Efficient Communication in Internet Slang

Emma Rafkin
Georgetown University
epr41@georgetown.edu

Abstract

In 1935, Zipf hypothesized that language becomes abbreviated over time due to the competing pressures of accuracy and efficiency, demonstrating that word frequency and length are inversely correlated. In later years, others hypothesized that word length is better explained by surprisal, or negative log probability given context. To test surprisal theory in a controlled setting, I looked at words that have the same meaning but different lengths: abbreviations and their full length counterparts. If language is in fact optimized towards efficient communication, then the amount of compression that a word experiences as it becomes abbreviated should be directly correlated to the amount of information lost between the long and short forms. I measured the frequency and surprisal of short and long forms of common abbreviations on a corpus of 3M Tweets. This experiment confirmed that word length is correlated to surprisal, shortened forms often occur in more predictable environments than their long form counterparts, and a meaning's compression rate is correlated to the difference in surprisal between its word forms. These results further cement idea that the amount of information in a single word is directly encoded in the amount of effort it takes to produce: its word length[1].

1 Introduction

Zipf's principle of least effort posits that language is optimized towards efficient communication. He demonstrated that more frequently used words tend to be shorter (Zipf, 1935, 1949). Zipf's theory was a predecessor of an information theoretic point of view: shorter words carry less information. Information content can be estimated probabilistically using surprisal. Surprisal measures the predictability of a word $\mathcal{W}$ given its context $\mathcal{C}$: $-log(P(\mathcal{W} = w|\mathcal{C} = c))$ (Hale, 2001). According to information theory, word lengths should be optimized based on their predictability in context (Gibson et al., 2019). Piantadosi et al. 2011 formally explored this, showing that a word's surprisal was actually a better correlate to word length than its frequency. While Piantadosi et al. 2011's findings have been supported by some (Mahowald et al., 2013, 2018), they have also been disputed by studies with more robust corpora or updated language models which found that frequency is a better correlate to word length than surprisal (Meylan and Griffiths, 2021; Levshina, 2022; Pimentel et al., 2023). Despite this disagreement, all of these experiments support Zipf's theory that language is optimized towards speaker efficiency as measured by word length (Mahowald et al., 2018).

Uniform Information Density (UID) predicts that in order to efficiently communicate in noisy or capacity-limited channels, "speakers will make strategic use of the flexibility allowed by their languages" to maintain "a uniform rate of information transmission close to the channel capacity" (Frank and Jaeger, 2008). Therefore, if there are word forms that convey the same meaning but differ in length, this should allow the speaker to strategically use the word form with more or less information when communicating in order to achieve UID. Mahowald et al. 2013 tested this with both a corpus and behavioral study on a set of known abbreviated/long form pairs. For meanings like "hippo" vs "hippopotamus" or "frat" vs "fraternity", they found that the short forms were generally used in more predictive environments. This finding supports UID as it means that the context already contained some of the information carried by the long form, therefore the short form was used to maintain a uniform information communication rate (Mahowald et al., 2013). This experiment demonstrated a consistent difference between the long and short forms of words with the same meaning. However, it did

[1] https://github.com/erafkin/twitter_abbreviation_word_length

not explore consistencies in the rate of information loss in process of abbreviation. If a word's length is correlated to information content, then another prediction that surprisal theory makes is that more information loss in abbreviated forms should result in more word length compression. Therefore, the difference in surprisal between the long and short forms should be correlated to the compression rate (as measured by $1 - \frac{length(shortform)}{length(longform}$).

I hypothesize that abbreviations are more common in a channel with a lower information capacity. When there is a limit to the amount of information that can transmitted over a channel, it would make sense that all redundant long forms should be abbreviated especially if a common short form exists and is predictable in context. Due to its character limits (X) and its strong foundation in "internet speak", I posit that the Twitter is one such channel with a lower information capacity, especially in comparison that of books (as were studied in the previously mentioned experiments). Therefore I expect that Twitter will have more abbreviations than a more "standard" corpus. This paper expands on Mahowald et al. 2013's work, taking into account the criticism on Piantadosi et al. 2011's findings, exploring the correlation between word length, frequency, and surprisal in Tweets. Furthermore, this paper explores the correlation of frequency, surprisal, and long form length with the rate of compression that a meaning can experience. A correlation between the difference in surprisal between the short and long forms and compression rates would show that shortening the word length of a meaning is in fact reducing the amount of information content contained a word form's representation.

2 Methods

This paper combines all of the configurations and splits of the Cardiff TweetEval datasets, totaling in over 3M Tweets (Barbieri et al., 2020, 2018a,b; Van Hee et al., 2018; Basile et al., 2019; Zampieri et al., 2019; Rosenthal et al., 2017; Mohammad et al., 2016). In order to compare the Twitter corpus with a more traditional corpus (with a higher information capacity), I also look at average word lengths in the English Wikipedia dump (Foundation). I randomly sampled documents to match the total Twitter corpus word length. I used BERTweet (Nguyen et al., 2020), a open-source BERT-based model trained on 850M Tweets[2] to estimate surprisal. Following Levshina 2022, I chose a model that has access to future context. This choice was made because the act of compressing words to meet a character limit intuitively should require a knowledge of the Tweet as a whole. This intuition is supported by findings that future context significantly correlated to word length as well as past context (Jurafsky et al., 2001). A list of 55 internet abbreviations (and their full length forms) were compiled using an online blog as a source and inspiration (Beal). 15 of the short forms of the meanings were unknown to the BERTweet tokenizer (e.g. "icymi" ("in case you missed it") and "fomo" ("fear of missing out")). This means the short form surprisal would be calculated using the <unk> token and would not be specific to the relevant meaning, therefore they were dropped from analysis. This indicated that these remaining short forms in the abbreviation list were present enough in the training data for BERTweet to develop specialized tokens for them. Their relative frequency in the training corpus in comparison to their real-life usage in Tweets directly affects how accurate BERTweet is at estimating their information content.

The question of how to calculate surprisal for multi-token words and phrases using a language model is open ended. A popular method for this calculation is the *Pseudo Log Likelihood (PLL)* described in Salazar et al. 2020, which involves individually masking each token in the phrase and taking the sum their surprisals. It has been shown that using PLL to measure the surprisal of a multi-token OOV word[3] can artificially lower the result as the bidirectional nature of the model might provide more information to the model than is appropriate for this calculation (Kauf and Ivanova, 2023). Therefore, Kauf and Ivanova, 2023 suggest following the pattern that a generative model might use: initially masking all of the tokens in the phrase and then unmasking them once the prediction as happened. This PLL calculation, dubbed *PLL-word-L2R*, was initially developed for calculating surprisal for a single OOV word, but in this paper I will call it *PLL-L2R* as it will be used to calculate surprisal for a phrase. Table 1 walks through an example of the different masking procedures used

[2]BERTweet can be accessed at `https://huggingface.co/docs/transformers/en/model_doc/bertweet`

[3]In this case I am treating a phrase (e.g. "i do n't know" or "oh my god") as a multi-token OOV word

Index	PLL				PLL-L2R			
0	<mask>	do	n't	know	<mask>	<mask>	<mask>	<mask>
1	I	<mask>	n't	know	I	<mask>	<mask>	<mask>
2	I	do	<mask>	know	I	do	<mask>	<mask>
3	I	do	n't	<mask>	I	do	n't	<mask>

Table 1: Different masking patterns for calculating the surprisal of the phrase “i don't know"

in this paper to calculate the surprisal for phrases.

Following Piantadosi et al. 2011; Mahowald et al. 2013; Pimentel et al. 2023; Meylan and Griffiths 2021, I calculated Spearman correlation between variables in question.

3 Results

The Wikidump corpus had an average word length of 5.26 characters while the Twitter data had an average word length of 4.88 characters (medians 5 and 4 respectively). This small test confirms the intuition that Twitter is a platform that encourages shorter words than a more “standard" written channel. See Appendix A for more analysis on the cause of language compression on Twitter specifically.

Table 2 displays the mean surprisal (PLL and PLL-L2R) and frequency for short and long forms for the corpus as a whole as well as split into single- and multi-token full length meanings (e.g. “oh my god/omg" vs “brother/bro"). For the rest of the paper, I refer to multi-token full length meanings as “phrases" and single-token full-length meanings as “words". This table shows that for the corpus as a whole, the long forms had a higher surprisal than the short forms. Additionally, the results of running correlation tests match Piantadosi et al. 2011's results: surprisal (PLL and PLL-L2R) was more correlated with word length than frequency for this corpus with a Spearman correlation of 0.63 for both surprisal calculations and -0.38 for the frequency calculation ($p < 0.01$). The rest of the results from the table do not necessarily match expectations from earlier experiments. The two calculations for surprisal, PLL and PLL-L2R did report different results, although they demonstrate the opposite of Kauf and Ivanova 2023's intuition. The PLL-L2R reported higher surprisal rates for the phrases than PLL. Additionally, there are many more corpus-wide instances of the long forms, with a significant amount coming from the word long forms. This is likely due to the fact that the word long form list included extremely common words such as “you", “very", “with", and “are" which significantly raised the average frequency counts for the word long forms. Despite this, it was still true that the phrase long forms were more common on average than their abbreviations, which was not the case in Mahowald et al. 2013's experiment.

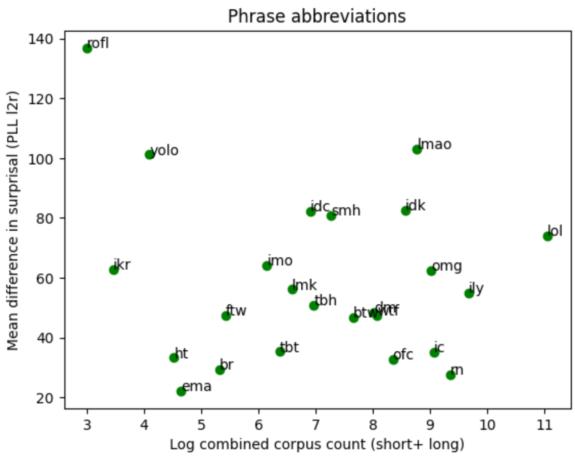

Figure 1: Phrase abbreviation surprisal compared to frequency counts. This shows that the short form phrases occur in *more* predictive environments than their long form counterparts. Meanings are labeled with their short forms to save space.

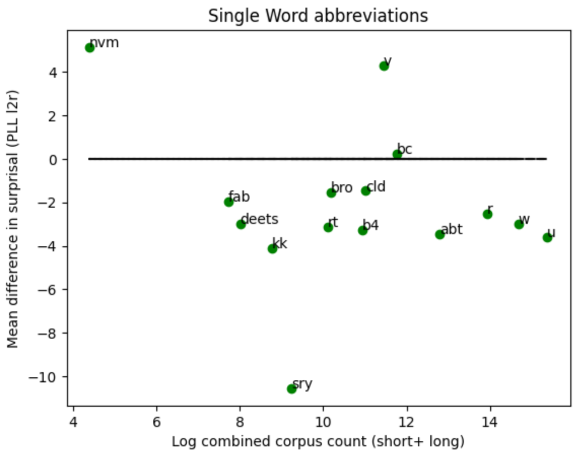

Figure 2: Word abbreviation surprisal compared to frequency counts. This shows that the short form words occur in *less* predictive environments than their long form counterparts.

Table 2 also shows that short forms as a whole occurred in more predictable environments. In Ma-

	All Short	All Long	Phrase Short	Phrase Long	Word Short	Word Long
PLL	34.88	72.63	36.04	91.22	35.06	32.90
PLL-L2R	34.88	70.52	36.04	88.32	35.06	32.90
Counts	397.25	2425.05	101.39	139.83	781.93	6599.5

Table 2: Mean PLL, PLL-L2R, and frequency counts across the corpus and split into phrase and single-word ("oh my god" vs "brother") forms.

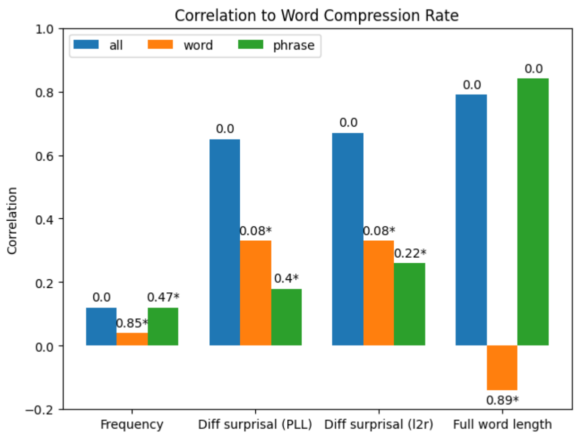

Figure 3: Corpus, phrase, and single-word Spearman correlations to compression rate. p values are located on top of the bars. * next to the values where $p > 0.05$ to emphasize the unreliability of those results.

howald et al. 2013's results, when comparing the log of the sum of the frequency of the short and long forms to the difference in surprisal, the majority of the meanings fell above the x-axis. The Twitter results, however, are not as clear. When split into word abbreviations and phrase abbreviations, all of the phrases fall above the x-axis (see Figure 1) while the words fall mostly fall below the x-axis (see Figure 2). Therefore, I find that Mahowald et al. 2013's results hold only for phrases. This makes sense due to the manner by which PLL is calculated: as a sum. The fact that this was not true for the single word abbreviations is interesting counter-evidence to Mahowald et al. 2013's findings. That being said, the scales of Figures 1 and 2 differ significantly, with the word calculations being much closer to 0. The reason for this might be due to BERTweet's inability to model the short forms with fewer examples of them in training, as discussed further in Section 4.

Figure 3 displays the correlation between the compression rate and the frequency, difference in surprisal (PLL and PLL-L2R), and length of the long form[4]. These correlations are also broken down for the corpus as a whole, the phrases, and the single words. Across all three analyses, difference in surprisal is a consistent correlate to compression rate. The length of the long word/phrase was also a strong correlate for the whole corpus and the phrases. This is intuitive as the longer the phrase, the more likely it is to be drastically abbreviated if possible ("rolling on the floor laughing" (20) becomes "rofl" (4) whereas "oh my god" (9) becomes "omg" (3)). All of the results for the full corpus had $p < 0.01$. It should be noted that once the corpus was split into word v.s. phrase abbreviations, the p values significantly increased. Despite this inconsistency in some of the correlation calculations, these results as a whole support an information theoretic perspective in that the amount of compression that a word experiences is correlated to its information "loss" as measured by a difference in surprisal.

4 Discussion

I have shown that for this corpus, word length is correlated is to surprisal and that abbreviated phrases occur in more predictable environments than their full length counterparts. Additionally, this experiment shows that compression rates are correlated to the difference in surprisal between the short and long forms. This final discovery supports the intuition that the amount of information lost in an abbreviated form is directly encoded by the amount of efficiency the speaker gains in using the shortened form. All of these results support the the information theoretic claim that language is optimized towards speaker (or in this case writer) efficiency as measured by surprisal.

This experiment attempts to qualify the use of a Twitter corpus due to the assumption that there are more abbreviations in the corpus because of the character limits that apply to the majority of users[5]. While the validity of this assumption is explored more thoroughly in Appendix A, Twitter is

[4]The long form length was included in this experiment due to the intuition that it might just be the longest words that frequently compressed.

[5]More recently X (the new name for Twitter after a change

nevertheless a complicated channel capacity limiter because the information contained in a single Tweet is not necessarily constrained to the Tweet itself. Information can come through an attachment or a previous Tweet (if the Tweet at hand is a “Retweet" or a part of a “Thread"). All of these contextual information channels could mean that Twitter is not a particularly limited information channel. Even if the assumption that Twitter as a platform uniquely encourages abbreviation more than other online platforms is false, this Twitter corpus did contain shorter words on average than the Wikipedia corpus. This Twitter corpus can then be seen as merely an “internet speak" corpus, where it is assumed that abbreviation on the internet is encouraged simply to minimize keystrokes and because it is the norm (Barseghyan, 2013).

Another conceptual assumption that was made in this experiment is that the long and short versions of the word meanings have the same semantic value. While of course a change in the amount of information that a word form holds is a change in meaning, it is assumed that the versions refer to the same entity or have the same grammatical function. This assumption is a simplification of the process of language change. The phrase “direct message" is generally a compound noun while “dm" can be a noun or a verb. For example,“send me a dm" and “dm me". In contrast, it would be unlikely for someone to produce “direct message me" over the simpler “message me" which already implies that the action will be direct. Therefore the long form does not act as a verb in the same way that the short form can. Some of the phrases have a much larger shift, take “bro" as an example. “Brother" generally means a male sibling or a close friend. “Bro" has come to mean both “brother" but is also often used as a discourse marker (e.g. `` ``bro im fuming right now" ``). While many of the short forms do maintain the meanings of their long forms, it cannot be dismissed that many short forms can also carry meaning that deviates from the original form. A future experiment might want to define a threshold for semantic similarity between the short and long forms before including them in the analysis.

An interesting result from Table 2 shows that the PLL-L2R scores for the phrases were lower than the PLL scores. This result goes against the intuition of Kauf and Ivanova 2023 that PLL unfairly exposes the model to too much context leading to artificially low surprisals. A reason for this could be that these multi-word phrases often are slightly idiomatic and taken literally do not contain predictable words when masked one at a time. This result in combination with the overall low mean counts for the long form phrases could mean that they were not prevalent enough in the training data for the model to have learned their idiomatic meaning or the fact that they often are produced as a group. If this is the case, then the PLL calculation scheme might lead to higher surprisal counts because of unpredictable nature of the content words of the phrase. Without seeing many examples of the phrase “I know right", “I `<mask>` right" might have an extremely high surprisal for the word “know" in comparison to other verbs such as “was", or “went" which might normally accompany “right". Sequentially unmasking the tokens might lead to a more predictable (i.e. lower surprisal) sequence of words because the “right" would be masked and wouldn’t affect the predictions for “know". This analysis leads to two conclusions for future work:

1. The abbreviation curation process should be more robust. Ideally, abbreviations should be detected dynamically. As this is quite a difficult and unsolved NLP task (Veyseh et al., 2020), there should at least be some count threshold on the abbreviations being studied. This would ensure that they were seen frequently enough in the training set so model is aware of the word groupings in a phrase. This count threshold should also apply to the short forms so that the model has more examples to learn where they might be used. This might help account for the inconsistencies displayed in Figure 2 where single word abbreviations were generally found in less predictable environments than their full-length counterparts. Due to the sheer number of instances of the full-length version of words such as “you", or “with", it is unlikely that a model will ever predict the abbreviated version over the full length version. It is unclear whether this is a reflection on the general use of the abbreviated forms, the method of surprisal calculation, or simply lack of representation in the training data. A model trained on a more representative set of examples on both sides of the comparison might lead to a more reliable understanding of how the abbreviated forms are

in ownership) has allowed paid users up to 4,000 characters (X).

used.

2. Although the use of BERT for the next token context was motivated by past research, the differences in the surprisal calculations in Table 2 show that maybe causal models that only see past context would be more suited for this task. Especially as data scarcity seems to be impacting the results of this experiment, it would be beneficial to look at causal language models as those tend to be the models that are trained on the most data. This might solve both problems in one-fell-swoop.

Running the same experiment a larger, generative model using a balanced corpus and more robust list of abbreviations might clarify some of the inconsistencies in these findings. Despite the limitations of this experiment, the results overwhelmingly support surprisal theory and demonstrate an optimization towards efficient communication in internet slang.

References

About different types of posts.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, page 1644–1650, Online. Association for Computational Linguistics.

Francesco Barbieri, Jose Camacho-Collados, Francesco Ronzano, Luis Espinosa-Anke, Miguel Ballesteros, Valerio Basile, Viviana Patti, and Horacio Saggion. 2018a. Semeval 2018 task 2: Multilingual emoji prediction. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 24–33.

Francesco Barbieri, Jose Camacho-Collados, Francesco Ronzano, Luis Espinosa-Anke, Miguel Ballesteros, Valerio Basile, Viviana Patti, and Horacio Saggion. 2018b. Semeval 2018 task 2: Multilingual emoji prediction. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, page 24–33, New Orleans, Louisiana. Association for Computational Linguistics.

Liana Barseghyan. 2013. On some aspects of internet slang. *Graduate School of Foreign Languages N*, 14:19–31.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, page 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Vangie Beal. Twitter dictionary: A guide to understanding twitter lingo.

Wikimedia Foundation. Wikimedia downloads.

A. Frank and T. Florian Jaeger. 2008. Speaking rationally: Uniform information density as an optimal strategy for language production. *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, pages 939–944.

Edward Gibson, Richard Futrell, Steven P. Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. 2019. How efficiency shapes human language. *Trends in Cognitive Sciences*, 23(5):389–407.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001 - NAACL '01*, page 1–8, Pittsburgh, Pennsylvania. Association for Computational Linguistics.

Daniel Jurafsky, Alan Bell, Michelle Gregory, and William D. Raymond. 2001. *Probabilistic relations between words: Evidence from reduction in lexical production*, volume 45, page 229. John Benjamins Publishing Company, Amsterdam.

Carina Kauf and Anna Ivanova. 2023. A better way to do masked language model scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, page 925–935, Toronto, Canada. Association for Computational Linguistics.

Natalia Levshina. 2022. Frequency, informativity and word length: Insights from typologically diverse corpora. *Entropy*, 24(2):280.

Kyle Mahowald, Isabelle Dautriche, Edward Gibson, and Steven T. Piantadosi. 2018. Word forms are structured for efficient use. *Cognitive Science*, 42(8):3116–3134.

Kyle Mahowald, Evelina Fedorenko, Steven T. Piantadosi, and Edward Gibson. 2013. Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, 126(2):313–318.

Stephan C. Meylan and Thomas L. Griffiths. 2021. The challenges of large-scale, web-based language datasets: Word length and predictability revisited. *Cognitive Science*, 45(6):e12983.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, page 31–41, San Diego, California. Association for Computational Linguistics.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, page 9–14, Online. Association for Computational Linguistics.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Tiago Pimentel, Clara Meister, Ethan Wilcox, Kyle Mahowald, and Ryan Cotterell. 2023. Revisiting the optimality of word lengths. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, page 2240–2255, Singapore. Association for Computational Linguistics.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, page 502–518, Vancouver, Canada. Association for Computational Linguistics.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 2699–2712, Online. Association for Computational Linguistics.

Cynthia Van Hee, Els Lefever, and Veronique Hoste. 2018. Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, page 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

Amir Pouran Ben Veyseh, Franck Dernoncourt, Quan Hung Tran, and Thien Huu Nguyen. 2020. What does this acronym mean? introducing a new dataset for acronym identification and disambiguation. In *International Conference on Computational Linguistics*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, page 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

George K Zipf. 1935. *The Psycho-Biology of Language*. Routledge, London.

George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley Press, Cambridge, Mass.

A Tweet Length Distributions

One hypothesis in this paper is that the character limits of Twitter create a low capacity channel that encourages abbreviations (referred to as compression throughout this paper). If this were the sole reason why there was more abbreviations in Tweets, then there should be a visible spike in the distribution of Tweets near the character limit containing the short forms of words that does not exist in the distribution of Tweets containing the long forms of words. This is because a user would hit the character limit and then need to go back through their tweets and abbreviate whatever they could without losing the original meaning. Figure 4 displays the distribution over Tweet length for Tweets containing the short and long forms of the words of interest across the corpus. Figure 5 displays the distribution over Tweet length for Tweets near the character limit[6] (Tweet length > 270). The shapes of these distributions match quite closely. Additionally, while there is an increase in the short forms near the character limit, that is matched by the long forms and the spike in the long forms is more dramatic, as is clearly displayed in Figure 5. Based on these distributions, it does not appear that the character limits were a catalyst for compression on this dataset. Therefore, there must be other reasons to explain the fact that word lengths in Twitter are shorter than other mediums. This could due to the effort that it takes to communicate via typing, especially on a phone rather than a computer (which is how Tweets are often crafted), as internet slang often originates to save on keystrokes. It could also just be that abbreviations are part of the Twitter/internet lingo and a signal of being in the internet-savvy in-group (Barseghyan, 2013). Additionally, it is always possible that these distributions would change and display more of the character-limit-driven narrative given a more robust set of abbreviations for analysis. If it is the case that character limits do not uniquely drive the need for abbreviations, then this is further evidence that a causal language model could be preferable for this measuring surprisal.

[6]Tweets with more than 280 characters were dropped from analysis. There were 756 Tweets containing the relevant long and short forms that were longer than 280 characters, leaving 76,551 unique Tweets for analysis

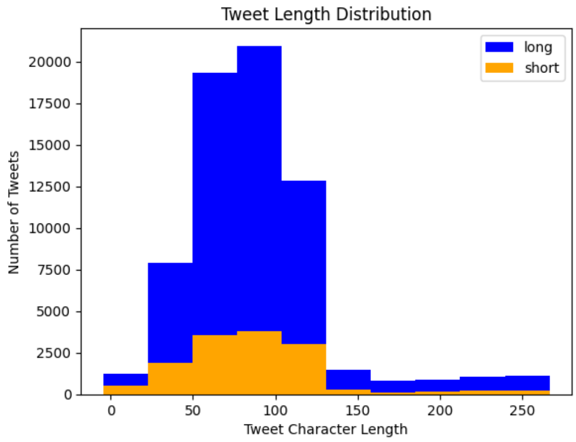

Figure 4: Distributions of Tweets containing the long and short form of the words

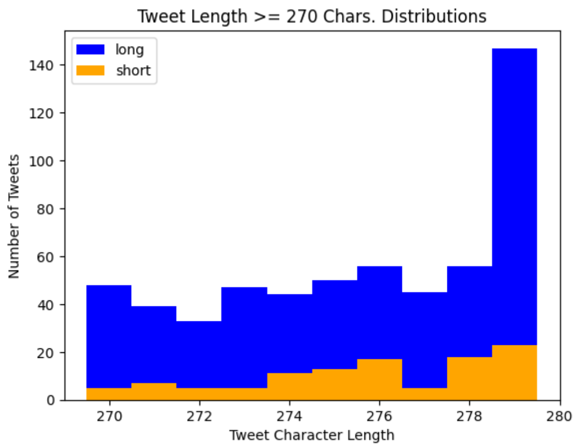

Figure 5: Distributions of Tweets >= 270 characters containing the long and short form of the words.