Team 2
Jack Hanke, Daniel Plotkin, Nicole Birova
1 May 2025

NLP Seminar Project Proposal

George Zipf, in his 1932 book *The Psychobiology of Language,* introduced the remarkable fact that the word frequencies in many human languages exhibit the same distribution. Nearly a century later, modern large language models (LLMs) generate text comparable to human communication. Unlike humans, LLMs digest text using a fixed vocabulary of word fragments, called *tokens*. These tokens are most commonly generated using the Byte Pair Encoding algorithm, which is an algorithm designed to avoid allocating a specific encoding to rare words. This motivates the question that is the title of our paper: *Are Modern Byte Pair Tokenizers Zipfian?*

Though Zipfian analyses have been conducted on many corpuses, none have yet been done at the token level. Our team was motivated by the novel question of whether or not LLMs learn from the same distribution shape that humans do.

To answer our question, we study two tokenizers, the RoBERTa tokenizer (via Hugging Face), and OpenAI's GPT-4 tokenizer (via tiktoken). We choose the RoBERTa tokenizer because the exact BPE training dataset is publicly known, which will allow for comparisons of Zipf-ness on seen and unseen corpuses. We also choose OpenAI's GPT-4 as a representative for cutting edge industry tokenizers, though the BPE training dataset is not public.

We will conduct our frequency analysis by tokenizing the first ~10M words of the BooksCorpus dataset (Zhu et al, 2015) for both tokenizers, as it is known to be used for RoBERTa's tokenizer (and likely for GPT-4's). For unseen-to-RoBERTA data, we also tokenize the first ~10M words of the cleaned C4 dataset hosted on Hugging Face.

We determine Zipf-ness using the Kolmogorov-Smirnof goodness-of-fit test statistic on the Zipf, exponential, and lognormal distribution of frequency of tokens, for each tokenization. We determine a specific tokenizer is Zipfian if the test statistic indicates a higher significance on the Zipf distribution than the other two distributions considered. We hypothesize that both the RoBERTa and GPT-4 tokenizer will affect the shape of the token frequency distribution on seen and unseen data enough to deviate from a Zipf distribution. We believe this because BPE is designed to segment rare words into more common chunks, which may lead to a bias towards more common tokens than a Zipfian distribution would expect.

Finally, with this paper we hope to spur more research into the subject, with larger tests run among more tokenizers and larger corpuses. It is possible that a universal Zipf-like law exists for BPE token distributions that takes on a different shape than the word-level trend in human languages.

(Potential) Citations:

- The Psychobiology of Language
- Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books
- RoBERTa: A Robustly Optimized BERT Pretraining Approach