The Role of Song Lyrics in Linguistic Change

Ria Prakash

Ben Prystawski

Abstract

What role do song lyrics play in linguistic change? We might expect that lyrics are a particularly important medium for the cultural diffusion of new trends due to music's prominence in culture and cognition. In this paper, we investigate the hypothesis that changes in words' usage frequencies occur earlier in song lyrics than in books, another common form of language use. We use time-lag cross correlation and Granger causality testing to quantify the relationship between trends in lyrics and books. Our results are somewhat inconclusive, but we find evidence that word use changes in emerging genres, as Rap and Electronic music were in the 1980s and 1990s, predict future usage changes in books. Furthermore, we investigate how musicians' gender and age influence how ahead of overall trends they are.

Keywords: music; language; lyrics; linguistic change; lexical innovation

Introduction

Music holds a special place in both culture and cognition. We listen to music while driving, shopping, and working. As a result, we might expect that new trends in language use occur earlier in song lyrics than in other media like books, television, or newspapers. There are two specific reasons why music might be especially important to the diffusion of new words. First, music is ubiquitous. People hear music while exercising, driving, grocery grocery shopping, and going about various other daily tasks. The mere repetition of popular songs might ingrain their lyrics in the mind of the listener.

Second, previous work has shown that people remember the lyrics of songs they listened to better than non-musical words (Balch et al., 1992) and that novel words are better integrated with the mental lexicon when incorporated in music (Tamminen et al., 2017). These results might lead one to expect that when people hear new words in music, they are more likely to remember them, integrate them, and ultimately use in their everyday language use. Therefore, music might play an outsized role in the dissemination and popularization of new words.

Music can also be at the forefront of broader social changes that popularize new words. For example, El Sanyoura & Xu (2020) found that the words in love songs sung by men and women became more similar between 1960 and 2009. In this case, the shift in gender norms over the late 20th century likely led to these changes in vocabulary. In addition to reflecting these social trends, music often plays an active role in social changes, such as Vietnam War protest music (James, 1989). Words used in song lyrics can clearly reflect broader social changes, but what role does music play in driving these changes and popularizing relevant vocabulary? For example, it is possible that people integrate new slang words they hear in music into their vocabulary and ultimately use them in spoken language.

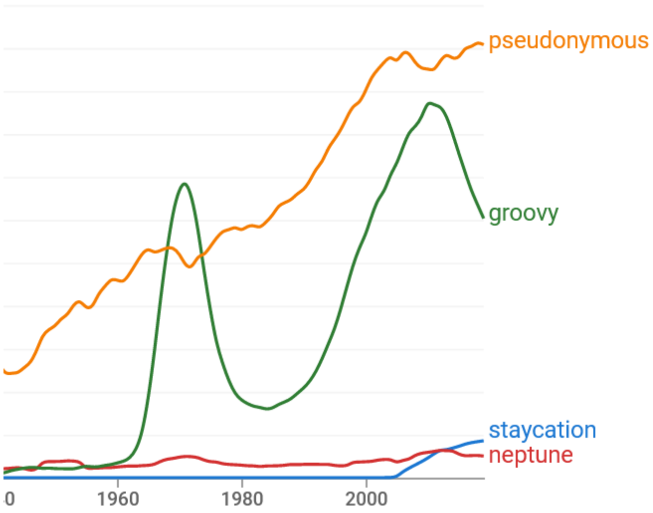

Figure 1: Google Ngram Viewer trends for selected interesting words. "pseudonymous" consistently increases , "groovy" has two distinct peaks, "staycation" has no frequency before about 2003, then emerges, and "neptune" is roughly constant over time.

Research on the origin of new words has categorized the processes by which new words enter the lexicon and shown that affixation and compounding are the two most common process in English (Ratih & Gusdian, 2018). In affixation, a new word is created by appending an affix that was previously never used before. For instance, the words "ghosting" and "googling" were formed by appending the suffix -ing to the pre-existing words "ghost" and "google." In compounding, speakers combine two or more words together to form a single word. This category includes words like "staycation" and "doomscrolling." The rhythm and rhyme structure of song lyrics might create stronger pressures to innovate linguistically. If you only have two beats in which to convey the idea of taking a break at home, inventing the word "staycation" or searching for an emerging two-syllable word conveying that meaning might be especially appealing. Likewise, if a lyricist needs a word that rhymes with -ing when writing lyrics, they might feel pressure to invent a word like "ghosting," or at least search far and wide for such a word. For these reasons, we might expect song lyrics to use either newly-invented words or newly-emerging words more often than written prose or general speech. Of course, as the relative frequencies of new words increases, the relative frequencies of other words must decrease to compensate. As a result, some words increase in their relative frequencies over time while others decrease. These changes in word frequency can reflect broader social trends. See Figure 1 for trends of a few example words in the Google Ngram viewer (Michel et al., 2011).

Past work has also shown that young women are disproportionately responsible for linguistic innovation (Tagliamonte & D'Arcy, 2009). Features of language such as using "like" as a filler word were pioneered by young women. We might expect similar patterns in song lyrics: songs sung by women and young performers might be more likely to incorporate novel vocabulary than those sung by men or older performers.

We hypothesize that changes in the corpus frequencies of words will occur earlier in song lyrics than in books, since music is likely to lead in linguistic trends. Furthermore, we expect popular genres to be further ahead of these trends than less popular genres. For instance, pop music is likely more plugged into social trends and heard by more people, so it can disseminate its conventions to general language use more easily than less popular genres like metal and new age music. Finally, we hypothesize that songs by younger musicians will be further ahead of general trends than songs by older musicians, and songs by women will be further ahead than songs by men.

Methods

Since the main goal of our project is to test for synchrony between two time series: words' relative frequencies in song lyrics and their relative frequencies in books. We want to account for one time series possibly lagging behind the other and, if so, quantify the direction and magnitude of the lag.

For robustness, we used two methods that solve this problem: time-lag cross-correlation and Granger causality testing. Before applying these methods we combined data from all songs of each year, giving us the relative frequencies of each word across all songs for each year.

Code to replicate our analyses is available at `https://github.com/benpry/COG403-songlyrics`.

Time-Lag Cross Correlation

Time-lag cross correlation (TLCC) is a method that quantifies synchrony between time series with a possible offset (Menke & Menke, 2012). It measures the Pearson correlation between elements of two series, with one series shifted by a given offest. After trying all offsets in a specified range, we take the offset that maximizes the correlation. This offset is an estimate of how much the second series lags behind the first. Negative offsets mean that the first series is ahead of the second, while positive offsets mean that the second series is ahead of the first.

In our analyses, we used offsets between -10 and 10. In doing so, we are assuming that if word frequencies in one medium lag behind trends in the other, the lag is probably not longer than 10 years. We also used song lyric frequencies as the first series and book frequencies as the second, meaning we will get negative offsets if song trends are ahead of book trends and positive offsets if book trends are ahead of song trends.

Since the time series are each specific to one word, we run this test for all English words in our lyric corpus and average the offsets.

Granger Causality Testing

Granger Causality Testing is another metric that tests for relationships between two time series where one series might lag behind another (Granger, 1969). While the name might imply that we are testing for a causal relationship between the first and second variable, Granger causality only tests whether one variable can predict the future values of another variable. This method takes an offset i and uses an F-test to test whether the value of the first series at time t contain information about the values of the second series at time $t+i$. By repeating this with a series of offsets, we can find the offset that maximizes the F-statistic. Offsets must be non-negative, so we ran two versions of the test: one where lyric frequencies are the first series and song frequencies are the second and one where the series are reversed. We then find the offset from all of these tests that maximizes the effect size of the F-test. If the best offset comes from the first round of tests, where we test whether lyric frequencies are ahead of song frequencies, then we count the offset as negative.

Bootstrap Tests

We use bootstrap tests to test whether the mean offset from TLCC or the mean Granger lag is significantly different from the null hypothesis that the mean lag is either 0 or on the opposite side of 0. After computing the mean offset and Granger lag for each word, we sample 1,000,000 times with replacement from the list of all offsets (or Granger lags) for all words. The p-value is the proportion of bootstrapped instances for which the mean effect size is in the opposite direction of the mean we computed.

Multiple Linear Regression

Multiple linear regression is a statistical technique that attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. In the case that there are n explanatory variables $x_1, x_2, ..., x_n$, the formula for multiple linear regression is given by $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon$. In this model, y is the predicted value of the response variable. The y-intercept, β_0, is the value of y when all other parameters are set to 0. The regression coefficient β_i of the i^{th} independent variable x_i represents the effect that increasing the value of that independent variable has on the predicted y value. The model error, ε, represents how much variation there is in our estimate of y. To find the best-fit line for each independent variable, multiple linear regression calculates the regression coefficients that lead to the smallest overall model error, the t-statistic of the overall model and the associated p-value which represents how likely it is that the t-statistic

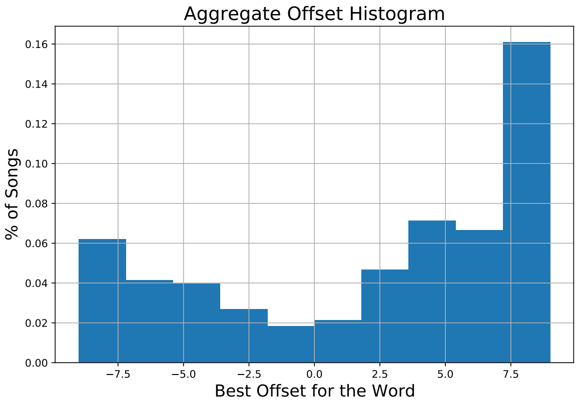

Figure 2: Histograms of the best TLCC offset for all words in the MusixMatch corpus.

would have occurred by chance if the null hypothesis of no relationship between the independent and dependent variables was true.

In our analysis, we used multiple linear regression to study how the age and gender of a song's artist impact the offset of words between songs and books.

In order to study how the gender of a musician impacts the difference in word usage trends between books and music, we conducted multiple linear regression with the best TLCC offset as the response variable, and word frequencies within songs by male and female artists as the explanatory variables. A positive coefficient in this model would imply that words said more by artists of the relevant gender tend to be further behind books in their frequency trends, while a negative coefficient would imply the reverse. Essentially, we are asking "How does the gender makeup of the artists using this word influence the lag between this word's frequency trends in songs vs. books?" Since we did not have equal amount of songs by artists of each gender, we calculated the frequencies for each gender using random samples of 3000 songs.

In order to study how age of the artist impacts the difference in word usage trends between books and music, we conducted multiple linear regression across different words with the best TLCC offset as the response variable and the word's frequency in songs by artists within each age group as the explanatory variables. We used five different age groups: under 20, 20 to 30, 30 to 40, 40 to 50 and over 50. A positive coefficient in this model would imply that frequency trends in song lyrics for words said more by that age group tend to be further behind books, while a negative offset indicates the opposite relationship. Coefficients with high magnitude indicate that the relevant age group has a large effect on the lag between words' trends in songs and books. Since we did not have equal amount of data for each age group, we calculated the frequencies for each age group using samples of 100 songs, which is the approximate number of songs for the age group with the fewest songs.

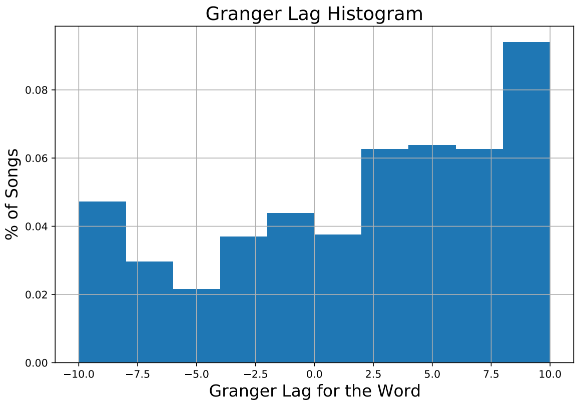

Figure 3: Histograms of the best Granger lag for all words in the MusixMatch corpus.

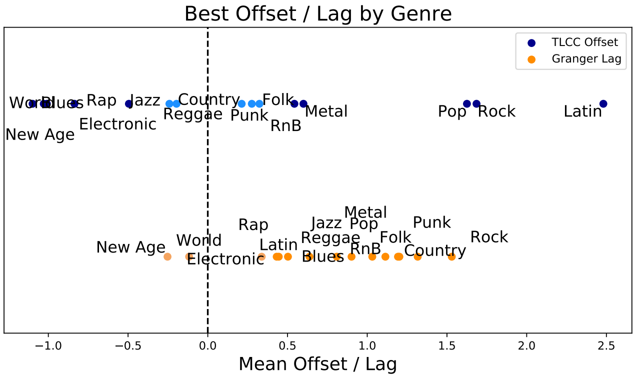

Figure 4: Mean offset (top) and Granger lag (bottom) by genre. Genres to the left of the dotted line are ahead of books, while genres to the right of the dotted line are behind books. Dark dots denote statistically significant effects at $p < 0.05$, while light dots denote statistical insignificance.

Data

Our analysis relies on three main datasets: the MusixMatch dataset, the Google Ngram dataset, and Wikipedia.

The MusixMatch dataset is a subset of the Million Song Dataset (MSD) that contains bags of words for the lyrics of 235,662 songs published before 2011 (Bertin-Mahieux et al., 2011). These bags of words contain only the top 5000 most common words occurring across all songs for copyright reasons. For our purposes, the data is quite sparse from before 1980, so we restrict our analysis to the period from 1980 to 2011. The MSD also contains genre information for most of its songs. We used the MusicBrainz API to retrieve the release date of each song in the corpus (Swartz, 2002). This dataset contains songs in several different languages, so we filtered out all songs where fewer than 80% of the words are included in the top 10,000 words of the Google Ngram dataset, which is described below.

We used the Google Ngram dataset to capture word frequency trends in books (Michel et al., 2011). This dataset

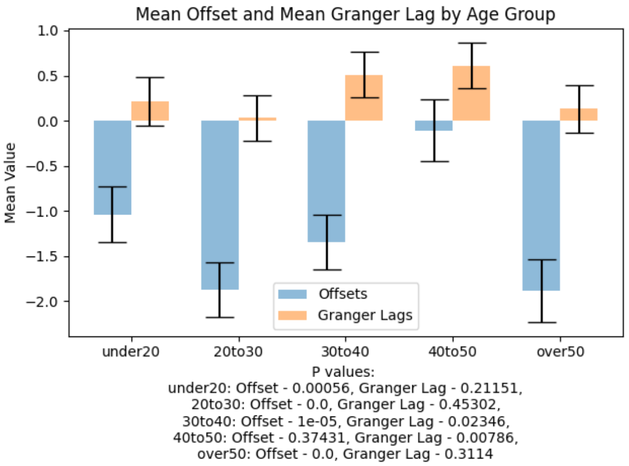

Figure 5: Mean offset and mean Granger lag by age. Error bars denote standard error of the mean.

contains relative frequencies of n-grams in each year. These frequencies are estimated from the Google Books corpus, which consists of approximately 189 billion tokens of text from books, spanning from 1800 to 2019.

We scraped Wikipedia to collect the year of birth of each artist. We extracted the year of birth from an artist's Wikipedia page by checking for the existence of a valid year in the 150 characters following the first occurrence of the word 'born' in first two paragraphs of text on their page. Through this method, we were able to collect the year of birth of 1038 artists. Many artists from the MSD did not have Wikipedia pages, and a birth year was available in the text for only a small portion from those who did.

Since there was no publicly available data on the gender of artists, we used a bigram Naive Bayes Classifier trained on National U.S. baby names from 1879 to 2015 in Social Security records to determine the gender of each artist.[1]

In order to maximise the accuracy of gender and age, we only used solo artists when collecting age and determining gender. To distinguish solo artists from bands, we selected all artists names with two or fewer words and a first name from the National U.S. baby names dataset. Using this method, we found 15057 of all artists were solo artists.

Results

In aggregate, we found that word trends in songs lag behind trends in books. The mean offset for TLCC was 2.22 ($p < 0.0001$) and the mean lag from Granger causality testing was 1.49 ($p < 0.0001$). This is contrary to our hypothesis that trends in lyrics would be ahead of trends in books. Figures 2 and 3 show histograms of the TLCC offsets and Granger lags for all words in the corpus. A large proportion of the words have offsets or lags of $+10$. This might be because earlier lyric data is somewhat noisier, as the MusixMatch dataset contains more songs from the 1990s and 2000s than the 1980s. Shifting the years 1980-1990 out of the window used to compute correlations or explain the variance of in F-tests might reduce noise. Therefore this result could just be an artifact of the different frequencies in the corpus.

Table 1: Regression on Age

	Offset	**Granger Lag**
log freq under 20	0.0572^{*}	0.0259
	(0.083)	(0.302)
log freq 20 to 30	0.0101	0.0109
	(0.749)	(0.651)
log freq 30 to 40	-0.0945^{**}	-0.0241
	(0.005)	(0.349)
log freq 40 to 50	-0.0246	-0.0123
	(0.399)	(0.579)
log freq over50	-0.0197	-0.00040
	(0.489)	(0.855)
const	1.0824^{*}	0.7108
	(0.067)	(0.114)
N	878	878
R^2	0.007	0.001

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Results by Genre

We can also divide the corpus by genre, then compute offsets and lags where the song frequencies are estimated only using songs of a particular genre. Figure 4 shows the mean offset and lag for each genre. Results differ somewhat between metrics, but there are some consistencies. Both Rap and Electronic music are significantly ahead of books using TLCC and are not significantly behind books using Granger lag. Pop and Rock are significantly behind books using both methods. This contradicts our hypothesis that popular genres of music, like Pop, would be further ahead of book trends than less-popular genres, like Electronic.

Results by Artist Age

Our analyses of mean offset and mean Granger lag by artist age are reported in Figure 5. In these reports, a negative offset or Granger lag means that word frequency trends in songs were ahead of books, while a positive offset means that word frequency trends in songs were behind books.

We found that the mean TLCC offsets for all age groups were negative and the age groups 20to30 and over50 had the highest-magnitude mean offsets. We found that the mean Granger lags for all age groups were positive and the age group 40to50 had the highest-magnitude mean Granger lag.

The results of our linear regression analysis are reported in Table 1. The first column includes the results of regressing TLCC offsets on age, while the second column includes the results of regressing the Granger lags on age. The values

[1] U.S. baby name data is available here: https://www.kaggle.com/kaggle/us-baby-names

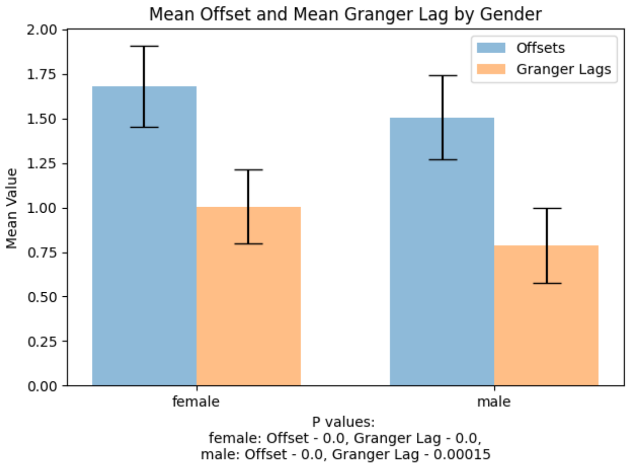

Figure 6: Mean offset and mean Granger lag by gender. Error bars denote standard error of the mean.

in the brackets are the *p*-values of the corresponding coefficient. A negative coefficient means that an increase in word frequencies within the corresponding age group leads to a decrease in the TLCC offset or Granger lag for those words. On the other hand, a positive coefficient means that an increase in word frequencies within the corresponding age group leads to an increase in the offsets for those words.

We found that the age group with the strongest influence on TLCC offset was group 30to40. In the offset model, groups under20 and 20to30 had positive coefficients, while the remaining groups had negative coefficients. We found that the age group with the largest magnitude of influence on Granger Lag was group under20. In the Granger lag model, groups under20 and 20to30 had positive coefficients, while the remaining groups had negative coefficients.

Results by Artist Gender

Our analyses of mean offset and mean Granger lag by artist gender are reported in Figure 6. We found that the mean offsets and Granger lags for both genders were positive, and female artists have a higher-magnitude mean offset than male. However, the gender difference is not statistically significant.

Table 2 contains the results of our linear regression analysis for gender. The first column includes the results of regressing TLCC offsets on gender, while the second column includes the results of regressing the Granger lags on gender. The values in the brackets are the *p*-values of the corresponding coefficient. In the first model, a negative coefficient means that an increase in word frequencies within the corresponding age gender leads to a decrease in the offsets for those words. A positive coefficient means that an increase in word frequencies within the corresponding gender leads to an increase in the offsets for those words. In the second model, a negative coefficient means that an increase in word frequencies within the corresponding gender leads to a decrease in the Granger lags for those words. A positive coefficient means that an increase in word frequencies within the corresponding gender leads to an increase in the Granger lags for those words.

Table 2: Regression on Gender

	Offset	Granger Lag
log freq male	0.7260**	-0.1797
	(0.002)	(0.330)
log freq female	-0.0918	0.1772
	(0.696)	(0.2329)
const	9.0698***	0.7192
	(0.000)	(0.313)
N	878	878
R^2	0.036	0.001

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

We found that a word's frequency in lyrics by male artists had a stronger influence on TLCC offset. In the offset model, male artists have a positive coefficient, while female artists have a negative coefficient. We found the opposite result using Granger Lag. In the Granger lag model, female artists have a positive coefficient, while male artists have a negative coefficient. Due to conflicting results with different methods and a lack of statistical significance, we cannot conclude that these results either confirm or falsify our hypothesis.

Discussion

Our results largely contradicted our hypothesis: word usage trends in song lyrics lagged behind trends in books. These results were consistent using both time-lag cross correlation and Granger causality testing.

One consistent result from both methods is that word frequency trends in Rap and Electronic music were relatively ahead compared to other genres. These genres both increased in popularity in the 1980s and 1990s, which might suggest that genres that newly enter into public consciousness exert a stronger influence on how language is used than already-popular genres. In other words, changes in overall usages are driven by new genres becoming more popular rather than trends within already-popular genres. This is contrary to our initial hypothesis that the most popular genres would drive linguistic change, but is understandable nonetheless. If Pop music represents what is already popular and Rap represents what is becoming more popular, we might expect word usage trends in Rap to predict future word usage trends in books more strongly than in Pop. Still, these results are not definitive and the lack of significance found in Granger causality testing indicates that more information is needed to draw any definitive conclusions.

Our age analysis results using offsets were slightly unexpected since all ages had negative TLCC offsets and artists over 50 had the highest-magnitude offset. All ages having negative offsets was consistent with our hypothesis that songs would be ahead of books in word frequency trends, but contrary to our finding that the mean TLCC offset in our aggregate analysis was positive. The group over50 having the highest-magnitude offset was unexpected. We had expected offsets for younger artists to be larger since previous research suggests that young adults are more likely to create and propagate emerging words. Our age analysis results using mean Granger lags were unexpected as well. All groups had positive Granger lags which was inconsistent with our hypothesis that songs would be ahead of books in word frequency trends and contrary to the results using mean offsets. However, the offset was the smallest for the 20to30 age range, meaning that these younger artists were the least behind trends in books.

Our gender analysis results using mean offsets were unexpected since mean offsets for both genders were positive and female artists were further behind books than male artists, though the difference was statistically insignificant. All offsets being positive was unexpected since this means books were ahead of songs for both genders which is inconsistent with our main hypothesis. This result is borne out both with TLCC and Granger lag. These results also contradict our hypothesis that trends in lyrics by female artists are further ahead of books are than trends in lyrics by male artists, since those trends were actually further behind .

The regression analyses regressing offset on age, offset on gender and Granger lag on age all yielded inconclusive results since the *p*-values for most explanatory variables were insignificant. Our results for regression analysis regressing offset on gender were expected since we had expected the word trends in songs to be more ahead in books for female artists than for male artists. However since the *p*-value for female artists is very high, this effect is insignificant and we cannot make any conclusions based on gender using this regression either.

The high *p*-values and low R^2 values in all regression results can probably be attributed to the small sample sizes. One major limitation of our analysis is that we were able to find age for only 0.7% of all solo artists, and thus used 100 artists per age group for regression analysis. Furthermore, it is likely that the samples we collected and used in our age and gender analyses were not representative of all the artists in the MXM dataset since TLCC offsets were negative for all age groups while the average TLCC offset using all data was positive.

In addition, the MusixMatch dataset only contains frequencies of the top 5000 words across all songs for copyright reasons. Word frequency trends in these words might not be representative of overall word frequency trends, especially in the case of emerging words. The dataset spans many decades, so a word that does not occur before, for example, the early 2000s is less likely to make it into the top 5000 words than a word that is present throughout the whole timespan of the dataset. The lack of emerging words in the dataset might explain why word trends in songs appear to lag behind trends in books. Songs might be ahead of books in their adoption of emerging words, but our dataset contains very few of these words.

Finally, our method only measures changes in the frequency of word types over time, while words' meanings and usages can change in more nuanced ways. Words often take on new senses, like the word "fire" meaning really good, which originated in Rap music. These trends might not show up in overall usage frequencies, but are nonetheless important aspects of linguistic change.

Future work in this area should seek out or compile datasets of song lyrics that contain more than just the top 5,000 words. The emergence of new words is an important part of linguistic change, so compiling a dataset which contains emerging words is a crucial future direction. This would enable a more fine-grained analysis of how new words emerge in song lyrics and books.

Since our ultimate goal is to investigate the relationship between trends in song lyrics and general language use, rather than just books, future work should also look for datasets from which word frequencies can be computed for other media, like television and newspapers.

References

Balch, W. R., Bowman, K., & Mohler, L. A. (1992). Music-dependent memory in immediate and delayed word recall. *Memory & Cognition*, *20*(1), 21–28.

Bertin-Mahieux, T., Ellis, D. P., Whitman, B., & Lamere, P. (2011). The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011).*

El Sanyoura, L., & Xu, Y. (2020). Gender convergence in the expressions of love: A computational analysis of lyrics. In *Proceedings of the 42nd annual meeting of the cognitive science society.*

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, 424–438.

James, D. (1989). The vietnam war and american music. *Social text*(23), 122–143.

Menke, W., & Menke, J. (2012). 9 - detecting correlations among data. In W. Menke & J. Menke (Eds.), *Environmental data analysis with matlab* (p. 167-201). Boston: Elsevier. doi: https://doi.org/10.1016/B978-0-12-391886-4.00009-X

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., ... others (2011). Quantitative analysis of culture using millions of digitized books. *science*, *331*(6014), 176–182.

Ratih, E., & Gusdian, R. I. (2018). Word formation processes in english new words of oxford english dictionary (oed) online. *Celtic: A Journal of Culture, English Language Teaching, Literature and Linguistics*, *5*(2), 24–35.

Swartz, A. (2002). Musicbrainz: A semantic web service. *IEEE Intelligent Systems*, *17*(1), 76–77.

Tagliamonte, S. A., & D'Arcy, A. (2009). Peaks beyond phonology: Adolescence, incrementation, and language change. *Language*, 58–108.

Tamminen, J., Rastle, K., Darby, J., Lucas, R., & Williamson, V. J. (2017). The impact of music on learning and consolidation of novel words. *Memory*, *25*(1), 107-121. Retrieved from `https://doi.org/10.1080/09658211.2015.1130843` (PMID: 26712067) doi: 10.1080/09658211.2015.1130843