# On the Sophistication of Tweets About Heroin, Cocaine, or Marijuana

**Michael Chary B.S., Nicholas Genes, M.D., Ph.D, Alex Manini, M.D., M.S. FACMT**

*The Icahn School of Medicine at MOUNT SINAI*

**Department of Emergency Medicine**

## Introduction

Characterizing new and emerging drugs can be difficult, because information about those drugs is often unavailable through traditional means such as national surveys. Social networks such as Twitter have been shown to contain messages from recreational drug users, and might provide a source of information to characterize newer drugs. As a first step in this process, we sought to investigate Twitter message complexity.

## Research Question

**Do tweets discussing cocaine, marijuana, or heroin differ in their structure?**

## Methods

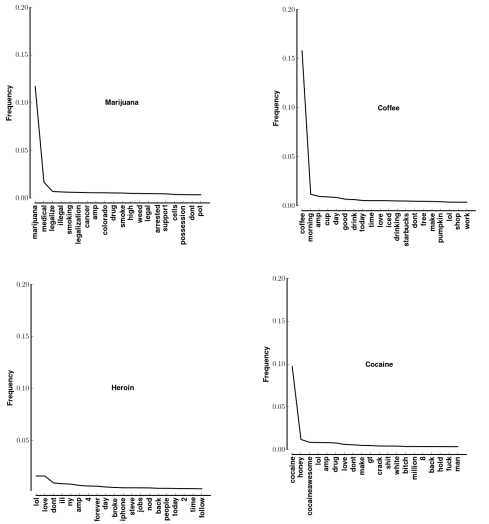
**Data Source** We analyzed two months of tweets from Twitter's streaming API for tweets that mentioned the words 'heroin', 'cocaine', or 'marijuana'. We excluded tweets that only contained links to websites.

**Preprocessing** Once acquired, all text was converted to lowercase, non-ASCII characters ignored, stopwords removed, and all words lemmattized. Lemmatization refers to converting variations on a word, such as 'ran','run', and 'running' into their simplest dictionary representations, e.g. 'run'.

**Flesh-Kincaid Grade Level** FK measures the difficulty of a reading passage on a scale that corresponds to the earliest U.S. grade level required to easily understand the passage.

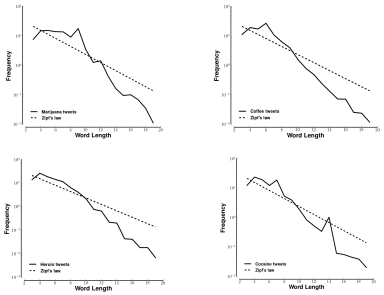
**Lempel-Ziv Complexity** LZ is the number of unique words, or more broadly symbols, in a reading passage. Whereas FK tracks the average length of words, LZ tracks the variance in word length.

## Results



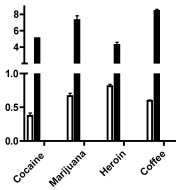

**Frequency of words in tweets containing drug keywords.** Each panel shows the frequency of words in all tweets containing that keyword. To allow comparison, the frequency for each category is normalized so that the area under each curve is 1.

## Results




**Relationship between word frequency and word length.** Similar to the previous figure, the solid line in each panel shows data for all tweets containing a certain keyword. The dashed line indicates the distribution of word lengths as predicted by Zpif's law for Standard American English. All y-axes use a logarithmic scale.

**Lempel-Ziv Complexity and Flesch-Kincaid Grade Level.** The clear bar show the Lempel-Ziv Complexity and the solid bar show the Flesch-Kincaid grade level. The y-axis jumps from 1 to 3 to display two measures with different scales on the same axis.




## Limitations

**Biased Sampling** Drug users who tweet may be a minority of drug users. Or, they may act differently than other drug users.

**Abbreviations** Internet users typically abbreviate long words and phrases. We labeled both words and abbreviations of a given length *words*. The use of abbreviations may explain the deviations from Zipf's law

**Level of Description** These measurements can characterize require more data than single tweets provide.

## Conclusions

The first panel suggests that a **few keywords dominate tweets about marijuana, cocaine, and coffee.** In contrast, no words dominate the tweets about heroin. The x-axes in each panel are different because the most common words for each group of tweets did not substantially overlap.

The second panel suggests that the **distribution of word length in tweets differs substantially from Standard American English**. In particular, shorter words are slightly more common and longer words are much less common.

The third panel suggests **that Lempel-Ziv complexity captures unique information about the word choice in each drug class**. Neither Lempel-Ziv complexity nor the Flesch-Kincaid grade level can distinguish tweets about coffee from those about illicit drugs.

Our results suggest that **tweets containing explicit references to drugs show systematic differences in measures of textual complexity**. These measures can be calculated automatically and may help efforts to mine social media for toxicologic data by helping to filter out irrelevant tweets.