Spring 2018

Automatic Analysis of Musical Lyrics

Joanna Gormley
Merrimack College, gormleyjo@merrimack.edu

Recommended Citation

Automatic Analysis of Musical Lyrics

Merrimack College Honors Program - Senior Capstone
Fall 2017

Joanna Gormley
Dr. Zachary Kissel

Table of Contents

Abstract

Is music getting less sophisticated over time? That is the question which this study aims to answer, with the goal of improving upon previous analysis done on the topic. The blog posts which inspired this project lacked accuracy and dimensionality. Realizing that a larger data set of songs would make a significant difference in the precision of our analysis, we set out to design a piece of software constructed with the capability to analyze several thousand songs. Mimicking previous works which analyzed sophistication of music, the software focuses on the lyrics of songs. Three metrics were used in order to measure the sophistication of a specific song. The first being the Flesch-Kincaid Grade Level Index. This was the baseline of our study, as it was the only metric used in the previous analysis done on the topic. In an attempt in increase dimensionality, we also calculated the Coleman-Liau Grade Level Index and the Shannon Entropy measurement for each song. These metrics required that our software be able to count the total number of syllables, characters, words, and sentences in each song, as well as calculate the probability distribution for each unique word and stanza. The results show that, for the most part, songs have not gotten more or less lyrically sophisticated over time. Lyrics from 1959-2016 have been fairly consistent in terms of grade level and entropy measurements. However, it is important to consider that analyzing songs based on their lyrics might simply be an inaccurate measure of sophistication. Therefore, the overall conclusion of this study is two sided. Either music has not gotten more or less lyrically sophisticated over the years or the metrics used in this study do not capture the entire picture.

Introduction

The majority of the most popular songs in the United States all have one thing in common: they have lyrics. Lyrics, the words sung in a song, are one of the most defining characteristics of pop music. Music is often created with the listener in mind; with the goal of creating a song that will be well liked by many. The key to this is successfully grabbing and maintaining the listeners' attention. The lyrics of a song can play a large role in this effort, being that "Music needs two things to maintain a listener's attention: repetition and contrast. If it repeats too much, it will get boring; if it is too changing and doesn't ever repeat, a listener isn't able to maintain focus." (Pruett, 24) It is for this reason that lyrics are often studied when attempting to analyze music.

Is popular music becoming less sophisticated over time? Although this question has been previously investigated by several bloggers, their analysis is not easily reproducible and there is quite a bit of room for improvement. Therefore, the goal of this project is to not only answer the question of trends in musical sophistication, but to improve upon previous work done on the topic. To do this, additional metrics were used and the data set was refined in order to increase accuracy. All analysis was automated by a piece of software constructed with the capability to analyze several thousand song lyrics.

Background

At the basis of this study are three main aspects: the question being considered, the data, and the metrics. Previous work done on the topic of lyrical analysis played a large role in the shaping of the base on which this project lies. Andrew Powell-Morse, the publisher of a statistics blog which performs in-depth analysis about sports and entertainment topics, investigated Lyric Intelligence in Popular Music. This blog post influenced the question, goal, and metrics at the center of this study. The data set, on the other hand, was modeled on the work of Kaylin Walker, a statistics grad student and "curious data explorer." Each of these blog posts, while at the center of this project's background and base, left a great deal of room for improvement and future work.

Andrew Powell-Morse asked three questions in his blog post: Which genre is the most sophisticated?, Which artists are the dumbest?, and Can any hit songs be comfortably read by a first grader? To answer these questions, he formed a data set consisting of songs that spent at least 3 weeks on the Billboard charts for Pop, Country, Rock, and R&B/Hip-Hop from 2005 to 2014. About 225 songs were pulled from each of these four charts, resulting in a data set consisting of just over 2000 songs. Morse used only one metric to analyze the sophistication of a song; the Flesch-Kincaid Grade Level index. Once the grade level for each song was calculated, Morse compared songs from different genres, songs by male and female artists, the top specific artists, the lowest scoring songs, and the highest scoring songs. His study produced results such as the ones shown in Figure A, which indicates that the average grade level for songs released between 2005 and 2014 did not varied greatly. According to his study, the music of most years has an average reading level of around third grade. Overall, Morse concluded that music has not gotten more or less intelligent over the years.

How Has Lyric Intelligence Changed Over the Last 10 Years?

average grade level

We've calculated the average grade level for the top ranking songs each year according to Billboard.
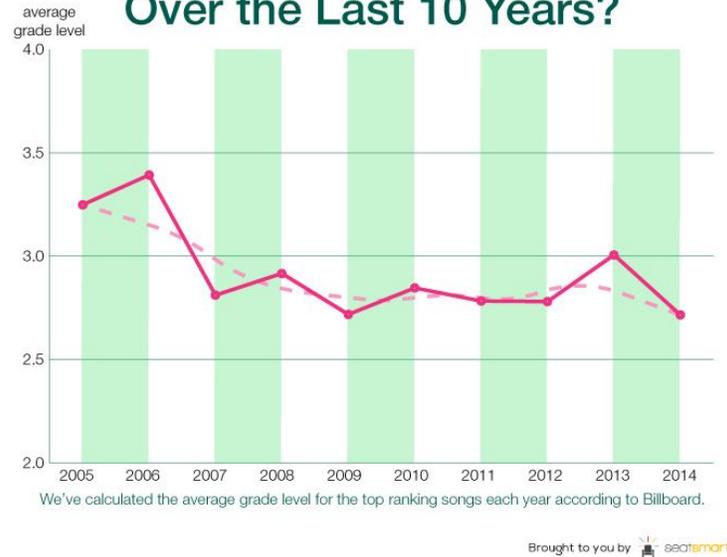
Brought to you by seatsmart

Figure A:

Unlike Morse's article, Kaylin Walker's blog post, titled *50 Years of Pop Music,* does not explicitly explore the question "Is music getting less sophisticated?" Walker surveys data points such as most frequent words, percent of one-hit wonders, career spans of top 20 most-charted artists, and the number of total and unique words per song. For example, as can be seen in Figure B, a graph from her study, it appears that word count per song is steadily increasing year by year. However, Walker does not investigate trends or hypothesize what her findings might indicate about lyrical sophistication. The most valuable aspect of Walker's post to this study is the data set which she created. Walker collected the lyrics of the songs from the *Billboard* Year-End Hot 100 Songs for each year from 1965-2015, resulting in a data set of 4,913 songs. Due to covering a much larger time span and consisting of many more songs, it was concluded that Walker's data set is superior to Morse's. Thus, Walker's data set was chosen to be the superior data set for this research.
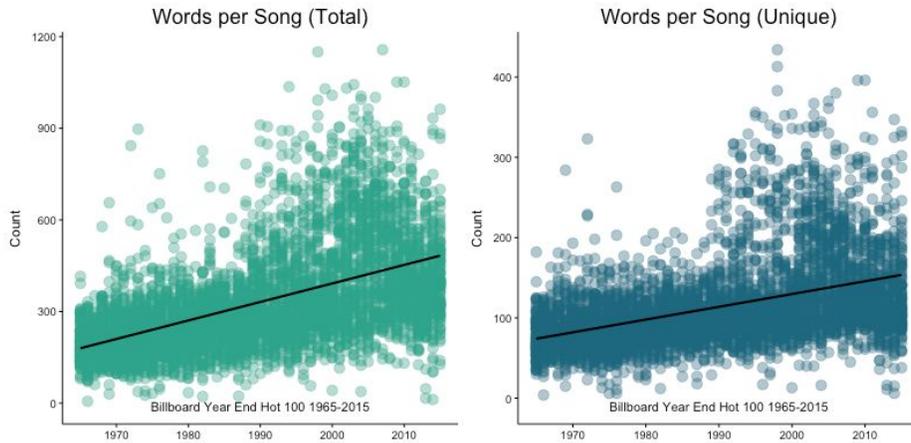
Words per Song (Total)   **Words per Song (Unique)**

Figure B:

The work of both Andrew Powell-Morse and Kaylin Walker heavily influenced main aspects of this study. The goal of the study, as well as the question which we try to answer, and the metrics used to analyze the lyrics were inspired by Morse. The data set which we shaped was inspired by Walker. However, both blog posts left much room for improvement. Morse could have had more accurate results had he created a larger data set, investigated a larger time span, and used more metrics to analyze the lyrics. Walker, although she had a superior data set, could have improved her study by focusing more on the analysis of what her results indicated. This study aimed to improve upon these previous works done on the topic.

Data

The analysis done in this project relies on several data sets. The first, Data Set 1, being a collection of each song's artist, album, year released, rank on the *Billboard* charts, and lyrics. This data set, like Kaylin Walker's, consisted of songs from the *Billboard* Year End Hot 100 Singles lists. However, in an effort to improve upon Walker's study, we widened her time span - investigating songs from 1959 to 2016. The song's lyrics for the data set were scraped from metrolyrics.com, songlyrics.com, and lyricsmode.com. However, once initial analysis was done on this corpus of lyrics, it was realized that there were flaws in many songs' lyrics. Oftentimes, the chorus of a song is repeated several times. Ideally, all of the lyrics which were scraped would have the chorus written out as many times as it was repeated while sung. However, many song's lyrics used labeling such as "[Chorus x 2]" to indicate that the chorus repeated twice. This issue led to inaccurate calculations of the count of syllables, words, sentences, and several other aspects that the metrics used to measure sophistication are dependent on. In an attempt to eliminate this flaw for the sake of accuracy, modifications were made to the corpus, forming Data Set 2. Lyrics were no longer scraped from lyricsmode.com, as it was observed that this site in particular often uses chorus labeling to indicate repetition. Also, to catch any flaws that slipped through the cracks, any songs from this new corpus that's lyrics contained "Chorus" were excluded.

The third data set is a dictionary of words and the corresponding number of syllables in each word. The total number of syllables is required in order to calculate the Flesch-Kincaid Grade Level, a metric which was used in order to analyze a song's sophistication. Syllables, however, are difficult for a computer program to detect and count. This dictionary of words

originated from a list of 340,000 hyphenated words containing all UK and US permitted scrabble words, the Moby Hyphenated Word List, and words from various dictionaries and spell-check lists. However, 8667 words in the corpus of lyrics were not included in the syllable dictionary, such as slang and profanity. Simply discounting such a large number of words would lead to very inaccurate results. Therefore, the 8667 words and the number of syllables in each word were added to the syllable dictionary manually.

Andrew Powell-Morse made no mention in his blog post of either of the issues faced regarding the flawed data sets. If one assumes that Morse did not attempt to resolve the lack of repeated choruses in many lyrics or the nonexistence of a dictionary containing every word in the lyrics, it can be concluded that Morse's analysis is flawed and partially inaccurate. The rebuilding of the corpus of lyrics and the addition of 8667 words to the syllable dictionary were the first steps towards the goal to improve upon Morse's study. The second step, rather than regarding the data sets, involved adding metrics to the analysis in an attempt to increase dimensionality.

Metrics

   The initial goal of the project was to recreate Andrew Powell-Morse's analysis and then

compare the results with his. As mentioned, Morse used only one metric to analyze the

sophistication of a song, the Flesch-Kincaid Grade Level index. The measurement, designed to

indicate comprehension difficulty when reading a passage of contemporary academic English,

corresponds to a U.S. grade level. The total number of syllables, words, and sentences in a body

of text are necessary in order to calculate the Flesch-Kincaid Grade level. The formula is as

follows:

$$0.39(word \div sentences) + 11.8(syllables \div words) - 15.59$$

   In an attempt to increase dimensionality of the analysis of lyrical sophistication, two

metrics were used in addition to the Flesch-Kincaid Grade Level index; the first of which being

the Coleman-Liau Grade Level Index. Similar to the Flesch-Kincaid Grade Level index, the

measurement corresponds to a U.S. grade level. The index was designed to help the U.S. Office

of Education calibrate the readability of all textbooks for the public school system. The total

number of characters, words, and sentences in a body of text are necessary in order to calculate

the Flesch-Kincaid Grade level. The accuracy of this measurement compared to Flesch-Kincaid

is debated. However, characters can be more easily and accurately counted by a computer than

syllables can. It is the fact that Coleman-Liau depends on characters, rather than syllables, that

some prefer it over Flesch-Kincaid. The formula is as follows:

$$0.0588(letters \div words \times 100) - 0.296(sentences \div words \times 100) - 15.8$$

The third metric used in the study to analyze lyrical sophistication is Shannon Entropy. In general, entropy measures the average amount of bits required to represent information. In this project, entropy is used to measure the amount of repetition in a songs lyrics. A high entropy measure indicates little repetition, and a low entropy measure indicates frequent repetition. If a song repeats one word many times, it would not need many bits to represent the lyrics because it only needs to store each unique word once. Therefore, the more unique words a song has, the more bits are needed to represent the lyrics. If one interprets frequent repetition as an indicator that a song is not very sophisticated and little repetition as an indicator that a song is sophisticated, a song's entropy measurement can be used to capture this dimension. The probability distribution of each word, being the probability that a specific word will occur in a song, is necessary in order to calculate entropy. In order to compute this, we counted the number of occurrences of each unique word in a song. The probability that a specific word will occur is the count of occurrences of that word divided by the total number of words in the song. Once the probability distribution is calculated, one can use the following formula to calculate entropy:

$$H(X) = -\sum_{x \epsilon X} Pr[X = x] \times lg Pr[X = x]$$

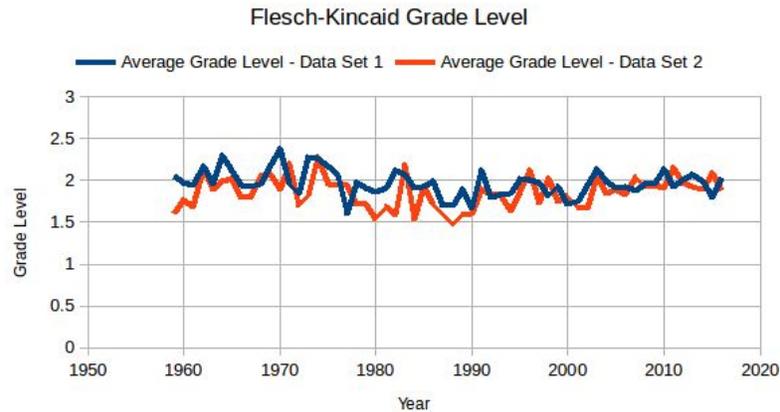# Results

Flesch-Kincaid Grade Level Index
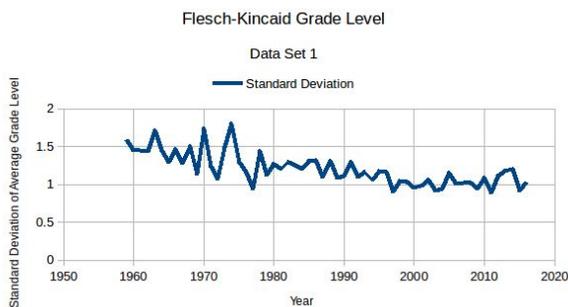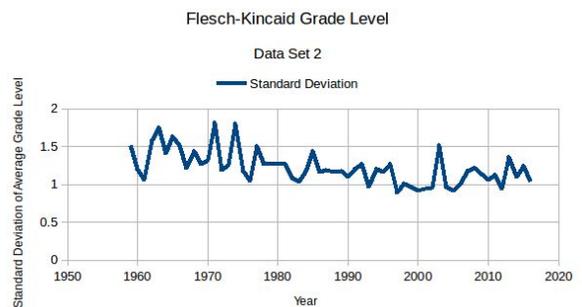

Figure 1:

Figure 2:

Figure 3:

Above are the results of the Flesch-Kincaid Grade Level Index calculation. Figure 1 depicts the average grade levels for each year from both Data Set 1 and Data Set 2. As the graph indicates, the average grade level of lyrics has been fairly consistent over the years, falling around a second grade reading level. Also, the results from the two different data sets of lyrics are very similar. It was anticipated that Data Set 2 would yield higher reading levels due to the fact that the syllable, word, and sentence could would have increased for the affected songs. However, this is not the case, indicating that our attempt to account for the lack of chorus

repetition was either not successful or it did not influence the Flesch-Kincaid Reading Level much.

Figure 2 depicts the standard deviation of grade levels for each year from Data Set 1. As the graph shows, this has been consistently low over the years. The standard deviation of grade levels is usually between 1 and 1.5. This indicates that the majority of songs from each year have similar grade levels. The same can be said for the results from Data Set 2, depicted in Figure 3.
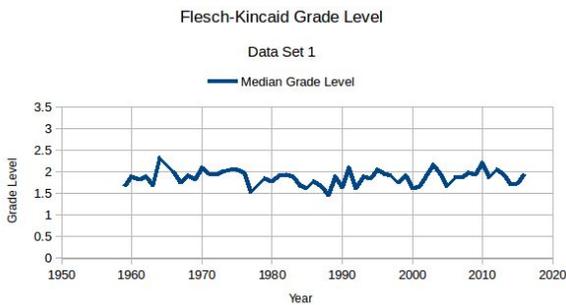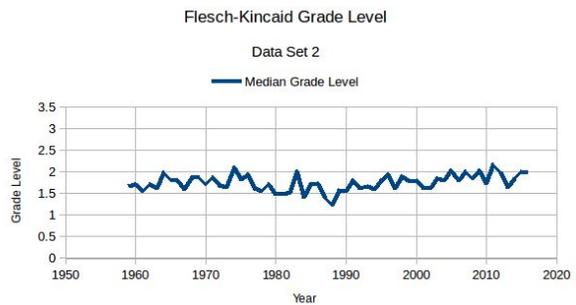
Flesch-Kincaid Grade Level

Data Set 1

Data Set 2

Figure 4:   **Figure 5:**

Above, Figure 4 and 5 show the median grade levels for each year from Data Set 1 and Data Set 2. For both datasets, the median consistently falls around a second grade reading level. These results match those of the average grade levels for each year, indicating that the average grade levels are an accurate representation of the data sets.

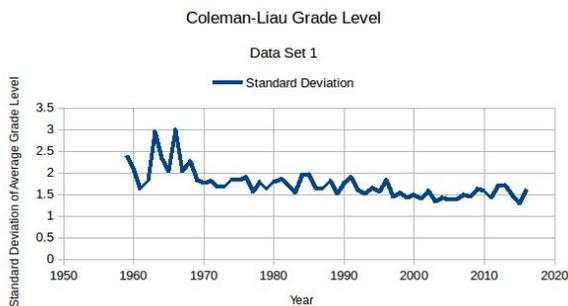Coleman-Liau Grade Level Index
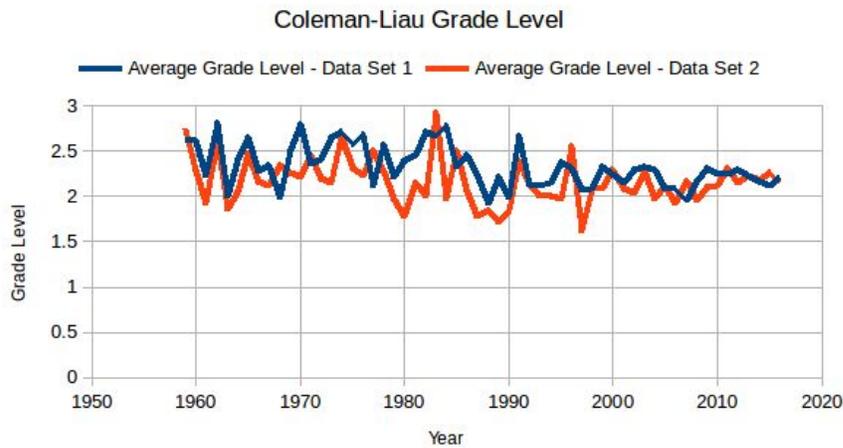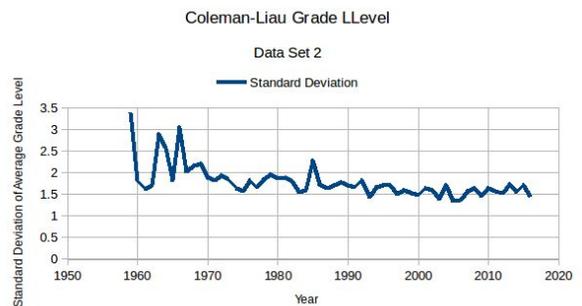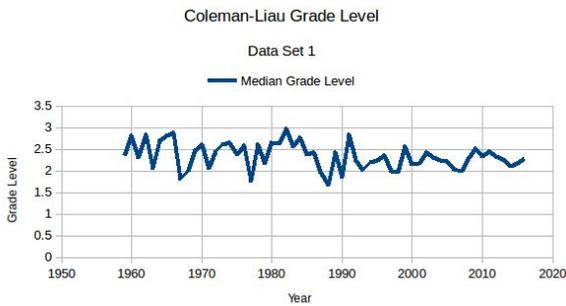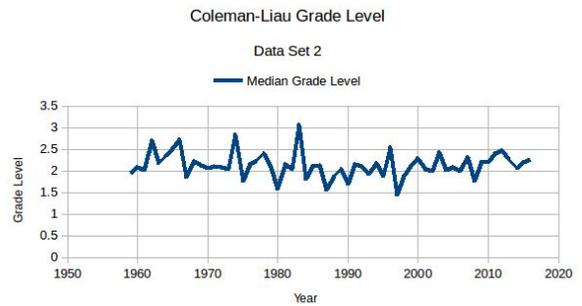

Figure 6:

Figure 7:

Figure 8:

Above are the results of the Coleman-Liau Grade Level Index calculation. Figure 6

illustrates the average grade levels for each year from both Data Set 1 and Data Set 2. As the

graph indicates, the average grade level of lyrics has been fairly consistent over the years, falling

between a second and third grade reading level. Also, like the Flesch-Kincaid Grade Level

results, the results from the two different data sets of lyrics are very similar. Again, the

conclusion can be drawn that the attempt to account for the lack of chorus repetition was either

not successful or it did not influence the Coleman-Liau Reading Level much.

Figure 7 depicts the standard deviation of grade levels for each year from Data Set 1. As the graph shows, this has been consistently low over the years. The standard deviation of grade levels is usually between 1.5 and 2. This indicates that the majority of songs from each year have similar grade levels. The same can be said for the results from Data Set 2, shown in Figure 8. However, in both data sets the standard deviation spikes between 1960 and 1970, indicating that the range of grade levels in this decade are wider than in future years.

Figure 9:

Figure 10:

Above, Figure 9 and 10 show the median grade levels for each year from Data Set 1 and Data Set 2. For both datasets, the median grade levels over the years ranges between a 1.5 and third grade reading level. Although the median grade level is not as consistent as the average grade level over the years, the two results still line up for the most part. This indicates that the average grade levels are an accurate representation of the data sets.

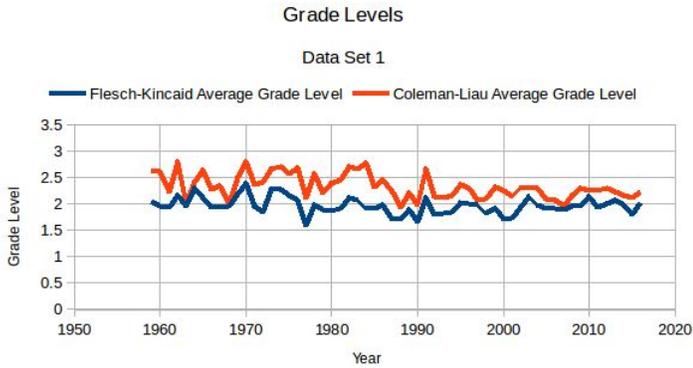**Figure 11:**   **Figure 12:**
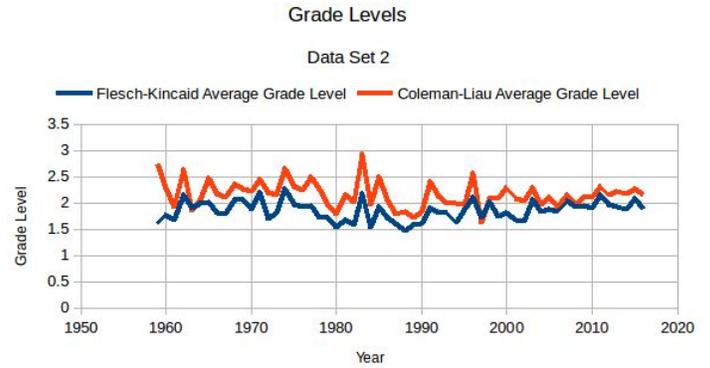
Figure 11 and 12 are graphs showing both the Flesch-Kincaid and Coleman-Liau Grade Level Index results for each year for both data sets. Neither graph shows any significant increase or decrease of either reading level over the years. When comparing the two different measurements, in both datasets, the Coleman-Liau Grade Level is consistently about half a grade level higher than the Flesch-Kincaid Grade Level. However, the results of the two measurements do not very greatly. In both datasets, the average grade level for either measurement is always between 1.5 and 3.

Figure 13:

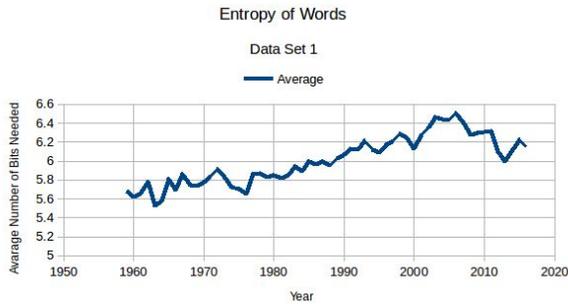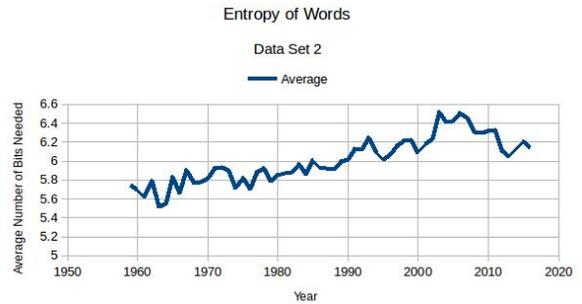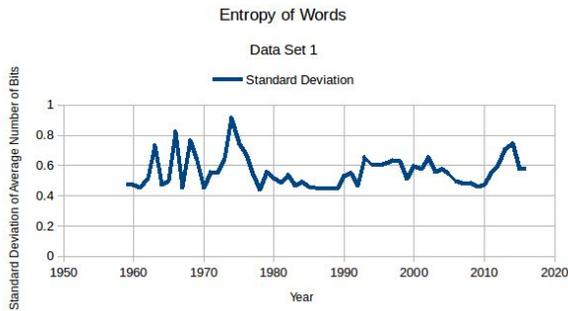

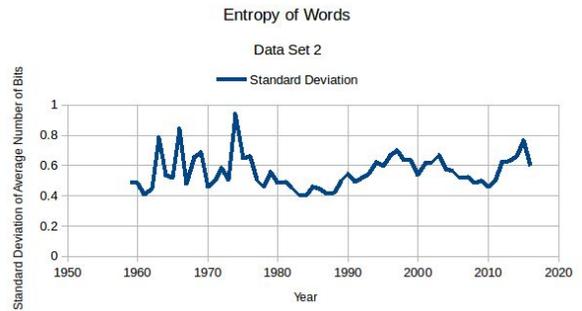**Figure 14:**

Figure 15:

Figure 16:

Above are the results of the entropy measurement of words from each year. Figure 13 shows the results from Data Set 1. It appears that the entropy measurement is steadily increasing as time goes on. The same can be observed from Figure 14, which depicts the results from Data Set 2. The fact that the entropy measurement is increasing shows that lyrics have gotten less repetitive over the years. However, it is important to note that the increase is not by much, going from around a 5.6 to a peak of about 6.4.

Figure 15 and 16 illustrate the standard deviation of entropy measurements from each year from Data Set 1 and Data Set 2, respectively. In both data sets, the standard deviation is

below one every year. This is a very low standard deviation, indicating that the majority of songs over the years have entropy measurements that are very similar.
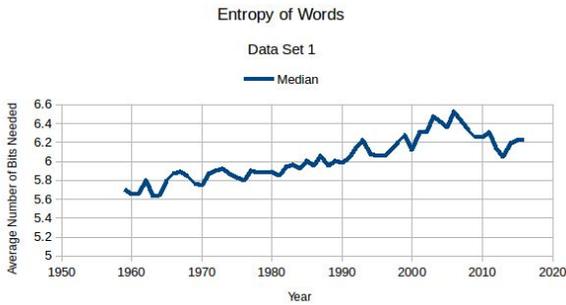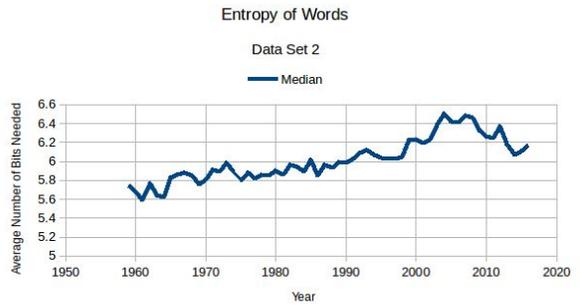

Figure 17:

Figure 18:

Above, Figure 17 and 18 depict the median entropy measurements of words from each year. The results from both data sets are very similar to the average entropy measurements. In both datasets, the average and median entropy measurements steadily increase from 1959 to 2016. This further confirms the conclusion that lyrics have gotten slightly less repetitive over the years.
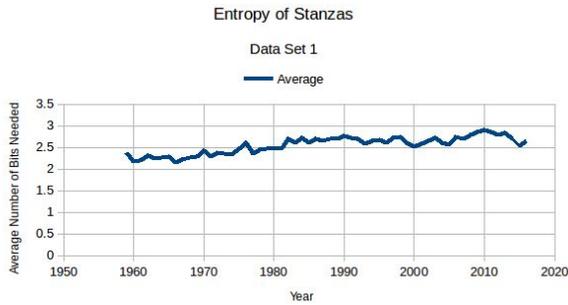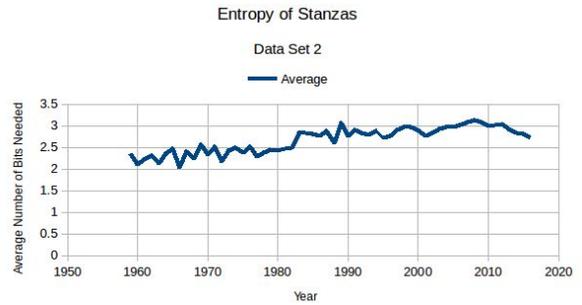
Entropy of Stanzas

Data Set 1

Average

Figure 19:

Entropy of Stanzas

Data Set 2

Average

Figure 20:

Entropy of Stanzas

Data Set 1

Standard Deviation

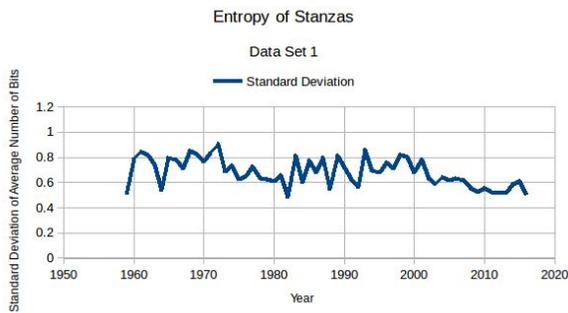**Figure 21:**

Entropy of Stanzas
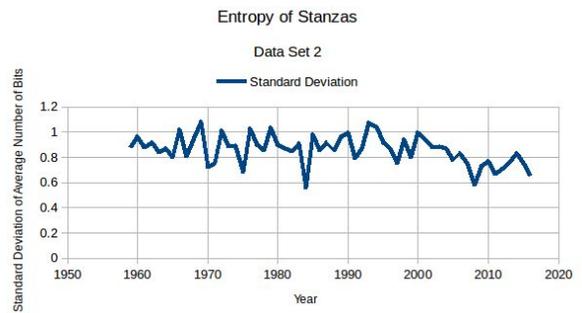
Data Set 2

Standard Deviation

Figure 22:

Above are the results of the entropy measurement of stanzas from each year. Figure 19 shows the results from Data Set 1. The analysis suggests that the entropy measurement is steadily very slightly increasing as time goes on. The same can be observed from Figure 20, which depicts the results from Data Set 2.

Figure 21 and 22 depict the standard deviation of entropy measurements from each year from Data Set 1 and Data Set 2, respectively. In Data Set 1, the standard deviation is below one every year. This is a very low standard deviation, indicating that the majority of songs from Data Set 1 have entropy measurements that are very similar. Data Set 2 depicts many standard deviations that are slightly higher than most in Data Set 1. Although they are still very low,

always lying below 1.2, this indicates that the range of entropy measurements of stanzas is slightly wider in Data Set 2.
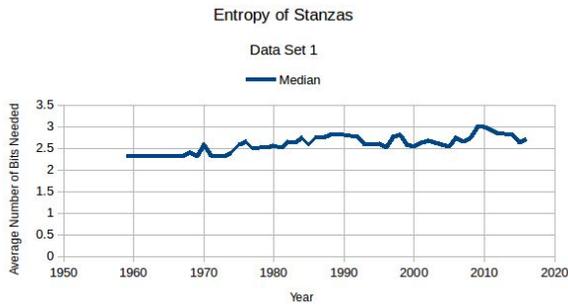

Figure 23:

Figure 24:
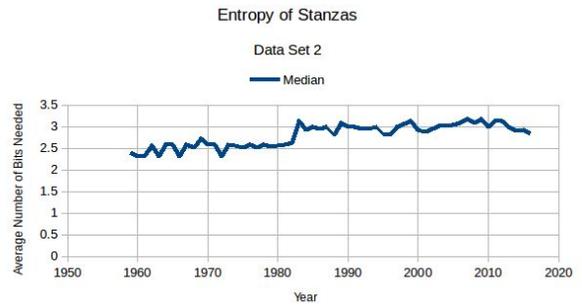
Above, Figure 23 and 24 illustrate the median entropy measurements of stanzas from each year. The results from both data sets are very similar to the average entropy measurements. In both datasets, the average and median entropy measurements steadily very slightly increase from 1959 to 2016. This further confirms the conclusion that lyrics have gotten slightly less repetitive over the years.

Conclusion

Our analysis shows that, for the most part, songs have not gotten more or less lyrically sophisticated over time. Lyrics from 1959-2016 have been fairly consistent in terms of grade level and entropy measurements over the years. Both the Flesch-Kincaid and Coleman-Liau Grade Level Index results showed that the average reading level for music over the years has been approximately a second to third grade reading level. The entropy measurements of words and stanzas only very slightly increased over the years. While looking at all of the metrics combined to judge the sophistication of a song based on the lyrics, one can conclude that it has been consistent from 1959-2016.

Although this study improved upon previous studies regarding the topic of lyrical sophistication, the conclusion that was reached is equivalent to that stated by previous studies. Andrew Powell-Morse asked questions similar to those investigated in this study. Namely, he sought to determine if specific genres or years had more sophisticated music than others. In his blogpost, he concluded that music is not getting less sophisticated. As stated, the same conclusion was reached in this study.

It is also important to consider that analyzing songs based on the number of syllables, characters, words, and sentences in the lyrics might simply be an inaccurate measure of sophistication. There may be better ways to approach song lyrics that could more accurately represent their level of sophistication. Also, the lyrics are only one dimension of a song. A song's meter, rhythm, orchestration, and a number of other musical elements are not considered. Therefore, the overall conclusion of this study is two sided. Either music has not gotten more or

less lyrically sophisticated over the years or the metrics used in this study do not capture the entire picture.

Future Work

Although this study greatly improved upon the previous work done on the topic by Andrew Powell-Morse and Kaylin Walker, there are still several aspects of the analysis that could be refined further. First, one could attempt to further account for lyrics using "[Chorus x 2]" rather than repeating the actual stanza. This would have a large impact on both grade level analysis and entropy measure analysis, due to the fact that syllable, character, word, and sentences counts would increase. Of course, the brute force approach would be to manually change the lyrics of each song that used labeling to indicate repetition. However, given the size of the corpus of lyrics used in this study, this would be extremely time consuming. If a better approach to this problem were found, a more accurate analysis of lyrical sophistication could be done. Second, one could attempt to add another dimension to the analysis. Currently, this study's measure of a song's sophistication is dependent on the lyrics alone. A song's meter, rhythm, orchestration, and a number of other musical elements are not considered. Two of the main goals of this project were to increase accuracy and dimensionality of Morse's and Walker's works. While these goals have been reached to an extent, there is still room for improvement.

References

Powell-Morse, Andrew. "Lyric Intelligence in Popular Music." *SeatSmart*, SeatSmart, 7 Dec.
    2017, www.seatsmart.com/blog/lyric-intelligence/.

Pruett, Laura Moore. *The Elements of Music: An Introduction to Listening*. iBook, 2016.

Walker, Kaylin. "Text Mining 50 Years of Popular Music." *K Walker*, K Walker, 8 May 2016,
    kaylinwalker.com/50-years-of-pop-music/.