up a tree, so while we know with certainty whether a tweet is a reply (even if it is to a user outside our communities), we do not know with certainty that a tweet has received no replies, especially from users outside our communities. If anything, this fact makes our analyses conservative, as they may understate differences between reply and non-reply tweets. The remaining 2 million tweets were replies whose conversations could not be traced back to the root.

2.3 Information content estimation

To estimate the information content of a tweet, we first tokenized the tweets using Twokenizer (Owoputi et al., 2013). We then removed any number of mentions at the beginning or end of a tweet, as these are usually used to address certain users rather than to convey information themselves. (Tweets that only contained mentions were removed.) Tweet-medial mentions were retained but masked with the single type *[MENTION]* to reduce sparsity. Links were similarly masked as *[URL]*. Punctuation and emoji were retained. We then built trigram language models using SRILM with default settings and Kneser-Ney discounting. Types with fewer than 5 tokens were treated as out-of-vocabulary items.

For each community, the training set was the set of all tweets from all other communities. This training set provides tweets that are contemporaneous to the test set and cover some of the same topics without containing the same users' tweets.

3 Analyses

We describe the results of three sets of analyses looking at the influence of audience size and available context on apparent tweet entropy. The first examines the effect of expected audience size at a coarse level, comparing tweets directed at a small subset of users, all one's followers, or the wider realm of a hashtag. The second examines the effect of finer differences in known audience size on apparent informativity. The third examines the effects of conversational context and length on informativity.

3.1 Expected audience size

First, we consider three different types of tweets and their expected audience size. Tweets whose first character is a mention (whether or not it is a reply) do not show up by default when browsing a user's tweets, unless the browser follows both the tweeter and first-mentioned user.[2] We will refer to these as "invisible" tweets as they are invisible to followers by default. A tweeter making an initial-mention tweet thus should expect such a tweet to have a relatively limited audience, with a focus on the mentioned users.[3]

On the other side, a hashtag serves as a categorization mechanism so that interested users can discover new content. Hashtags are often used to expand the potential audience for a tweet to include the feeds of users tracking that hashtag, regardless of whether they follow the original tweeter, and so a tweeter using a hashtag should expect a larger audience than

[2]This behavior varies slightly depending on what application is used to view Twitter. On the website, mention-first tweets do not appear in lists and only appear after clicking the 'tweets & replies' option on a timeline. On the Twitter mobile app, mention-first tweets appear by default on a timeline but still not in lists.

[3]Some Twitter users consciously manipulate audience using these markers: many tweets have an initial period or other punctuation mark to prevent it from being hidden. Some users routinely switch between initial-mention replies and "dot"-replies in the course of a conversation to change the audience, presumably depending on their estimate of the wider relevance of a remark.

Type	Tweet	Per-word entropy
invisible	[MENTION] [MENTION] this is so accurate tho	6.00
	[MENTION] can you come to my high school ? ;3	7.61
	[MENTION] Hi Kerry , Please send us your email address in order to discuss this matter further . Thanks !	8.58
baseline	post your best puns in the comments of my latest instagram photo : [URL]	7.44
	I wish I could start a blog dedicated to overly broad and sweeping introductory sentences	9.98
	this new year's eve in NYC , keep an eye peeled 4 Sad Michael Stipe . [URL] already found him : [URL]	7.17
hashtagged	I will probably be quitting my job when #GTAV comes out	7.63
	#UMAlumni what is the number one thing graduating seniors should know ? #MGoGrad	6.80
	Brilliant interactive infographic : shows cone of uncertainty for #climatechange [URL] #howhotwillitget	12.1

Table 2: Example tweets from each category.

normal.[4] Finally, we have baseline tweets which contain neither mentions nor hashtags and whose expected audience size is approximately one's followers.

Intuitively, common ground is higher for smaller audiences. It should be highest for the invisible tweets, where the audience is limited and has seen or can readily access the previous tweets in the conversation. It should be lowest for the hashtagged tweets, where the audience is the largest and will likely contain many users who are completely unfamiliar with the tweeter. If contextualized UID is the driving force affecting information content, then the invisible tweets should have the highest entropy and hashtagged tweets should have the lowest.

In this analysis, we use the full 5.5 million tweet database. Figure 1 plots the entropy of tweets for these three audience sizes. Per-word and per-tweet entropy both significantly *increase* with expected audience size ($p < .001$ by likelihood-ratio test), the opposite direction of our prediction. We discuss this finding below in the context of our next analyses.

[4]Not all hashtags are intended for categorization; some are used for emphasis or metalinguistic comment (e.g. #notmyfavoritefridaymeal, #toomuchinformation). These comments are probably not intended to broaden the tweet's audience. The presence of such hashtags should, if anything, cause our analysis to underestimate variability across audience types.

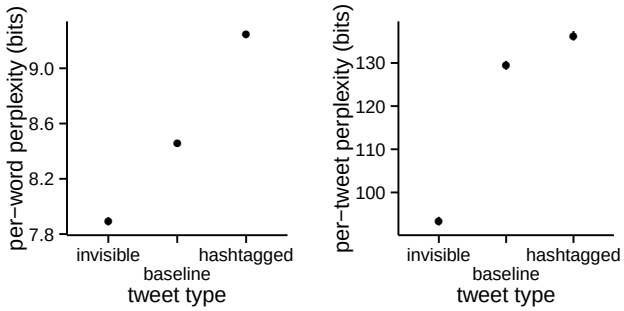

Figure 1: Per-word (left) and per-tweet (right) entropy are higher for tweets with larger expected audience size. Error bars (in some cases smaller than plotting marker) show by-user 95% confidence intervals.

3.2 Known audience size

The results from expected audience size in Section 3.1 have a potential explanation: different tweet types are received and viewed in different ways, which may encourage different kinds of communicative behavior. Tweets with mentions are highly likely to be seen by the mentioned user (unless the mentioned user is very popular), whereas the likelihood of a given hashtagged tweet being seen through the hashtag-searching mechanism is very low. This uncertainty about audience may lead a rational tweeter to package information into tweets differently: they may include more redundant information across tweets when the likelihood of any

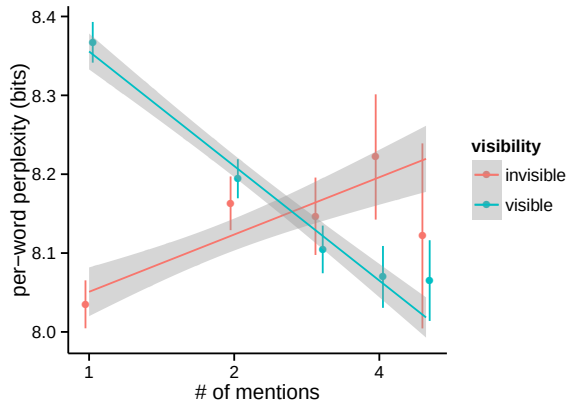

Figure 2: Per-word entropy of tweets with different numbers of mentions and different visibility. Invisible tweets' entropy increases with mentions, while visible tweets' entropy decreases. Logarithmic fits with 95% confidence intervals; x-axis is log-scaled.

given tweet being read is low.

To assess audience size effects in a more controlled setting, we look at invisible tweets with varying numbers of mentions. Invisible tweets provide a quantifiable audience size; those with few mentions have a smaller audience than those with more mentions. Visible tweets, on the other hand, have approximately the same audience size regardless of the number of mentions, since all of a user's followers can see them. Visible mentions can be used for a wide range of discourse functions (e.g., self-promotion, bringing one's followers into an argument, entering contests), and so we do not have a clear prediction of their behavior. But invisible mentions should, under the UID hypothesis, show decreased common ground as the number of conversation participants grows and it is harder to achieve consensus on what all participants know.

Figure 2 shows that the per-word entropy of invisible tweets goes up with the logarithm of the number of mentions. We look only at tweets with between one and five mentions, as invisible tweets must have at least one mention, and five mentions already substantially cut into the 140-character limit.[5] This leaves 1.4 million tweets.

The fact that invisible tweet entropy increases

[5]Usernames can be up to 15 characters (plus a space and an symbol per mention); even if each username is only 7 characters, five mentions use almost one-third of the character limit.

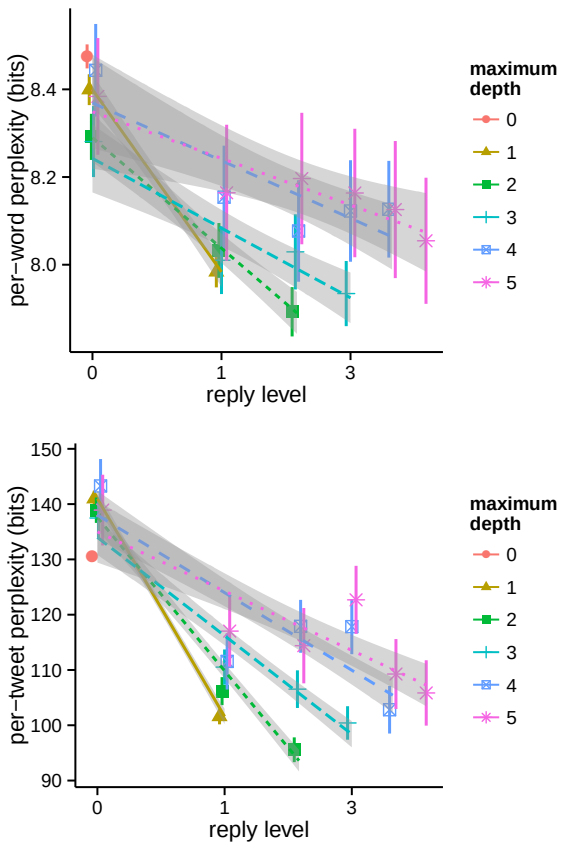

Figure 3: Per-word (top) and per-tweet (bottom) entropy decrease with reply level and increase with conversation length. Logarithmic fits with 95% confidence intervals; x-axis is log-scaled.

with number of mentions, even as visible tweet entropy decreases, suggests that audience size is having an effect. However, this effect is causing entropy to increase as common ground should be decreasing due to the larger number of conversation participants. Furthermore, this effect is not driven by reply level (Sect 3.3); there is a significant increase ($p < .001$) in explanatory power from adding number of mentions to a mixed-effects model with fixed-effects of reply level and a by-user random intercept.

3.3 Reply level and conversation length

We next turn to our second UID prediction: that information content should increase as common ground increases. As common ground is assumed to increase in dialogues (Clark, 1996), we thus predict that Twitter conversations should show increases in

information content that scale with reply level, the number of replies between the current tweet and the conversation root. Such a result would constitute a replication of Genzel and Charniak (2002) in the discourse context, and would confirm preliminary results on the Switchboard corpus by Vega and Ward (2009). As is clear from our analysis below, that is not what we found.

Figure 3 plots mean perplexities for different reply levels and conversation lengths, with confidence intervals based on by-user means. Increasing the reply level decreases the information content of the tweet, while increasing the conversation length increases the information content.

We fit a linear mixed-effects regression model to per-word and per-tweet perplexity. Control factors were the logarithm of the tweet reply level and the logarithm of the conversation length, along with a separate binary variable for whether the tweet was part of a conversation at all, and random by-user intercepts. Both log reply level and log conversation length had significant effects by likelihood-ratio tests.

Log reply level had negative effects on per-word and per-tweet perplexity (per-word: $-.341 \pm .009$; per-tweet: $-39.6 \pm .3$; both $p < .001$). Log conversation length had positive effects on per-word and per-tweet perplexity (per-word: $.285 \pm .010$; per-tweet: $35.1 \pm .3$; both $p < .001$).

To summarize these effects, all conversations lose entropy as they go along, and tweets that start longer conversations tend to have higher entropy to start. Whereas previous work has suggested that messages become more unpredictable as context builds up, Twitter conversations appear to shift to more predictable messages as context builds up, and seem to go until messages get sufficiently predictable. We discuss these results below.

4 General Discussion

Previous work supports the UID hypothesis that rational communicators adjust their messages so as to spread information as uniformly as possible in response to local context (Aylett and Turk, 2004; Levy and Jaeger, 2007), as well as to discourse-level context in monologic writing (Genzel and Charniak, 2002; Qian and Jaeger, 2012).

Our current work synthesizes these two bodies of work by looking for evidence of discourse-level UID effects in a large corpus of Twitter conversations, including dialogues and many-party conversations. Contrary to expectations, we failed to find UID effects; in fact, we often found information rate *increasing* when context changes would have predicted decreasing information rates. Specifically, we found that messages to smaller audiences, which should have lower noise and greater context and hence higher information density, actually have *lower* information density than messages to larger, noisier, and less context-sharing audiences. Furthermore, we found that later messages within a reply chain, which should have greater context, have less information. This last result is especially surprising because UID context effects have been repeatedly found in less conversational texts.

So should we give up on UID? While our results were unexpected, as we discuss below, we believe that they instead encourage more reflection on how speakers conceptualize information for conversational UID. We consider two aspects of these conversations: first, that the collaborative nature of conversation introduces rational uses for messages with low (lexicosyntactic) information density; second, that rational behaviors resulting from the nature of noise in social media communication complicates our evaluations of UID.

4.1 Information in terms of contributions

Why do Twitter conversations look different from the increasingly-informative texts studied in previous work? For one, our dataset contains true conversations, whereas almost all of the sentence-level context informativity results were based on single-author texts.

In monologues, there is neither an ability nor a need to check that common ground has been established. Participants in a conversation, though, must both produce content and establish grounding (Clark and Schaefer, 1987). Participants employ many methods to establish grounding, including backchannels (Yngve, 1970; Schegloff, 1982; Iwasaki, 1997), which have little lexicosyntactic information but provide crucial turn-taking and grounding cues.

Dialogues are often reactive; for instance, a re-

ply may be a clarification question, such as this reply in our dataset: *[MENTION] What do you mean you saw the pattern?* Such replies are typically shorter and more predictable than the original statement, and are often part of adjacency or coordinate pairs (Schegloff and Sacks, 1973; Clark and French, 1981), where one participant's utterance massively constrains the other's next utterance. Such pairs cover a wide range of low-entropy messages that are likely to appear in multi-party conversation but not in monologic text, including question-and-answers, offer-and-acceptances, and goodbyes.

As a result, Clark and Schaefer (1987) argue for a collaborative view of conversation structure, with conversations best viewed not as a series of utterances but as a series of contributions—sets of utterances that, combined, both specify some new content and establish it as part of the common ground. UID as a rational behavior is based on the idea that a rational speaker seeks to maximize linguistic information transfer, which would seem to be the primary goal during the content-specification portion of a contribution. During the grounding portion of a contribution, though, the primary goal is likely to be to establish common ground as quickly as possible. This goal is potentially more complex, as it depends on a variety of factors including the quality of the content-specification portion. If the content specification was simple and clear, grounding can be acheived with low-entropy backchannels (*mm-hmm, right*, etc.); if it was complex or unclear, grounding will require more messages and greater message entropy.

Furthermore, conversations may contain exchanges that have little linguistic context but serve important social ends. Many response tweets in our dataset are single, common words (*haha, lol*) or emoji/emoticons. These provide important emotional information in a very low lexicosyntactic entropy package, much as a backchannel or metalinguistic cue (e.g., a smile) might in face-to-face conversation. This suggests that the information measure within UID may not be strictly based on literal lexicosyntactic information but rather a combination of linguistic and metalinguistic information.

In sum, this conception suggests that in discourse, UID may operate as usual for parts of a contribution, but not necessarily throughout it. Rational conversational behavior may resemble an error-checking system in which UID may be observed at a contribution-by-contribution level.

4.2 Multiple types of noise

In most of the previously-studied genres, the authors of the texts could reasonably expect their readers to be both focused and unlikely to stop while reading. Tweets, however, are often read and responded to while doing other tasks, reducing focus and increasing disengagement rates. Interestingly, the one genre where Genzel and Charniak (2003) found a negative effect of sentence number on informativity was tabloid newspapers, where readers are likely to be distractable and disengaged.

It may be that Twitter requires an idiosyncratic adjustment to the noisy-channel model: perhaps the locus of the noise in tweets should not be in comprehension of the tweet per se (or at least not exclusively on comprehension). Instead, the main source of noise for Twitter users may be whether a reader engages with the tweet at all. Many Twitter users follow an enormous number of users, so outside of directed mentions and replies, there is a substantial chance that any given tweet will go unread by the larger part of its intended audience.[6]

The decreases we observed may have to do with users optimizing the amount of information content relative to the likelihood of an audience-member seeing more than one message. For tweets that go the largest audience, it is unlikely that multiple tweets would all be seen; thus it makes more sense to send information-rich tweets that can stand alone. In contrast, for replies, the intended audience should notice each sent tweet.

Evidence in favor of the conversation- or noise-based explanation could be obtained by comparing the Twitter reply chain effects against a corpus of conversations in which message reception is essentially certain, as in person-to-person chat logs (e.g., Potts 2012). If noise at the message level accounts for the anomalous Twitter behavior, then

[6] As a result, tweeters often create tweets that include their own context; for instance, a reply may quote part of its preceding tweet, or a user may talk about a recent event and include an explanatory link. This example from the corpus does both: *Pls help if you can! RT [MENTION]: Henry broke his foot [URL] Please donate: [URL].*

chat logs should show the UID effect of increasing entropy through the conversation. If turn-taking or meta-linguistic discourse functions drive it, chat logs would show decreasing entropy, as in our data.

4.3 Conclusions

We tested the Uniform Information Density hypothesis, which has been robustly demonstrated in monologue-like settings, on dialogues in Twitter. Surprisingly, we failed to find the predicted effects of context within these dialogues, and at times found evidence for effects going in the opposite direction. We proposed that this behavior may indicate a crucial difference in how information flow is structured between monologues and conversations, as well as how rational adaptation to noise manifests in different conversational settings.

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.

Herbert H. Clark and J. Wade French. 1981. Telephone goodbyes. *Language in Society*, 10:1–19.

Herbert H. Clark and Edward F. Schaefer. 1987. Collaborating on contributions to conversations. *Language and Cognitive Processes*, 2:19–41.

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

Vera Demberg, Asan Sayeed, Phillip Gorinski, and Nikolaos Engonopoulos. 2012. Syntactic surprisal affects spoken word duration in conversational contexts. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 356–367.

Gabriel Doyle and Michael C. Frank. 2015. Shared common ground influences information density in microblog texts. In *Proceedings of NAACL-HLT*.

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

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.

Shoichi Iwasaki. 1997. The northridge earthquake conversations: The floor structure and the 'loop' sequence in japanese conversation. *Jouranl of Pragmatics*, 28:661–693.

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

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.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL-HLT*.

Christopher Potts. 2012. Goal-driven answers in the Cards dialogue corpus. In Nathan Arnett and Ryan Bennett, editors, *Proceedings of the 30th West Coast Conference on Formal Linguistics*, Somerville, MA. Cascadilla Press.

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

Emanuel Schegloff and Harvey Sacks. 1973. Opening up closings. *Semiotica*, 8:289–327.

Emanuel Schegloff. 1982. Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. In Deborah Tannen, editor, *Analyzing Discourse: Text and Talk*, pages 71–93. Georgetown Univ. Press.

Alejandro Vega and Nigel Ward. 2009. Looking for entropy rate constancy in spoken dialog. Technical Report UTEP-CS-09-19, UTEP.

Victor Yngve. 1970. On getting a word in edgewise. In *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, pages 567–578. Univ. of Chicago Dept. of Linguistics.