2016

Quantifying Cultural Changes Through A Half-Century Of Song Lyrics And Books

Robert Bruce Woodward III
University of Vermont

QUANTIFYING CULTURAL CHANGES THROUGH A HALF-CENTURY OF SONG LYRICS AND BOOKS

A Thesis Presented

by

Robert B. Woodward, III

to

The Faculty of the Graduate College

of

The University of Vermont

In Partial Fullfillment of the Requirements
for the Degree of Master of Arts
Specializing in Mathematics

October, 2016

Defense Date: May 2, 2016
Thesis Examination Committee:

Christopher Danforth, Ph.D., Advisor
Sean Field, Ph.D, Chairperson
Peter S. Dodds, Ph.D.
Cynthia J. Forehand, Ph.D., Dean of the Graduate College

Abstract

Music is an ever-changing cultural reflection. It is deeply integrated into our society, ubiquitous in movies, television shows, restaurants, sport venues, churches and a plethora of other places. This thesis proposes that we consider the lyrics in popular music, as determined by Billboards Hot 100 chart, as a natural medium to analyze the changes in culture over the past half-century. Using this collection of lyrics, we analyze the change in relative frequency of individual words over time, and compare to works of literature. Furthermore, we use the ranking in the Top 100 as a metric with which to explore the relationship between usage of particular words and the popularity of the respective songs. We find that our data coincides with a previous hypothesis that the relative happiness of lyrics has decreased over time, and find that this also applies to the relative happiness of popular music.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction and Literature Review

The English language has evolved from its original incarnation in many quantifiable ways and an exploration into these changes reveals certain aspects of the culture of the people at each period of time. Here, we will explain why analyzing the composition of music lyrics over the past half-century can also reveal cultural changes over that time as well as support the notion of ever-changing language. We will also discuss existing literature that similarly analyzed words used in various mediums over time.

1.1 Introduction

Words fluctuate in both meaning and usage over time at various rates. This is rather obvious to most when reading books from such authors as William Shakespeare or Geoffrey Chaucer and comparing language used then to language used now. However, sometimes words fall out of common use so quickly that a word used often just ten years prior could seem archaic to those observing with hindsight. Other words change their meanings in such

a drastic way that many analysts spend their time researching how exactly words such as "bad" or "sick" went from their typical negative connotation to being descriptors for positive things. Much of the research into the evolution of words relies on their usage in books or other printed documents, such as letters or newspapers, since these are quite reliable references for the language of the respective era. However, using many of these types of documents can lead to misinterpretation of the common language.

Though William Shakespeare and John Bunyan were two of the most prominent literary figures of the 1600's, it is doubtful that the common populace spoke in a manner similar to how those gentlemen wrote. If it was the case that any given Englishman in 1600 spoke similar to Henry V in Shakespeare's play of the same name, what proof would we have outside of second-hand or third-hand written accounts. For quite some time, it was near-impossible to record language change based on spoken word outside of informal letters, for obvious reasons. With the invention and implementation of recording equipment, keeping track of information through a "voiced medium" from the past century has become viable. We can easily compare such films as Casablanca (1942) and Cast Away (2000) and see how spoken language has changed and evolved over the decades. What we aim to analyze here is the progression of words according to their usage in popular music.

Music as an art form has existed for much longer than the recording equipment mentioned earlier, but with the spread of radio, television and other devices of the like, music has become much more readily available to the common populace and thus exploded in variety and amount. Also, since most individual songs are typically just a few minutes in length, the artist is almost forced to keep language simple, at least compared to classic or even modern prose, meaning it is more likely to be applicable to comparison with common speech. Thus, it behooves the curious to specifically address the evolution of word usage

or frequency in music to come to a conclusion about the evolution of common language as a whole.

1.2 Related Works

Analyzing the evolution of language by looking at word frequency is not a groundbreaking topic and has been studied for many years, even before the implementation of computers that make the gathering of words and phrases much easier. In 1944, what is considered the first large-scale corpus of word frequencies were collected by hand into what was called the Teacher's Word Book of 30,000 Words, featuring approximately eighteen million hand-collected words (Nation and Waring 1997). This corpus, although now outdated due to the many years that have passed since its publication, has been used repeatedly as a reliable resource for word frequency analysis. For example, one early study used the Teacher's Word Book as a basis for studying word association and familiarity, conducting experiments based on the frequencies listed in the corpus (Nunnally and Flaugher 1963). This study was conducted in the 1960's and ever since, word frequency research has gained the interest of many, particularly with the creation of many more frequency collections that are available for researchers to access.

Though the Teacher's Word Book is still reliable and is impressive for its size, it is considered outdated at this time, over seventy years after its creation. For example, the evolution of language over time is primarily responsible for the change in connotation of the word "gay", who's meaning was primarily used to indicated happiness or joyfulness and has since evolved in a way to become connected to homosexuality. Likewise, words such as "blog", "groovy" and "microchip", who's invention and meaning were birthed after 1944, would obviously not hold a considerable frequency in the Teacher's Word Book. A

study on this evolution of language has been conducted using a more updated word corpus that collects words published in various forms of media in the past half millennium, in which the authors analyzed the evolution of verbs and the frequency of various people and topics to gain an understanding on how language and culture changes over time (Michel et al. 2011).

The authors discussed the change of verb usage, both irregular and regular, and explored the rate at which certain irregular verbs become regularized, that is change in common usage to the more regular conjugation such as the evolution of "burnt" to "burned". Though they were able to find a large group of previously irregular verbs that have become regular, and even one instance of a regular verb becoming irregular. They note the fastest instance of a verb shifting from irregular to regular, "chide", took nearly 200 years for its evolution to occur, a time frame we unfortunately do not have when studying popular music. The authors also discussed the evolution of ideas and topics over time, such as the word "influenza", who's usage is tied strongly to historical events relating to the spread of strains of influenza. These results reflect the culture of the people at the time, to discover what were considered the important topics or events, and comparing to other periods.

Another study conducted explores not just the words themselves and their relative frequencies, but more importantly considers other words that tend to be attached to or linked with the words in question (Wijaya and Yeniterzi 2011). For example, consider the word "war". In the middle of the 19th century, one of the more common words associated with "war" would likely be "civil" due to the American Civil War of the 1860's. Likewise, in 1910 or 1920, the word "great" would likely have been a strong link with "war", as World War I was known as the Great War, at least until World War II broke out, in which case "great" would likely transition out of association. This concept is what the authors

wanted to analyze, looking at the words immediately proceeding and following each word in question and analyzing the frequency of these associated words. This gives a deeper insight as to how the words are used and in what context rather than simply looking at the words themselves and trying to draw conclusions about cultural shifts from their individual frequencies.

However, it is important to note that the previous two studies relied on a single corpus, the Google Books NGram dataset, whose validity has been called to question lately due to the nature of the dataset itself (Pechenick et al. 2015). A recent study of the dataset reveals various issues that must be noted before making certain claims using the dataset, specifically claims as to the nature of the culture at a specific date and time. The authors establish three key issues: 1) the method used to gather the information, 2) the unintentional and misleading rise of certain types of documents within the dataset, and 3) lack of accounting for the popularity of given works.

Of the three issues, the last is the most relevant to this study, as here we aim to deal with popularity in music. Pechenick explains that any given popular work, such as George Orwell's *1984* or Arthur Miller's *Death of a Salesman*, is only recorded once within the corpus. This may seem harmless, but when attempting to ascertain the important words or topics of a given time, one must include some degree of popularity to get a true understanding as to the important issues to the populace. For example, the two novels previously mentioned were both published in 1949 and were both immensely popular, becoming two of the most important texts at the time. However, each one only occupies a single entry in the corpus, the same as Margaret Wise Brown's children's book *The Color Kittens*, implying each hold the same amount of significance, which is obviously fraught. For this reason, and others including the rise of scientific articles and documents in Google's NGram cor-

pus, the authors recommend the usage of the English Fiction subsection of the corpus when attempting to make statements concerning the state of culture over any given period.

Two of the authors who discussed the issues of the NGram dataset have also done extensive research relating to the frequency of words in various other datasets, including song lyrics and song titles for a large collection of songs released between 1961 and 2008 (Dodds and Danforth 2009). The authors used their datasets to compare to a previous study relating the perceived scale or degree of happiness for a large collection of words and made conclusions regarding the happiness of overall song lyrics. They analyzed the dataset in various ways, including considering each song or artist's respective genres, and came to a general conclusion that the happiness of song lyrics has slowly deteriorated over time. This conclusion was supported by considering the words whose frequency changed the most over time, a concept they refer to as a word's valence shift. Of the top ten words with the highest valence shift, only one, "lonely", contributed to a boost in song lyric happiness over time due to the general decline in frequency of "lonely". The process by which this is determined is by using what is known as the Jensen-Shannon Divergence, which measures the degree of divergence between datasets (Pechenick et al. 2015). This process will be discussed in depth later in this paper, as we employ the same process in order to compare to the authors findings on this topic.

It should be noted that despite looking at separate genres, Dodds and Danforth also did not factor in popularity, which this paper will address. There have been other works that discuss similar topics for other subsections of music, such as a tool that was developed to analyze the change in focus of rap music since its conception in the late 1970's and 1980's (Rap Stats).

Chapter 2

Methods

Using the Billboard Top 100 Song Data, sometimes known as the Hot 100, we will begin to analyze the relative frequency change of various words over the past half-century. However, gathering the target lyrics for every song within that time period proves difficult due to the inconsistency of user-entered data, leading to some songs and their respective lyrics not being gathered. Using the lyrics that were gathered, we then begin to create relative frequency graphs to analyze the change in usage of each word, as well as Jensen-Shannon Diverge representations to quantify the difference between various periods of time within the past half-century.

2.1 Dataset

The experiments and analysis used in this paper largely revolve around the dataset of retrieved lyrics for the "Top 100" Songs, according to the charts constructed by Billboard. Billboard is the most prominent collector of music information based on popular consen-

sus, compiling and distributing charts depicting the top songs in various genres, and so is an obvious choice to use as our basis for what are the "most popular" songs. The charts were downloaded directly from the Billboard.com website (Billboard).

Originally, we analyzed the overall Top 100 songs per year, leaving us with just under six thousand songs to consider, most of which had lyrics we could access. This dataset proved to be ultimately too small to make any firm conclusions, so after some deliberation, it was decided to instead analyze the Top 100 songs Weekly data, which provides just over fifty times as much information.

To discover how Billboard constructs their popularity lists, we spoke with Alex Vitoulis, the Research Manager for Billboard at the time, who informed us that the way they collected their data differed from the method used at company's inception. When Billboard began releasing general Top 100 Charts in the 1950's and 1960's, the lists were created by the company based on their opinions on what the top songs were and were not based on any hard data. In the early 1970's, Billboard changed their analysis and now releases the Top Song charts based on record sales, radio play, and other factors, making the charts released after that period much more reliable. Because of this, much of the research we do focuses on music in the Top Charts following 1970 since that information is more reliable in terms of evaluating public opinion.

Though the range of years analyzed had been diminished about fourteen years, we are still analyzing one hundred songs per week from 1970 to 2014, the only complete and reliable years at the time of this research. This means that we are considering around 200,000 different songs, not including repeats. However, even if a song was repeated through many weeks, that merely means it was popular even longer than other songs, and its importance

needs to be considered more than once for the initial week it entered the Top 100, so all repeats will be considered for every instance of their repetition.

Year	Percentage	Year	Percentage	Year	Percentage	Year	Percentage
1970	56.6	1981	77.5	1992	78.3	2003	94.0
1971	59.6	1982	77.7	1993	78.6	2004	95.1
1972	66.9	1983	83.9	1994	81.9	2005	97.5
1973	68.9	1984	79.8	1995	75.4	2006	93.7
1974	68.0	1985	88.7	1996	69.8	2007	94.3
1975	64.3	1986	85.4	1997	75.2	2008	94.0
1976	70.1	1987	82.7	1998	85.7	2009	92.9
1977	75.6	1988	84.9	1999	96.8	2010	93.8
1978	76.8	1989	82.7	2000	99.2	2011	94.8
1979	74.4	1990	78.6	2001	94.4	2012	94.2
1980	73.5	1991	77.1	2002	92.7	2013	92.5
						2014	94.4

Table 2.1: **Percentage of Billboard Top 100 Lyrics Collected Per Year.** This table depicts the percentage of collected lyrics per year between 1970 and 2014. Because not all were found every year, or even a consistent percentage found per year, the data moving forward would have to be normalized to account for this inequality.

After the charts were collected, we needed to access the lyrics for every song in the Top 100 charts. To access these lyrics, we attempted to find the lyrics by accessing four different lyric websites: AZLyrics, MetroLyrics, SongLyrics and OldieLyrics. These four sites all rely on user-submitted lyrics, suggesting that the song in question has a large enough following to be included on the sites. However, since we are analyzing the Top 100 songs, we did not expect this to be an issue. Unfortunately, we were not able to find lyrics for every song searched, specifically due to the fact that the songs on the sites are all user-entered. For example, a song that charted by the artist "Tom Petty and the Heartbreakers" could be submitted as by "Tom Petty and the Heartbreakers", "Tom Petty & the Heartbreakers" or simply "Tom Petty". Because these is no standard system for instances such as these, it is difficult to search for and attain every lyric possible. Thankfully

we were able to consistently find over 70% of the lyrics for every year since 1976, a value that only increased as years became closer to the present, as depicted in Table 2.1. In total, about 84.4% of lyrics were found, or about 193,000 songs consisting of just over 69,000,000 words and about 60,000 unique words. Specifically, the most used word is the word "you" at approximately 2.73 million uses between 1970 and 2014.

In addition, we also accessed the Google Books NGram dataset, which is a collection of words used in millions of books stored in the Google Books database. This dataset is helpful to produce time-series graphs that depict the usage of a given word or words compared to all words used in each year, making it a natural comparison tool against the time series graphs made from the Lyric dataset. Unfortunately, the Google Books NGram dataset only goes to 2008 so we cannot use it to directly compare to the most recent songs, but we can still use the dataset for the appropriate time period of 1970 to 2008. Also, we would like it if the NGram dataset had a similar relation so that we may compare to similar popular books at the time, perhaps based on New York Times best seller lists, but since the dataset reportedly has data for 4% of all printed books, it is more likely that the books included in that dataset are the more popular or well-known ones, making for more natural and safe comparisons.

The NGram dataset includes information for a wide array of written works in many different languages. We are specifically using the "English Fiction" dataset, as that seems to be the most appropriate dataset to compare to song lyrics instead of other English-focused sets such as "English", which includes many scientific articles and documents that, for the most part, do not compare well to lyrics. Additionally, when considering any given word, it will be considered case-insensitive and thus will include all capitalization variations of the word.

2.2 Process

In order to examine both the dataset gathered from the lyric websites mentioned previously and the NGram dataset, we first needed to devise a way to meaningfully depict the information in a graphical manner. The dataset that was gathered included the amount of different types of words used per week. However, the total amount of words differed every week, party due to varying song lengths and also due to the fact that not all songs were gathered every year, such as the case where approximately 94.4% were gathered in 2001 and approximately 92.9% were gathered in 2009. In order to compare all of these together, we depended on the relative frequency of each word per week, meaning that the count for each word would be directly related to the number of words gathered in that week.

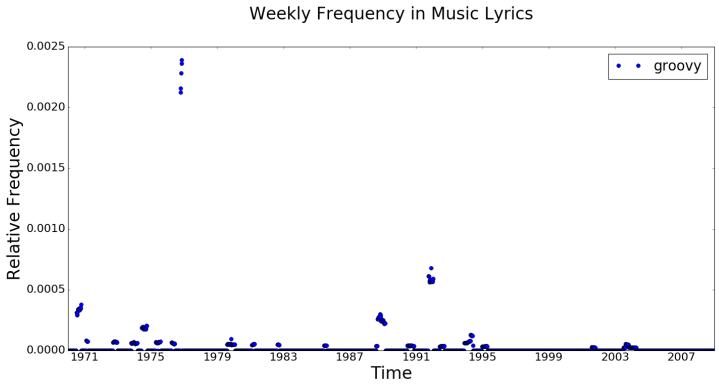

Figure 2.1: **Weekly Frequency of "groovy" in Music Lyrics.** This plot depicts the relative frequency of the word "groovy" between 1970 and 2008. It can be noted that "groovy" was used, albeit sparsely, between 1970 and roughly 1995, but then became almost non-existent, save for a few weeks in the early 2000's. Notice the odd frequency trend around 1976/1977, where the frequency jumps very high for a short amount of time and fluctuates between 0.0021 and 0.0024. This is attributed to the song "Groovy People" by Lou Rawls which uses the word 38 times, whereas the small spikes in 1988 and 1992 are attributed to "Groovy Kind of People" by Phil Collins and "Groovy Train" by The Farm, respectively.

This Relative Frequency plotting yields more helpful results, but it also reveals something that must be noted, namely a small sense of interpretation. Consider the word "groovy", a word typically associated with the 1970's era of music. "Groovy" ended up being a very sparsely used word, usually just appearing once or twice in a song, leading to just a few positive relative frequencies over the entire timeframe considered. However, note the strange occurrence in Figure 2.1 where the frequency of "groovy" tends to fluctuate between a relative frequency of approximately 0.0021 and 0.0024. This should not be interpreted as a fluctuation in frequency of that word through the various weeks around that time but should be considered a result of calculating the relative frequencies. One must

be aware that even between weeks within the same year, the number of words and lyrics collected will vary meaning that even if a word such as groovy was used the same amount of times in one week as the next, the relative frequency calculated will likely be different between those weeks. Thankfully, because of the sheer amount of words collected, this variance only yields a small degree of relative frequency change and does not result in great concern.

For a majority of these relative frequency plots, words tend to rise and fall in relative frequency in a wave-like fashion, sometimes with very abrupt rising or falling, which is typically attributed to the inclusion or removal of a song or group of songs that prominently featured that word. This however tends to produce graphs with large amounts of noise that become difficult to decipher, particularly when directly compared with another word or groups of words in the data set, such as in Figure 2.2. For a majority of the graph, it appears the word "he" is more frequent, but this is still unclear when the frequencies become too similar, particularly in the 1970's and 1990's, the latter of which make the two words seem almost the same. To remedy this, we consider a "binned" version of all of these plots, taking the average frequency found over the fifty-two weeks in each year and plotting that average as a single point for each year. This process yields Figure 2.2, which is a much more understandable representation of the data over the same time period.

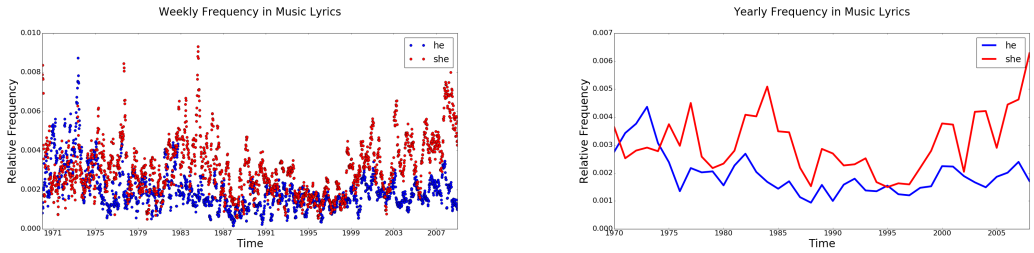

Figure 2.2: Frequency of "he" vs "she" between the years 1970 and 2008. The graph on the right depicts the weekly relative frequency, plotting each point according to that specific word's frequency in that given week. The graph on the right depicts the yearly frequency as a line graph, binning the collection of weekly points from the graph on the right into a single data point for ease of understanding.

The other reason we consider these binned versions of the relative frequencies is their relation to the NGram dataset, which similarly graphs a single relative frequency point per year considered. When comparing to the NGram dataset, we consider both the Pearson and Spearman coefficients to determine relatedness. Because the Spearman coefficient also considered the monotonicity of the sets in question, we suspect that will be the more accurate measure of comparison, but both will be considered regardless.

To quantify the differences between two given datasets, we will be employing the Jensen-Shannon Divergence which calculates how much one dataset diverges away from the other and, more importantly, what are the important individual contributors of this divergence. The Jensen-Shannon Divergence relies on the Kullback-Lieber Divergence, which also measures the difference between two datasets, but will always depict strong divergence if there exists an element in one set but not the other. To correct this, the Jensen-Shannon Divergence calculates a symmetrized version of the Kullback-Liber divergence,

$$JSD(P||Q) = \frac{1}{2}K(P||M) + \frac{1}{2}K(Q||M) \tag{2.1}$$

where $K(P||M)$ is the Kullback-Lieber Divergence of the set P given $M = \frac{1}{2}(P+Q)$. This calculates a numeric which explains how much one set diverges from another, but what we are more interested in are the individual contributions of each word in the dataset; i.e. the words that contribute the most to the divergence one way or the other (Dodds and Danforth 2009). To do this, we employ the following equation for each word i,

$$JSD_i(P||Q) = m_i \cdot \frac{1}{2}(r_i \cdot \log_2 r_i + (2 - r_r)\log_2(2 - r_i)) \tag{2.2}$$

where $m_i = \frac{1}{2}(p_i+q_i)$ is the average frequency of the word i in both sets and $r_i = p_i/m_i$ is the ratio of the contribution of the word in one set to the average m_i. In the case where $p_i = 0$, $r_i \cdot \log_2 r_i$ is set equal to zero, as $\lim_{x\to 0} x \log x = 0$.

2.3 Software

To both access the above information from the listed locations and to analyze the data gathered, all code was written in the Python programming language using Enthought Canopy and Jupyter. Much of the code to gather the datasets relied on the BeautifulSoup 3rd Party Module, which allows one to safely and cleanly parse HTML strings and access that which is important, such as the frequencies from each NGram site or the lyrics from each of the four previously mentioned lyric sites. See Appendix B for one of the functions prominently used in this analysis that gathers the information for a given set of words from each dataset and plots them with certain options included. This code produces and saves three different plots, the Weekly Frequency in Music Lyrics, the Yearly Frequency in Music Lyrics and

the Yearly Frequency in Books, as well as calculating and displaying both Pearson and Spearman coefficients for each pair of words or each word and its NGram counterpart.

Chapter 3

Results and Discussion

Using the relative frequency graphs as well as a devised weighting method to further apply popularity to our analysis, we begin to look at many specific words that either changed in usage drastically over the past half century or share a relatively strong correlation with the usage of that word in the Google NGram dataset. We then conclude with an comparison between lyrics in specific decades and the words that contributed the most to the difference between the high-ranked songs and the low-ranked songs in those decades.

3.1 Main Results

To begin the analysis, we needed a logical frame of reference to guide our investigations, some mechanism that instructs as to what words or groups of words to consider and analyze first. With the goal of analyzing how music has evolved over the given time period in mind, we initially examined the overall frequency shift of each word, gathering a list of the top 100 words whose overall absolute frequency shift was the greatest. To avoid producing a

list of singleton words that only appeared a few times throughout the entire time period, or "words" that are the result of user input error such as "\xe2\u20ac\xbdive", we stipulated that the list must only include words that appeared at least five times total between 1970 and 2008, thus eliminating many of the issues. Many of the top Absolute Ratio Change words are very common words, the top three being "you", "i" and "the", but some more interesting words that can lead to some interpretation, such as the fourth-greatest Absolute Ratio Change word, "love", are shown in Figure 3.1.

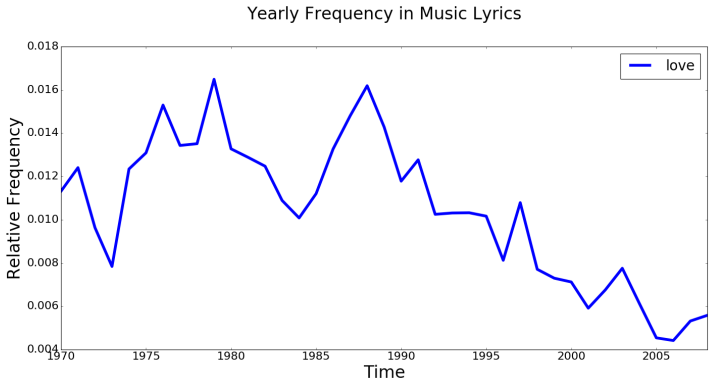

Figure 3.1: Frequency of "love" between the years 1970 and 2008 in Lyrics

"Love" is quite an interesting word to have such a sharp change in frequency over time due to what it could imply about the connotation of lyrics. Note that there is a sharp decrease in the frequency of "love" in the early 1970's but the relative frequency of the word increases again to a comparable level, with various spikes and dips until the final spike in 1987-1988. Every year following that is a marked decrease and, aside from small increases in frequency, consistently trends downward to an ultimate low in 2006. The years following 2008 follow this trend as well, implying that, in terms of decades, the 2000's and the 2010's, thus far, have a much lower level of relative frequency for the word when

compared to any decade prior, including the 1990's when the consistent decrease really began. With an average relative frequency of around 0.012 in the 1970's and 1980's and an average relative frequency of around 0.006 in the 2000's, the marked decrease in usage of the word becomes even more apparent.

To additionally analyze the word, as well as any other word, we began to consider each word's usage relative to the position of its respective songs in the Top 100 chart. For example, it could be true that "love" may have decreased in relative frequency, but perhaps it is still being used in more popular songs, indicating that it is still seen as an important word since more songs featuring that word rank higher in the charts. To do this, we assigned point values to every song in the Top 100 lists based on an inverse linear relationship starting with 100 points for each 1st place song down to 1 point for the 100th place song. Essentially, this leads to a simple formula for points of a given song, represented by

$$P_s = 101 - N_s \tag{3.1}$$

where P_s corresponds to the points assigned to song s and N_s corresponds to the Top 100 placing of song s. The idea here is to scale the word frequencies with a power law reflecting their popularity, in this case simply using their position on the chart as an indication of popularity. We also attempted to find a different weighting relation based on the number of times a song in a given position on the Top 100 was played on the music streaming site Spotify (Spotify). These attempts are shown in Figure 3.2, where we show this comparison for one of the more recent weeks, relative to the writing of this study, and also its log-log representation. Unfortunately, based on the log-log graph, a higher degree power law is not likely, so we will keep the law from equation 3.1.

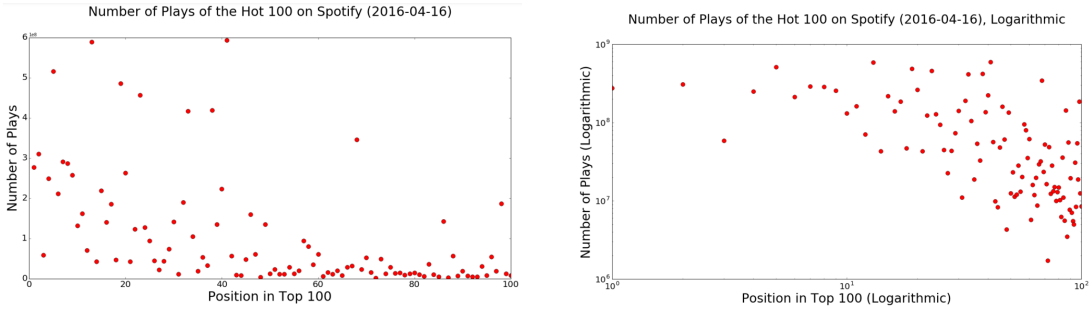

Figure 3.2: Number of Plays of the Hot 100 on Spotify (accessed 2016-04-16). The x-axis depicts the song position in the Billboard Top 100, and the y-axis depicts the number of times the song has been streamed on the music streaming site Spotify as of April 16, 2016 (Spotify). The left plot axes are linearly spaced wheras the right plot aes are logarithmically spaced; the right plot suggests a power law distribution for popularity.

This weighting was then applied to each word per song to achieve a new total frequency of each word per year. For example, say the word "groovy" appeared only twice in the year of 1973, in the third and seventy-fourth ranked song in a particular week. This would normally mean it would have a frequency count of two, but with the new weighting, "groovy" would have a frequency of $P_{s_1} + P_{s_2} = 98 + 27 = 125$ for that year. As before, the new weights were then divided by the new total weights to achieve the new weighted relative frequencies. Figure 3.3 depicts the new weighted relative frequency of "love" plotted using a dashed line against the previously-used relative frequency.

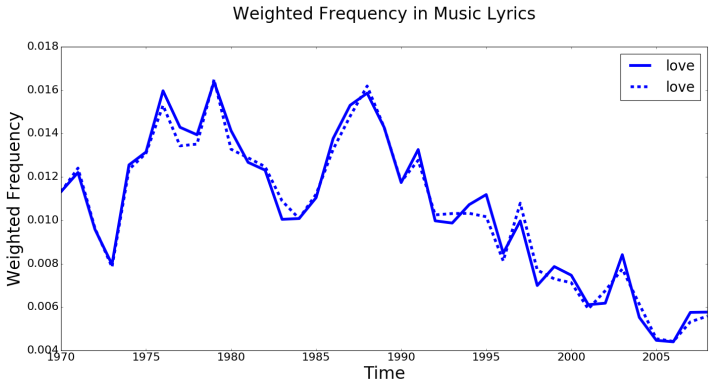

Figure 3.3: Frequency of "love" between the years 1970 and 2008 in Lyrics. Includes the weighted form of each year's relative frequency designed by the dashed line

Note that if the weighted relative frequency falls below the standard relative frequency for a given year, that implies that the word tended to be used more often in lower-ranking words, i.e. those closer to rank 100, such as in the case of "love" in the mid-late 1970's in Figure 3.3. The opposite also holds, so if a weighted relative frequency exceeds a standard relative frequency, that implies the word tended to be used in higher-ranked songs more often. However, as noted in Figure 3.3, this weighted relative frequency tends to not deviate from the standard relative frequency much, and sometimes even follows nearly perfectly along with one another. This means that we do not have much evidence to claim that "love" was used more or less often in higher-ranked songs, nor in low-ranked songs for that matter. as its relative frequency decreases, meaning that our claim of "love" decreasing in popular usage is strengthened.

Does the decrease in "love" imply that we tend to listen to less love-oriented songs than we did in the 1970's and 1980's? Not necessarily, it simply implies that the word itself is used less, relative to other words. For example, in 2006 the Irish rock band Snow Patrol released their Grammy-nominated love ballad "Chasing Cars" which lasted an astounding

166 weeks on the Billboard Top 100 charts, peaking at Number 6 in 2006. However, the lyrics of this love song fails to use to actual word "love" itself at all, instead describing it in other ways, meaning that this incredibly popular love song would not even contribute to the usage of "love" during any of the 166 weeks it was in the Billboard Top 100 and, in fact, would actually decrease the relative frequency of "love" over that time period by contributing over 100 non-"love" words to the corpus. Until we are able to read every lyric and gather the connotation or meaning of the lyrics as concepts, these graphs should not be interpreted in that fashion.

For further comparison, let us examine how "love" has been used in another medium, the English Fiction NGram dataset provided by Google. Comparing the standard relative frequency to the google books relative frequency below in Figure 3.4 leads us to believe there exists some slight inverse relationship, particularly during the 1990-2008 period of time when the standard relative frequency. To calculate the comparison, both Pearson and Spearman's correlation coefficients were calculated, resulting in a Pearson coefficient of -0.2786 and a Spearman coefficient of -0.258125, indicating that the two frequency graphs are inversely related, though not very strongly, meaning we cannot make any strong claims relating the two datasets.

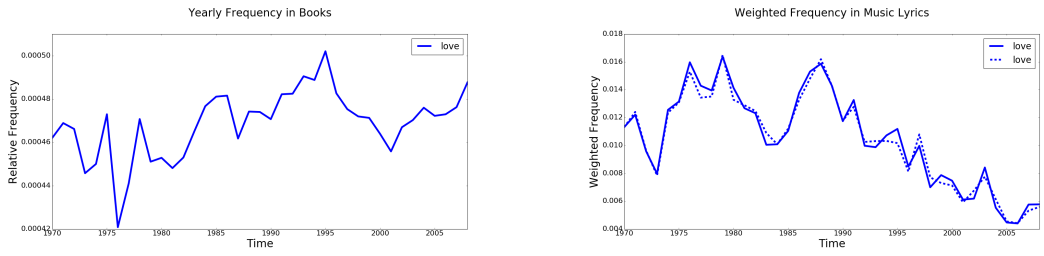

Figure 3.4: Frequency of "love" between the years 1970 and 2008 in Books vs Lyrics. Note that the y-axis is an order of magnitude smaller for books.

Though the "love" relative frequency plots did not relate very well to their Google Books counterparts, others lend themselves for much better comparison, particularly swear words. In fact, when calculating the average of the two correlation coefficients, the word with the highest absolute correlation between its standard relative frequency graph in lyrics and its Google Books relative frequency graph is the (censored) word "s***", with an average correlation of 0.853, indicating a strong positive relationship. This should not be all too surprising, particularly when you look at the frequency plot for "s***" and notice the word was practically non-existent in music prior to 1990, which also holds true for a plethora of other swear words. In fact, other swear words, such as "s***ty" and "bustera**", also are among some of the highest positively correlated words when compared to the Google Books dataset, and other swear words tend to be highly correlated with each other within the lyric dataset, indicating a relationship between the usage of swear words in the two forms of media and between each other within the lyric dataset. Again, this is not all too surprising, but could lend itself to further study analyzing how certain swear words became so commonplace in such a short amount of time in popular music, possibly one of the stronger examples of the evolution of language within a short timeframe.

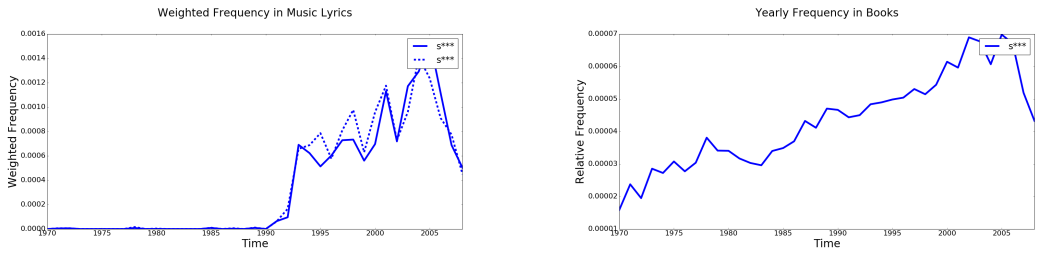

Figure 3.5: Frequency of "s***" between the years 1970 and 2008.

For more information on the Top Absolute Ratio Change words and the Top 100 Positive and Negative Average Correlation words between the lyric dataset and the Google

Books dataset, see tables A.1 through A.6 in the appendix. Motivation for Figure 3.6, Figure 3.7, and Figure 3.8 are all derived from Tables A.3 and A.5.

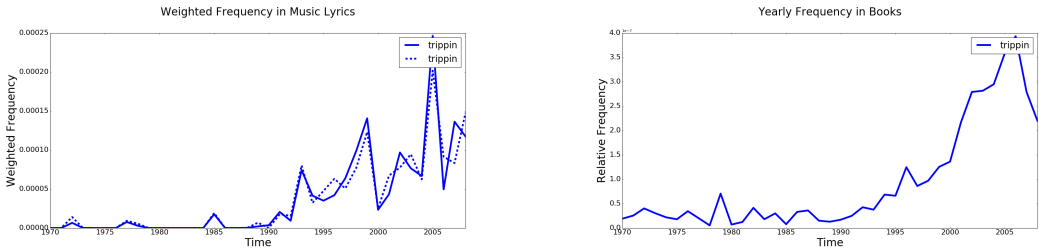

Figure 3.6: Frequency of “trippin” between the years 1970 and 2008.

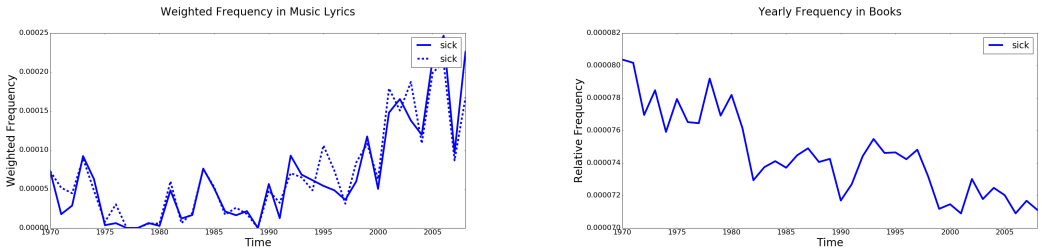

Figure 3.7: Frequency of “sick” between the years 1970 and 2008.

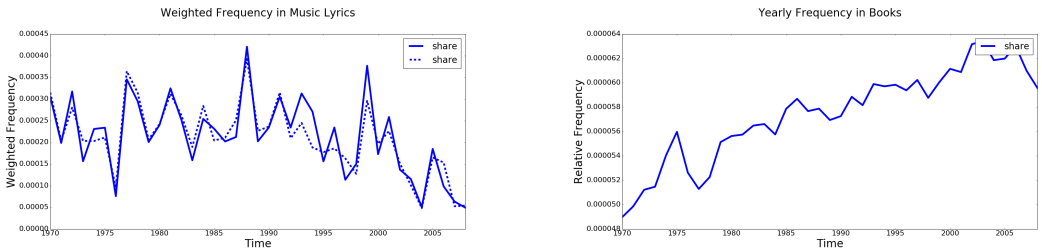

Figure 3.8: Frequency of “share” between the years 1970 and 2008.

Before we venture on into the final exploration into the lyrics database, we wish to cover one last topic pertaining to the relative frequency of certain words in lyrics. According to

Dodds and Danforth's paper from 2009, they discovered that, when analyzing lyrics of all genres over the similar time period, music has generally become more sad (Dodds and Danforth 2009). They came to this result using another dataset that included a happiness rating for a large amount of words found within their lyric dataset. In their result, they listed a group of eight words whose relative frequency shift over time contributed to this shift in happiness of lyrics: love, baby, home, lonely, hate, pain, dead and death. The two claim that the general decrease in usage of the first three words and the general increase in usage of the other five contributed more to this result than any other group of words.

The word "love" was analyzed previously, but we wish to address the other seven of these "Top 8 Words", as they henceforth shall be called, as well as depict their change over time in terms of relative frequency. The first two words mentioned, "love" and "baby", are by far much more frequent in lyrics than the other six, so the graphs shall be split into two to retain as much information as possible. In Figure 3.9, we can see that love is noticeably decreasing over time, as mentioned earlier, but the other word changes are not as clear. In fact, they all seem to settle around particular values of relative frequency and do not deviate in a way that we would deem as meaningful. Even "baby", which does vary in relative frequency over the time period, seems rather settled around a relative frequency of 0.005, though it should be noted that the relative frequency does decrease gradually along with "love" since 1990. However, from this, we cannot divine any distinct conclusion to correlate with Dodds and Danforth's Top 8 aside from the one word "love", which seems to follow along with their conclusions.

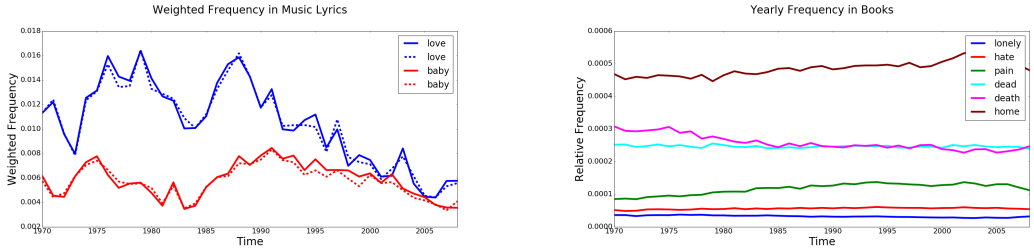

Figure 3.9: Frequency of the Top 8 words between the years 1970 and 2008

Though the Top 8 words did not lead to some strong conclusion as to whether popular happiness levels follow or deviate from the happiness of all songs, we can still approach the popular songs in another method to try to ascertain how popular songs change over time. We will now consider the Jensen Shanon Divergence models, much like those used by Dodds and Danforth, to illustrate the relative frequency difference between songs of varying levels of popularity of different time periods.

In terms of varying levels of popularity, we will be comparing how the lyrics of songs in the lowest Billboard ranking evolve from one time period to another with the songs in the highest Billboard ranking. Recall that Table 2.1 showed that various percentages of all lyric data was actually gathered, so more specifically this comparison will be comparing the lowest and highest Billboard ranked songs of those that were gathered per week. For example, if we were to use the Top 10 songs per week from the Billboard Top 100, if the Number 8 ranked song was not gathered, then the next highest rank would be used in its place. A similar situation would hold for a missing member of the Bottom 10, using the next-lowest ranked song.

Testing various amounts of Top and Bottom songs to compare, anywhere from 5 to 25 songs per group, we decided to settle with the Top and Bottom 10 Popular Songs of Billboard Top 100 lyrics gathered. Using the Jensen-Shannon Divergence testing, we col-

lected each word used in the Top and Bottom songs in the time period and calculated their relative frequency and determined the set of words that contributed the most to the overall frequency difference between the groups. Here, if a JSD Contribution factor is positive, then that means that it is a word that has a relative frequency greater for the Top 10 Songs, whereas a negative JSD Contribution factor implies a relative frequency greater in the Bottom 10 Songs. Using this method, we considered the frequency shift between decades, such as the frequency shift of words in the Top and Bottom 10 songs in the between 1980 and 1990 or from between 1970 and 2010.

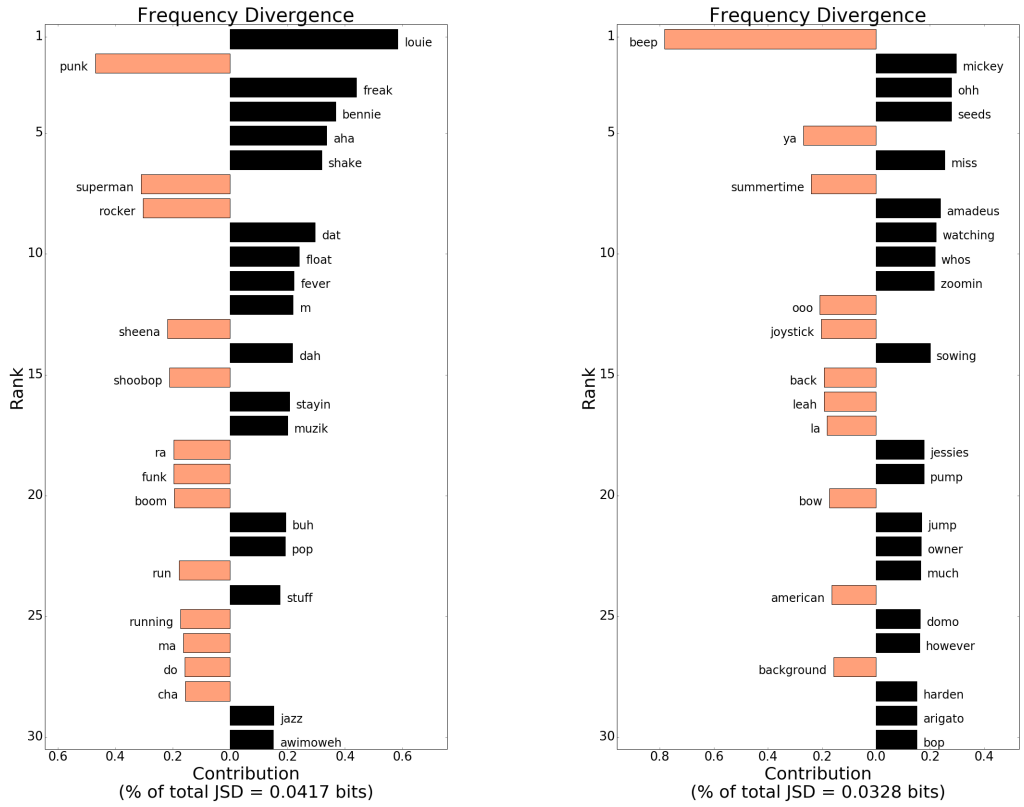

Figure 3.10: The Jensen Shannon Divergence for the Top and Bottom 10 songs present between 1970 and 1980 (left) and between 1980 and 1990 (right). Words that have a positive JSD Contribution direction are those whose relative frequency is greater in the Top 10 songs compared to the Bottom 10 songs. The reverse holds for JSD Contribution words and values in the negative direction.

In Figure 3.10, we have two representations showing the Jensen Shannon Divergence of the Top and Bottom 10 songs within two different decades. From here, we can see some interesting occurrences, such as in the 1970-1980 Divergence, we see that the genres or

styles of “funk” and “punk” both show much higher frequency in the Lower 10 songs. Once may compare this fact to the Top 10, whose common words include more dance words such as “shake” and “pop” and “freak”, the latter of which deriving mainly from Chic’s disco smash hit “Le Freak” from 1978. This comparison could imply that during that decade, the Top 10 songs tended to be dominated by more dance-oriented songs, despite the popularity of “funk” and “punk” oriented tracks.

However, one should remember that context is still lost in this frequency comparison, we are simply looking at what the usage of “punk” could be, another easy assumption could be songs that describe people using the word “punk” rather than being within the punk genre. One should also note that one song can significantly impact these types of frequency representations by repeating a word not often used in other songs, such as “Le Freak” repeating “freak” many times , “amadeus” deriving from “Rock Me Amadeus” by Falco and “whos” and “watching” strongly deriving from the chorus and bridge of Roswell’s “Somebody’s Watching Me”. The combination of the plots from Figure 3.10 is shown in Figure 3.11, where certain words with strong frequency in a particular set of songs either change in ranking or disappear completely, such as “beep”, which was present in the 1980-1990 ranking but not for the 1970-1990 ranking for either Top or Bottom 10.

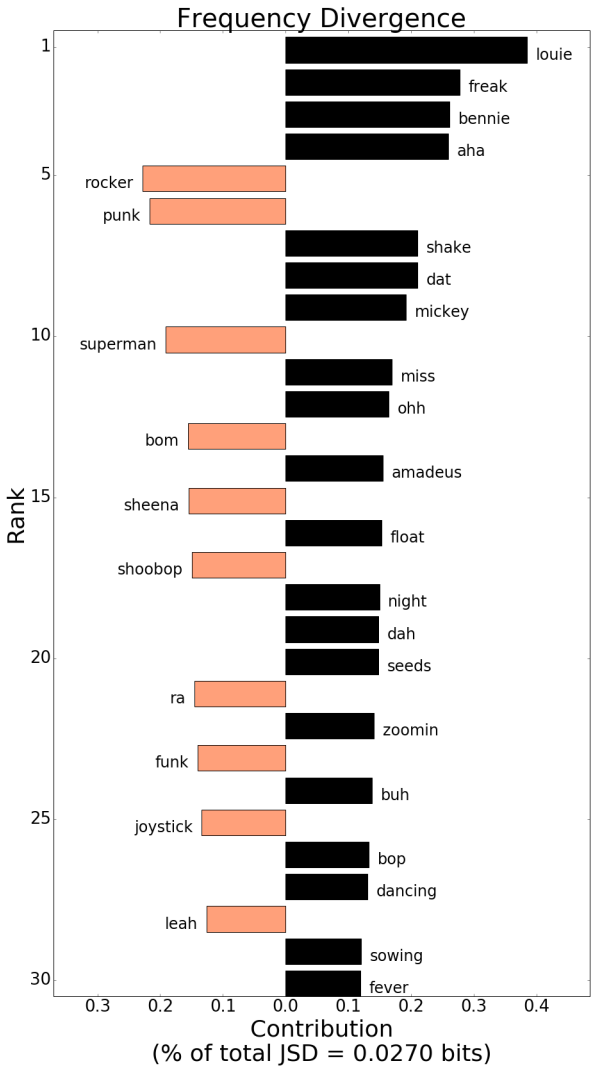

Figure 3.11: The Jensen Shannon Divergence for the Top and Bottom 10 songs present between 1970 and 1990. This plot is a combination of the two decades from Figure 3.10, but introducing the word “rocker” as a popular word in the Bottom 10 songs along with “punk”, possibly indicating some split between genres in the Top and Bottom 10.

Now looking at all Top and Bottom 10 songs between 1970 and 2010 in Figure 3.12, we find some other interesting circumstances, particularly when comparing to Figure 3.11. Note that in this plot,the word “funk” is still one of the more common words in the Bottom 10 songs. This was not only due to its high frequency in the Figure 3.10 graphs, but also from its high frequency in the 1990s as well. From here we can make the claim that “funk” songs, or at least words that specifically use the word funk, tend to be less popular than others.

A more interesting relation is the types of words that start appearing in the more popular songs. Note in Figure 3.12, words such as “duuh”, “na”, “ayy” and “badoop” tend to have higher frequency in the Top 10 songs, which is odd due to their arguably nonsense meaning. This most likely is the result of errors in user-entered data, since all lyrics are entered by people on each of the lyric sites, but no word was considered if it was used less than five times, so each of those nonsense words had shown up more than once and enough to affect the overall relative frequency. It is also important to note that these nonsense words do not start appearing in the JSD graphs until the 1990s, when words like “badoop” start showing up, but even more so in the 2000’s where a majority of the highest JSD contribution words for the Top 10 songs are nonsense or slang words. From here, one can start hypothesizing about the possible “slangification” of popular lyrics where large amounts of nonsense or slang words start appearing in the Top 10 songs.

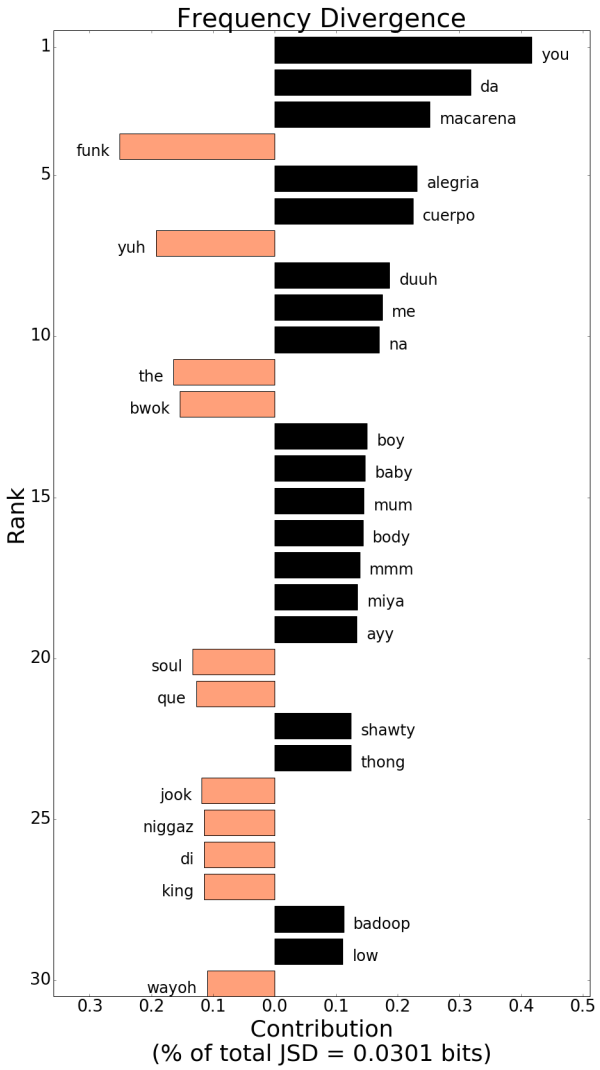

Figure 3.12: The Jensen Shannon Divergence for the Top and Bottom 10 songs present between 1970 and 2010.

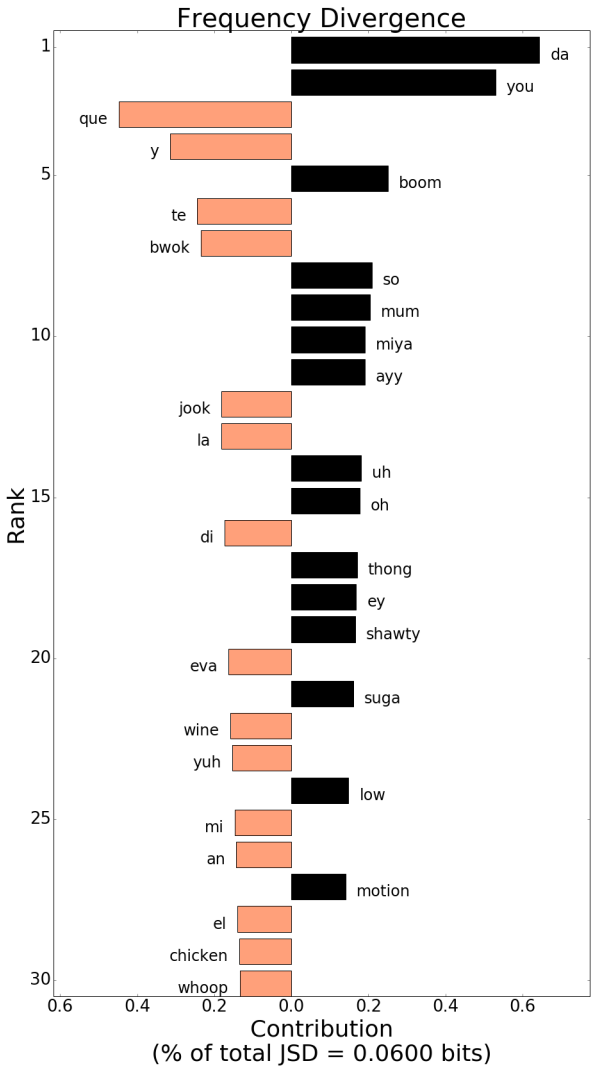

Figure 3.13: The Jensen Shannon Divergence for the Top and Bottom 10 songs present between 2000 and 2010.

3.2 Discussion

Analyzing the frequency shift of words in both books and music over time is a very useful identifier for analyzing the culture of the people who produced the given mediums. The practicality of collecting and analyzing the words that constitute these mediums can lead to some loss of contextual information and themes, but the information produced by the frequency of particular words could reveal more subtle changes that would be otherwise overlooked. We have shown that some commonly used words such as "love" has deceased, relatively, in music lyrics over time, more than any other word, leading to various conclusions. This decrease could be indicating that we as a culture have moved further away from the concept of "love" or it could be true that we have simply opted in favor of other words or phrases with a similar meaning, producing the concept of love without actually using the word itself.

We have also looked into the difference between the high-ranked songs and the low-ranked songs in the Top 100 in a given time period, attempting to find a connection between usage or particular words and the probable popularity of the song. For instance, Figure 3.12 indicates that the words "da" and "na" are more often used in the Top 10 songs compared to the Bottom 10 songs between 1970 and 2010, so a song using these words is more likely to be in the Top 10 than the Bottom 10. This information could be useful to professional lyricists behind artists who are attempting to create a very popular song. Having knowledge of these "hot" words can lead to what is perceived as good songwriting by the popular opinion.

Further research into this topic would likely find a stronger method of weighting the words. We used a simple linear function, weighing the words based solely on their position

in the chart. However, we suspect an exponential function to be more reflective of this relationship based on Figure 3.2. Deriving a more accurate weighting method could change the relative frequency graphs shown previously, though most likely by a small amount, as the weighted relative frequency deviated very little from the normal relative frequency.

Lastly, we suspect the next course of action would be to analyze groupings of words to hopefully integrate the context of the songs into the relative frequency of the words. For example, consider the word “love” that we have shown many times previously. Looking at the words used in the immediate vicinity of the word “love” in the Top 10 songs could reveal more as to the change in context of the word “love”. This concept was analyzed well for the Google NGram dataset and helped support the claim of the evolution of language over time, but has yet to be applied to the dataset of music lyrics (Wijaya and Yeniterzi 2011).

Bibliography

Billboard. www.billboard.com. Accessed: 2016-03-15.

Rap stats: Breaking down the words in rap lyrics over time. http://genius.com/Sameoldshawn-rap-stats-breaking-down-the-words-in-rap-lyrics-over-time-annotated. Accessed: 2016-01-28.

Spotify. https://www.spotify.com/us/. Accessed: 2016-04-17.

Dodds, P. S. and C. M. Danforth (2009, July). Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. In *Springer*.

Michel, J.-B., Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, T. G. B. Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden (2011, January). Quantitative analysis of culture using millions of digitized books. In *Science*.

Nation, P. and R. Waring (1997, September). Vocabulary size, text coverage and word lists. In *Vocabulary: Description, Acquisition and Pedagogy*.

Nunnally, J. C. and R. L. Flaugher (1963, May). Psychological implications of word usage. In *Science*.

Pechenick, E. A., P. S. Dodds, and C. M. Danforth (2015, October). Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. In *PLoS ONE 10(10)*.

Wijaya, D. T. and R. Yeniterzi (2011, October). Understanding semantic change of words over centuries. In *Detect'11*.

Appendix A: Parameters

Table A.1: Top 100 Absolute Ratio Change Words (1-50)

	Word	Ratio Change ($*10^{-3}$)		Word	Ratio Change ($*10^{-3}$)
1	you	30.1	26	your	12.5
2	i	27.4	27	of	12.4
3	the	27.0	28	that	12.3
4	love	22.9	29	no	12.3
5	it	22.7	30	whoa	12.0
6	a	21.6	31	in	11.8
7	la	21.3	32	da	11.5
8	my	20.5	33	be	11.4
9	on	19.3	34	na	11.3
10	and	19.0	35	up	11.2
11	me	18.7	36	little	10.9
12	to	17.3	37	like	10.8
13	we	17.0	38	go	10.7
14	mow	16.8	39	got	10.7
15	oh	16.8	40	all	10.6
16	baby	16.6	41	um	10.2
17	shimmy	15.7	42	is	10.1
18	yeah	14.0	43	down	9.91
19	do	14.0	44	its	9.88
20	can	14.0	45	dance	9.58
21	don't	13.9	46	girl	9.56
22	im	13.8	47	what	9.55
23	ba	12.8	48	come	9.53
24	never	12.8	49	her	9.40
25	get	12.8	50	he	9.32

Table A.2: Top 100 Absolute Ratio Change Words (51-100)

	Word	Ratio Change ($*10^{-3}$)		Word	Ratio Change ($*10^{-3}$)
51	gimme	8.99	76	every	8.03
52	who	8.95	77	night	7.95
53	she	8.82	78	ooh	7.93
54	know	8.80	79	way	7.92
55	help	8.73	80	shake	7.89
56	when	8.72	81	him	7.89
57	cry	8.71	82	you're	7.64
58	good	8.69	83	chain	7.58
59	just	8.66	84	rock	7.58
60	let	8.57	85	one	7.55
61	now	8.56	86	for	7.53
62	was	8.44	87	if	7.51
63	hitch	8.37	88	gonna	7.51
64	hike	8.37	89	make	7.49
65	give	8.32	90	wanna	7.49
66	heart	8.31	91	boogie	7.46
67	cant	8.30	92	time	7.33
68	bird	8.29	93	more	7.33
69	say	8.28	94	been	7.27
70	so	8.27	95	want	7.26
71	see	8.25	96	walk	7.26
72	back	8.23	97	bom	7.23
73	this	8.13	98	away	7.21
74	ill	8.07	99	take	7.14
75	right	8.03	100	wimoweh	7.13

Table A.3: Top 100 Positive Average Correlation Words (1-50)

	Word	Average Coefficient		Word	Average Coefficient
1	s***	0.853	26	cheri	0.553
2	trippin	0.771	27	natures	0.550
3	miscellaneous	0.757	28	overated	0.546
4	shawty	0.745	29	math	0.541
5	sip	0.728	30	pew	0.537
6	breathe	0.711	31	kelly	0.536
7	stress	0.706	32	leash	0.535
8	feds	0.703	33	x2so	0.529
9	okay	0.676	34	niros	0.527
10	buggin	0.651	35	karaoke	0.518
11	hangup	0.649	36	bustera**	0.507
12	apart	0.640	37	n****z	0.506
13	flipped	0.627	38	honkey	0.502
14	feeno	0.625	39	homicide	0.497
15	scary	0.616	40	lock	0.492
16	pager	0.605	41	drug	0.489
17	focus	0.599	42	gut	0.488
18	chuckers	0.594	43	bra	0.487
19	misswhat	0.590	44	fires	0.486
20	izzo	0.588	45	gm	0.484
21	truck	0.584	46	messed	0.483
22	slurp	0.564	47	suck	0.480
23	wind	0.561	48	locked	0.478
24	shut	0.560	49	forgetful	0.475
25	paging	0.556	50	bikes	0.464

Table A.4: Top 100 Positive Average Correlation Words (51-100)

	Word	Average Coefficient		Word	Average Coefficient
51	click	0.464	76	amok	0.428
52	soccer	0.463	77	birdman	0.428
53	continue	0.462	78	dogg	0.427
54	brrr	0.458	79	shine	0.425
55	s***ty	0.458	80	latte	0.425
56	mighty	0.457	81	boyz	0.424
57	nestle	0.457	82	focused	0.424
58	board	0.457	83	sucks	0.422
59	atlanta	0.454	84	smileaint	0.420
60	streetsfor	0.454	85	damaged	0.418
61	backed	0.452	86	posted	0.416
62	grandpa	0.451	87	outta	0.416
63	prom	0.449	88	huhchorus	0.414
64	bro	0.447	89	mahout	0.414
65	discouraged	0.444	90	dots	0.412
66	country	0.440	91	organized	0.412
67	ohh	0.439	92	constanent	0.412
68	pasternak	0.439	93	fashioned	0.410
69	dude	0.438	94	shall	0.408
70	d	0.435	95	dawg	0.407
71	micchecka	0.435	96	eff	0.405
72	movingto	0.433	97	yous	0.405
73	mirar	0.432	98	frightened	0.404
74	option	0.429	99	enjoys	0.403
75	raisin	0.428	100	tryin	0.401

Table A.5: Top 100 Negative Average Correlation Words (1-50)

	Word	Average Coefficient		Word	Average Coefficient
1	menage	-0.622	26	journal	-0.429
2	opposite	-0.577	27	lola	-0.429
3	slums	-0.538	28	sideshow	-0.428
4	jewels	-0.530	29	bach	-0.426
5	sick	-0.524	30	gunit	-0.424
6	3	-0.510	31	admission	-0.420
7	share	-0.499	32	produce	-0.404
8	chalk	-0.488	33	36	-0.402
9	thorough	-0.487	34	style	-0.402
10	traveling	-0.476	35	vvs	-0.399
11	propaganda	-0.472	36	simon	-0.399
12	edition	-0.465	37	begging	-0.398
13	teachers	-0.462	38	es	-0.398
14	slanging	-0.454	39	heartache	-0.394
15	warmed	-0.451	40	post	-0.391
16	properly	-0.445	41	30	-0.390
17	skill	-0.442	42	NHB[1]	-0.388
18	pillow	-0.442	43	blue	-0.386
19	gone	-0.439	44	added	-0.385
20	dimmed	-0.438	45	techniques	-0.383
21	chairman	-0.436	46	mindcouldnt	-0.383
22	trap	-0.433	47	smothered	-0.382
23	pimps	-0.432	48	night	-0.382
24	sharing	-0.430	49	sacred	-0.381
25	cuando	-0.430	50	compares	-0.381

Table A.6: Top 100 Negative Average Correlation Words (51-100)

	Word	Average Coefficient		Word	Average Coefficient
51	needs	-0.381	76	fury	-0.345
52	per	-0.380	77	quo	-0.342
53	drastic	-0.376	78	relived	-0.341
54	christmas	-0.375	79	grandpad	-0.340
55	babyso	-0.372	80	invent	-0.331
56	compulsively	-0.370	81	richer	-0.330
57	hitll	-0.369	82	invades	-0.330
58	cotton	-0.369	83	walking	-0.330
59	away	-0.365	84	mackallenium	-0.330
60	assassinate	-0.363	85	friction	-0.326
61	planned	-0.362	86	haves	-0.326
62	industry	-0.359	87	original	-0.326
63	biography	-0.359	88	holluh	-0.324
64	terrific	-0.359	89	gillianll	-0.324
65	recite	-0.359	90	swang	-0.322
66	warmer	-0.357	91	plow	-0.322
67	sizzla	-0.354	92	rode	-0.321
68	35	-0.353	93	maggie	-0.321
69	inevitable	-0.351	94	3braid	-0.319
70	viene	-0.351	95	casantheova	-0.319
71	accepts	-0.351	96	lured	-0.318
72	complaining	-0.347	97	frankly	-0.318
73	troop	-0.347	98	stoney	-0.318
74	digo	-0.345	99	coverin	-0.318
75	pipes	-0.345	100	nibbling	-0.317

Appendix B: Code

Commonality Plot

```
def commonality_plot2(keywords, lyric_word_count, year_s, year_e, \
    point_marker=1, spline_marker=0, book_marker=1, coeff_marker=0, \
    picture_marker=0,rename = None):
    '''
    Makes a commonality plot for word in lyrics from the range of given years
    Try not to use more than ten words in a list of keywords
    The lyric counter should be a dictionary from lyric_word_counter
    To show points, set point marker == 1
    To show spline or polyfit, set spline_marker = "spline" or "polyfit"
    To show information for NGrams, set book_marker == 1
    To show coefficients of correlation, set coeff_marker = 1
    To save the pictures as PNG files, set picture_marker = 1
    If want to rename objects in table, make rename = list with renamed
    '''

    picture_name = '_'.join(keywords)
    mydir = 'C:\\Users\\ROB III\\Documents\\Research\\pictures\\'
    os.chdir(mydir)
```

```
if rename != None:
    names = rename
else:
    names = keywords

X = [saturdays(year) for year in range(year_s,year_e + 1)]
X = [item for sublist in X for item in sublist]
X_datetime = [datetime.strptime( x, "%Y-%m-%d" ) for x in X]

Y = []
for word in keywords:
    Y.append([1.0*(lyric_word_count[x][unicode(word)])/\
        (sum(lyric_word_count[x].values())) for x in X])

if coeff_marker == 1:
    framelist = []
    for x in X:
        framelist.append([lyric_word_count[x][unicode(word)] for word \
            in keywords])
    frame = pd.DataFrame(numpy.array(framelist),columns = names)
    print "Pearson Coefficient"
    print frame.corr()
    print
    print "Spearman Coefficient"
    print frame.corr(method = 'spearman')
    print

d = mdates.date2num(X_datetime)
```

```
colors = itertools.cycle(["#0000ff","#ff0000","#008000","#00ffff",\
"#ff00ff","#800000","#ffd700","#dc143c","#ff8c00"])

xx = np.arange(min(d),max(d) + 1)

Year_datetime_songs = [datetime.strptime( str(year), "%Y" ) for year in \
    range(year_s,year_e+1)]
if book_marker == 1:
    if year_e >2008:
        year_e = 2008
    Year_datetime = [datetime.strptime( str(year), "%Y" ) for year in \
        range(year_s,year_e+1)]
    first_str = "%2C".join(keywords)
    second_str = "".join(["t4%3B%2C" + keywords[0] + \
        "%3B%2Cc0%3B%2Cs0%3B%3B" + keywords[0].capitalize() + \
        "%3B%2Cc0%3B%3B" + keywords[0] + "%3B%2Cc0" ] + ["%3B.t4%3B%2C" + \
        word + "%3B%2Cc0%3B%2Cs0%3B%3B" + word.capitalize() + \
        "%3B%2Cc0%3B%3B" +word + "%3B%2Cc0" for word in keywords[1:]])
    new_url = "https://books.google.com/ngrams/graph?content=" + first_str \
        + "&case_insensitive=on" + "&year_start=" + unicode(year_s) + \
        "&year_end=" + unicode(year_e) + \
        "&corpus=16&smoothing=0&share=&direct_url=" + second_str
        # corpus16 is English Fiction
    soup2 = BeautifulSoup(urllib2.urlopen(new_url).read(), "lxml")

    book_data = soup2(type="text/javascript")[4].string.strip()\
        .split("var data = ")[1].split(";")[0]
    book_dict_pre = ast.literal_eval(book_data)
    book_dict = []
```

```
        all_text = ["{0} (All)".format(word) for word in keywords]
        for d in book_dict_pre:
            if d['ngram'] in all_text:
                book_dict.append(d)

    #Point Plot
    fig1 = plt.figure(figsize=(20,10))
    plt.axes([0.12, 0.10, 0.80, 0.75])
    for n in range(0,len(keywords)):
        c = next(colors)
        if point_marker == 1:
            plt.plot(X_datetime,Y[n],'o', color=c, ms=7, \
                label = names[n])
        if spline_marker == "spline":
            spline = inter.UnivariateSpline(d, Y[n])
            plt.plot(xx, spline(xx),'-', c, linewidth=5, label = names[n])
        if spline_marker == "polyfit":
            list_betas = np.polyfit(d,Y[n],3)
            line = np.polyval(list_betas, d)
            plt.plot(X_datetime,line,'-',c, linewidth=5, label = names[n])

    formatter = DateFormatter('%Y')
    plt.gcf().axes[0].xaxis.set_major_formatter(formatter)
    plt.xlabel("Time", fontsize=30); plt.ylabel("Relative Frequency", \
        fontsize=30)
    plt.title("Weekly Frequency in Music Lyrics", fontsize=30, y=1.08)
    plt.tick_params(labelsize=20)
    plt.legend(loc=1, fontsize=25)
    if picture_marker == 1:
```

```
        fig1.savefig(picture_name + '_freq.png',dpi=fig1.dpi)
    print;

    #Binned Plot
    YY = []
    for n in range(0,len(keywords)):
        new_y = zip(*[iter(Y[n])]*52)
        new_y = [np.mean(new_y[m]) for m in range(0,len(new_y))]
        YY.append(new_y)

    fig2 = plt.figure(figsize=(20,10))
    colors = itertools.cycle(["#0000ff","#ff0000","#008000","#00ffff",\
    "#ff00ff","#800000","#ffd700","#dc143c","#ff8c00"])
    plt.axes([0.12, 0.10, 0.80, 0.75])
    for n in range(0,len(keywords)):
        c = next(colors)
        subtracted = np.array(YY[n])
        plt.plot(range(year_s,year_e+1),subtracted,'-',color=c, linewidth=5, \
            label = names[n])
    plt.xlabel("Time", fontsize=30); plt.ylabel("Relative Frequency", \
        fontsize=30)
    plt.xlim(year_s,year_e)
    plt.title("Yearly Frequency in Music Lyrics", fontsize=30, y=1.08)
    plt.tick_params(labelsize=20)
    plt.legend(loc=1, fontsize=25)
    if picture_marker == 1:
        fig2.savefig(picture_name + '_freq_binned.png',dpi=fig2.dpi)
    print;
```

APPENDIX B. CODE

```
#Book Plot
if book_marker == 1:

    fig3 = plt.figure(figsize=(20,10))
    colors = itertools.cycle(["#0000ff","#ff0000","#008000","#00ffff",\
    "#ff00ff","#800000","#ffd700","#dc143c","#ff8c00"])

    if coeff_marker == 1:
        framelist = []
        for n in range(0,year_e-year_s):
            framelist.append([YY[m][n] for m in range(0,len(keywords))] +
                             [book_dict[m]['timeseries'][n] for m in \
                             range(0,len(keywords))])
        frame = pd.DataFrame(numpy.array(framelist),columns = names + \
                [word + '_book' for word in names])
        print "Pearson Coefficient for Yearly"
        print frame.corr()
        print
        print "Spearman Coefficient for Yearly"
        print frame.corr(method = 'spearman')
        print
    plt.axes([0.12, 0.10, 0.80, 0.75])
    for n in range(0,len(keywords)):
        c = next(colors)
        subtracted = np.array(book_dict[n]['timeseries'])
        plt.plot(range(year_s,year_e+1),subtracted,'-',color=c,linewidth=5,\
            label = names[n])
    plt.xlabel("Time", fontsize=30); plt.ylabel("Relative Frequency", \
        fontsize=30)
```

```
        plt.xlim(year_s,year_e)
        plt.title("Yearly Frequency in Books", fontsize=30, y=1.08)
        plt.tick_params(labelsize=20)
        plt.legend(loc=1, fontsize=25)
        if picture_marker == 1:
            fig3.savefig(picture_name + '_book.png',dpi=fig3.dpi)
        print;
```
