Punchlines and Patterns: A Quantitative Analysis of Language Use in Stand-Up Comedy

Michael Robinette[*] and Danny Satterthwaite[†]

(Dated: December 8, 2024)

This paper explores the linguistic and statistical properties within stand-up comedy, focusing on the application of Zipf's Law and Heaps' Law. By analyzing each set's innovation rate, which quantifies the proportion of unique words relative to the total word count, we examine the relationship between innovation rate, transcript length, and performance duration. Through the lens of Zipf's and Heaps' Laws, we investigate vocabulary distribution and growth patterns in the unique context of stand-up comedy, a medium that combines crafted material with spontaneous improvisation. Our analysis reveals that the scaling of language in stand-up comedy mirrors patterns seen in other forms of narrative, such as those found in short stories by Edgar Allan Poe. This paper contributes a novel perspective to the study of humor and linguistics, highlighting the dynamics of language use in comedic performances and its connection to vocabulary diversity and frequency distributions.

I. Introduction

In 2002, to mark the 100th anniversary of George Kingsley Zipf's birth, a paper titled Zipf's Law Everywhere [1] was published, highlighting the wide-ranging applications of Zipf's Law across disciplines from human language to natural and physical phenomena. Although Zipf's Law has been shown to emerge in various fields, one area that has remained largely unexplored is the stand-up comedy stage. Like other art forms, stand-up comedy explores the depths of human experience, provoking laughter, tears, and even discomfort. However, unlike more traditional forms of art, stand-up comedy specials often blend highly crafted material with spontaneous improvisation, making it a particularly interesting medium through which to examine Zipf's law. This paper will examine the language of comedy through the lens of not only Zipf's Law but also Heaps' Law. It will argue that, much like the vocabulary growth observed in short stories by Edgar Allan Poe, the language used in stand-up comedy scales in a comparable manner, reflecting patterns found in well-known literary works.

In the study of complex systems, scaling is a core analytical lens used to examine these systems. While scaling itself may seem relatively simple, the existence of scaling within complex systems—which often involve many interconnected and interrelated parts—can signal underlying structures and mechanisms. It is not merely the observation that scaling exists that makes a system complex; rather, it is through exploring how and why scaling arises that we learn the true complexity of the system. Language is a complex system—letters, words, and tokens all coming together to have meaning and convey feelings and emotions. Comedy, as a form of language, provides an engaging context to study this complexity. Before unpacking the mechanisms that lead to the inherently complex nature of language and comedy, we explore the various scaling behaviors that exist within our system of comedy specials.

Arguably the most important type of scaling is power law scaling. In the context of literature and language, Zipf's law and Heaps' Law are the most widely researched and explored types of power laws. Both Zipf's and Heaps' laws are core to our analysis, making an understanding of them essential.

In the context of language [2] and standup comedy as we will explore, Zipf's law states that $S_r \propto r^{-\alpha}$, where r is the rank of the word by the number of times it appears, S_r is the frequency of words with a given rank r, and α is a scaling exponent. In other words, the frequency of words that appear a given number of times is inversely proportional to their rank. An often-discussed mechanism for Zipfian distributions is the rich-get-richer mechanism, where words that are used more frequently in the present are more likely to be used in the future. In the context of comedy, this is particularly relevant as comedians often use language that resonates with audiences, leading to repeated usage. However, a crucial aspect of comedy is the introduction of new words and phrases. This rate of introducing new words is called the innovation rate, denoted ρ, and it connects the rich-get-richer mechanism to Zipf's law by $\alpha = 1 - \rho$. Thus, the larger the innovation rate, the larger the tail in the Zipf distribution.

Similarly, Heaps' Law states that the longer a piece of text is, the fraction of unique words used decreases. In other words, $N_t \propto t^{\gamma}$, where t is the length of the text (number of words) and N_t is the number of distinct words. There is an interesting parallel with the innovation rate, as Heaps' Law effectively describes a

[*] Michael.Robinette@uvm.edu
[†] dsattert@uvm.edu

Typeset by REVTEX

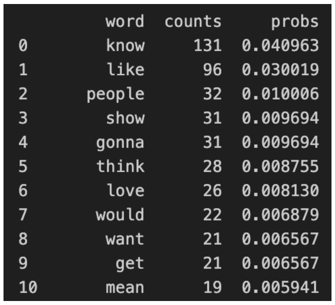
decaying innovation rate within a single work. This insight is particularly valuable when analyzing comedy specials, as comedians balance between familiar language that resonates with audiences and the introduction of novel language to keep their material fresh and engaging.

II. Description of data sets

The dataset consists of 50 comedy specials. In selecting our specials, we made an effort to be as inclusive and representative as possible. The data focus on three main features: title, length of time on stage, and transcript. In terms of corpus length, there is a range between 4,329 and 13,771 tokens. In this paper, we primarily discuss specials at the tail or midpoint of the data set. All data were obtained from Scraps From The Loft website.

III. Results

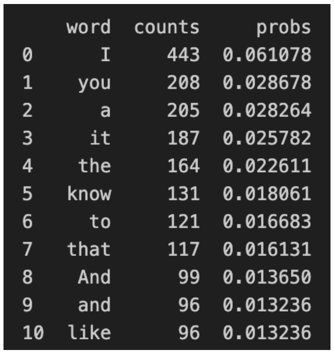
In an effort to abide by the principle of least effort, we set out to begin analyzing the data through the lens of Zipf's law. It should be noted that in tokenizing each corpus we chose to observe the data by removing stop words while the default was to keep them. To remove stop words, we used a Python library known as NLTK or natural language toolkit. This library removes by default 40 words that do not "enhance our understanding" of the corpus. Removing stop words yielded words that were more central to the ideas the comedian was trying to convey. See below.



```
      word  counts     probs
0     know     131  0.040963
1     like      96  0.030019
2   people      32  0.010006
3     show      31  0.009694
4    gonna      31  0.009694
5    think      28  0.008755
6     love      26  0.008130
7    would      22  0.006879
8     want      21  0.006567
9      get      21  0.006567
10    mean      19  0.005941
```

FIG. 1. Stop words excluded

After generating Zipf plots for all 50 comedians in our dataset, we sought to compare sets of various corpus lengths to well-known short stories of similar length. Below, we compare Sarah Silverman's Jesus Is Magic to Edgar Allan Poe's The Fall of the House of Usher—with stop words preserved. These selections represent the mid-range of our dataset in terms of corpus length. Similar results were observed across the lower and upper ranges of the dataset.



```
    word  counts     probs
0      I     443  0.061078
1    you     208  0.028678
2      a     205  0.028264
3     it     187  0.025782
4    the     164  0.022611
5   know     131  0.018061
6     to     121  0.016683
7   that     117  0.016131
8    And      99  0.013650
9    and      96  0.013236
10  like      96  0.013236
```

FIG. 2. Stop words included

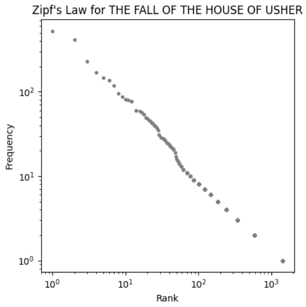
In the plots below, we observe the characteristic downward trend in word frequency as the rank of the word increases. Both plots exhibit approximate linear decay on a log-log scale, consistent with Zipf's law. This suggests that higher-ranked words occur less frequently than lower-ranked words. However, several key differences emerge upon closer examination.

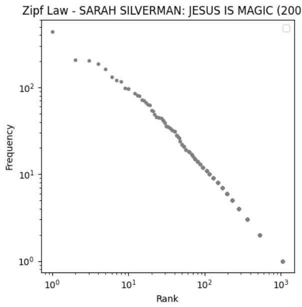
First, Jesus Is Magic shows a flatter slope for the highest-frequency words (low ranks), indicating greater repetition of common words. This aligns with expectations for spoken language, where conversational fillers and repeated phrases are more prevalent. In contrast, The Fall of the House of Usher adheres more closely to a steeper power-law decay in this range, reflecting the structured and deliberate word choice typical of literary prose.

Second, the tail behavior (low-frequency, high-rank words) differs between the two. Jesus Is Magic has a smoother, more compressed tail, which likely reflects the limited lexical variety and simpler vocabulary of a stand-up performance. Conversely, The Fall of the House of Usher exhibits a more jagged tail, consistent with a richer and more diverse vocabulary, as well as less uniform frequency among rare words.

Lastly, the overall slopes of the mid-rank sections suggest differences in vocabulary diversity. The middle ranks in Jesus Is Magic have a gentler slope, indicating a reliance on conversational language with a smaller range of distinct words. Meanwhile, The Fall of the House of Usher features a steeper slope in this range, demonstrating Poe's denser and more varied lexicon.

While both texts conform to the general trends predicted by Zipf's law, the differences in slope, tail behavior, and overall lexical diversity highlight the stylistic distinctions between spoken comedic performance and literary prose. These findings are representative of broader patterns observed across the dataset, reflecting the impact of genre and medium on word frequency distributions.





Next we wanted to examine innovation rate to get a sense of the richness of the vocabulary in the stand-up environment. To calculate innovation rate we capture the number of unique words across a corpus as well as the total number of words and take the ratio. Again we compare the same texts that represent our mid-range in terms of length - THE FALL OF THE HOUSE OF USHER with an estimated innovation rate of 0.2855 and Jesus is Magic with an estimated innovation rate of 0.2046. See empirical vs theoretical below.

TABLE I. THE FALL OF THE HOUSE OF USHER

	Empirical Estimate	Theoretical Estimate
n_1^g	0.639	0.583
n_2^g	0.162	0.172
n_3^g	0.071	0.078

TABLE II. Jesus is Magic

	Empirical Estimate	Theoretical Estimate
n_1^g	0.565	0.557
n_2^g	0.152	0.171
n_3^g	0.072	0.080

To further investigate, we examined how it differed with respect to the length of time on stage and actual corpus length and found that innovation rate seemed to peak and begin to diminish with an increase in time on stage and length of corpus. This trend is far more evident in the Innovation rate vs. corpus length plot.

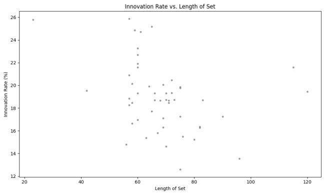


FIG. 3. Innovation rate vs. length of time on stage

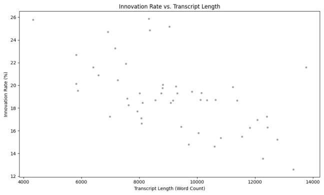
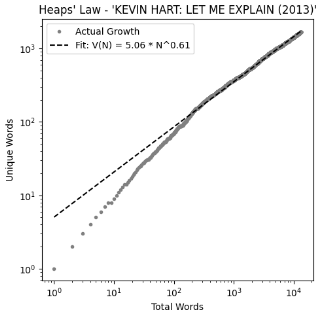


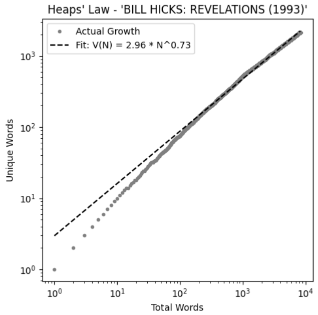
FIG. 4. Innovation rate vs. corpus length

Given the above findings, we chose to examine the comedy data by applying Heaps' law. From the graphs above it would seem that the more time a comedian spends on stage, the longer the corpus, resulting in a lower innovation rate. While this may be true, all comedians in our data set are introducing new vocabulary at a very fast rate where innovation ranges from $12.60 - 25.86$. Recall that $N_t \propto t^\gamma$ where t is the length of the text (number of words) and N_t is the number of distinct words. Below, we examine key comedians with the lowest and highest innovation rates, as well as exceptions to the pattern.

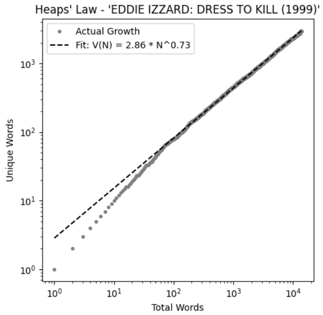
First, we will look at KEVIN HART: LET ME EXPLAIN (2013). His special had the lowest innovation rate at 12.60 with a corpus length of 13330. Despite this, his vocabulary growth follows a clear Heaps' pattern with a relatively low Heaps' exponent of 0.61. This suggests a steady but slower introduction of new terms compared to other comedians. The graph demonstrates that, while Hart's performance spans a long corpus, the rate of unique word introduction stabilizes earlier.



The highest innovation rate belonged to BILL HICKS: REVELATIONS (1993) at 25.86 with a corpus length of 8330 terms. Hicks' Heaps' exponent of 0.73 indicates a much faster vocabulary expansion relative to corpus size. This reflects his ability to introduce diverse vocabulary throughout his performance, as seen in the steeper slope of the Heaps' plot.

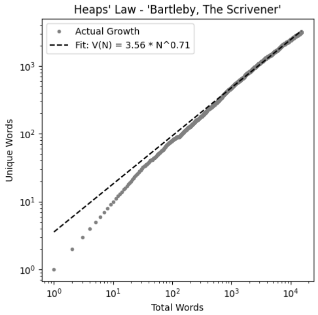


What is truly remarkable is the exception to the pattern of a diminishing innovation rate in relation to corpus length. Eddie Izzard in Dress to Kill introduced a total of 13771 words and had the longest corpus in our dataset. However, the innovation rate in this set was 21.57. Izzard also introduced new terms at the same rate as Bill Hicks.



All of this suggests that comedians introduce new vocabulary at a fast rate. This is likely due to the shorter length of the corpus. However, stand-up comedy differs from many other linguistic systems normally analyzed through the lens of Zipf and Heaps' law in that there is often steady topic transition as well as improvisation.

For comparison, we analyzed a classic literary work, Bartleby, The Scrivener. This text, with a corpus length of $15,202$ words, exhibited a Heaps' exponent of 0.71 The innovation rate, though comparable to Eddie Izzard's performance, aligns more closely with traditional literary expectations, where new vocabulary is introduced at a steady rate. This further highlights the unique dynamics of stand-up comedy as a linguistic system, characterized by rapid topic transitions and improvisation.



IV. Concluding remarks

V. Next steps:

- **Justify comedian selection**: We need a strong rationale for the comedians we chose to analyze, such as IMDB ratings, Netflix ratings, or other relevant metrics. While we spent significant time collecting data, we did not focus enough on justifying our selections. Given the limited availability of stand-up comedy data, a well-reasoned selection process would strengthen the study.

- **Clean and standardize the data**: Data cleaning and standardization are crucial. This was one of the reasons we focused on comedians at the extremes and middle of our dataset, where we could confidently analyze a smaller subset. Although collecting 50 specials was an accomplishment, it quickly became clear that this volume was too much for the scope of this project. Proper data preparation would also allow others to use the dataset in future research.

- **Use a more comparable medium for analysis**: The paper would benefit from comparing stand-up comedy to a medium more similar to it. We chose literature because we already had tools in place to analyze texts from Project Gutenberg, but alternatives like speeches or *The Moth Podcast* could provide a more relevant comparison. These formats have been subject to prior analysis and might yield insights into how stand-up comedy relates to other spoken-word media.

- **Clarify the focus**: The paper currently lacks a clear connection between Zipf's Law and Heaps' Law, requiring further work to integrate these concepts. Alternatively, we could structure the paper around innovation and Heaps' Law, providing a more focused narrative.

Acknowledgments

We are grateful for Peter Dodds and Ashley Fehr for letting us bounce ideas around.

[1] W. Li, Zipf's law everywhere., Glottometrics **5**, 14 (2002).

[2] L. Lü, Z.-K. Zhang, and T. Zhou, Deviation of zipf's and heaps' laws in human languages with limited dictionary sizes, Scientific reports **3**, 1082 (2013).