Towards Uncovering Language Usage Patterns in Song Lyrics:

A Case Study of the Relationship Between Function-Word

Frequency in Lyrics and Songwriter Age

Olivia Zhang (oliviazhang@college.harvard.edu)

Professor Fiery Cushman, TF Yelina Chen

1 Introduction

Is there a relationship between a musical artist's underlying language style and their age? This study seeks to provide potential insight into this question, with a basis in the work of social psychologist James Pennebaker on functional words analysis.

1.1 Background

Music as a Method of Personal Expression:

For hundreds of years, music has been an essential and popular method for artists to communicated stories with their audiences. Contemporary lyrical music is particularly evocative due to the direct use of language, in addition to instrumental components. The widespread genre of popular music has evolved and spawned many subgenres over the past few decades. The subset of music by singer-songwriters – artists who write their own song lyrics – is of interest in this preliminary study of language usage patterns.

Relationship Between Language Usage and Age:

The work of social psychologist James Pennebaker has been pivotal in the analysis of an individual's use of function words, i.e. pronouns, articles, prepositions, conjunctions, and

auxiliary verbs, in writing and speaking in relation to identity factors.[1] In two 2003 projects, Pennebaker and Stone discovered significant relationships between a person's age and their usage of certain linguistic style markers, such as personal pronouns and prepositions.[2] Higher age was correlated with more positive and fewer negative affect words, lower usage of first person pronouns, and more prepositions and nouns, among other findings. Analysis was conducted using the LIWC software developed by Pennebaker's team, which analyzes a text across word-frequency-driven language dimensions. Pennebaker offers possible explanations for the demonstrated relationships between age and language usage pattern, such as: increasing age leads to increasing complexity of thought, which may be reflected in language via an increase in cognitive word and noun usage; as we get older, we gain more experience and perspective and thus tend to self-reference less when reflecting than we did while younger, which corresponds with a decreased usage of first person pronouns, particularly "I". As with many correlation studies of real-world variables, it is difficult to establish causation, but even the discovery of statistically significant relationships can provide interesting and impactful insight of the individual in society.

Motivation:

Pennebaker's findings on function words usage patterns have been consistent across many types of expressive text, including lab-directed expressive writing, internet blog posts, fiction literature, and poetry. However, text in the form of song lyrics, which bears certain similarities to poetry, has not been analyzed for function word frequency patterns. This study aims to fill this gap and examine the relationship between age and language usage patterns of singer-songwriters.

1.2 Objective

The objective of this project is to study the relationship between age and linguistic style markers, such as personal pronoun, preposition, and noun use. This goal is the first step in determining whether Pennebaker's findings hold for the text form of song lyrics, and whether different significant patterns emerge.

1.3 Design

Logical Application to Song Lyrics:

Pennebaker previously found significant correlative results in studies of poets and their language usage in their published poetry.[1,2] Poetry is a creative discipline and consists of language crafted for artistic form and purpose, which is a key difference from other forms of natural language text, such as books, speeches, and personal journaling. Song writing bears many similarities to poetry as another creative form of personal expression and verbal communication. If word frequency analysis can be conducted on poetry for insights about language usage and author characteristics, one can imagine the possibility of applying such techniques to analyze lyrical data.

Taylor Swift as a Case Study:

Due to feasibility limitations on project scope, one artist was chosen as a case study for the function-word frequency analytical method. This is presented as a base point from which to perform a follow up study with a much larger dataset of songs from multiple artists. Taylor Swift was selected due to her relatively long 15-year musical career that began in her adolescence and her well-known self-penmanship of most of her released songs. Implications follow to apply the proposed methodology to a larger set of singer-songwriters with long-lasting careers for whom age-language correlation analysis is applicable.

Hypothesis:

Based on Pennebaker's findings on the relationships between age and language usage, particularly personal pronouns, prepositions, and nouns, a three-part hypothesis was proposed for this study of song lyrics: 1) The age of the artist at the time of song release is negatively correlated with first person pronoun frequency, 2) is positively correlated with preposition frequency, and 3) is positively correlated with noun frequency.

The three-part hypothesis can be tested by finding the frequencies of first person pronouns, prepositions, and nouns per song, and then computing the linear correlation coefficient between each variable and the age of the song artist at the time of song release. The p-value significance level is set as $p < 0.05$, such that the correlation between two variables is statistically significant if and only if the p-value is less than 0.05.

Ideally, the artist age (i.e. the year) at which a song was written is known and used in the analysis; however, there is no standard or requirement in the music industry to report this, and so the assumption is that in most cases, the year at which a song is release is reasonably close to the time of writing and can be analyzed instead.

Alternative Hypothesis:

An alternative hypothesis is the null hypothesis that there is no correlation between the variables. This would be supported by no statistically significant correlation coefficient values. If the null is supported, it could be the case that there truly is no correlation, or it could be the case that confounding variables had an impact and the study design could be improved in a future iteration to account for such.

2 Methods

2.1 Materials

Github Repository:

All data, code, and correlation analysis files for this project have been annotated and can be found in this public GitHub repository: https://github.com/Livy08/song_pronoun_analysis for reference and updates on future extension projects.

Corpus of Lyric Text:

Song lyrics were obtained from the AZLyrics website (https://www.azlyrics.com/) and manually cross checked with lyrics on Genius (https://genius.com/). Writing credits were obtained from Genius.

The initial dataset was chosen to include the full lyrics of songs with Swift as the sole writer and songs with exactly one co-writer in addition to Swift, with one .txt file for each song. It is very common in the music industry, even for prolific expressive songwriters such as Swift, to have co-writers. To balance this aspect of the available data with the goals of this project, an assumption was made that songs with exactly one co-writer in addition to Swift herself are still relatively representative of Swift's personal language expression choices (whereas songs with three or more co-writers are more likely to stray away from Swift's natural style of expression).

Songs from all of Taylor Swift's full-length albums and demos were reviewed; if a deluxe version was produced, then it was reviewed in place of the regular length album.

Table 1 shows the results of an extension of the initial study, where two additional datasets were analyzed: the Exclusive-Full Lyrics dataset, which exclusively includes songs with Swift as the sole writer, and the Extended-Cut Lyrics dataset, which includes the same

songs as the Extended-Full Lyrics dataset but only certain sections of lyrics. A full discussion of these two datasets is not within the bounds of this report, though the code and analysis files for each are available in the GitHub repository.

2.2 Python-based Word Frequency Calculation

Software and System:

A Python-based word frequency calculator for lyric-oriented processing was created for this project. Code was developed and run in the browser-based Jupyter Notebook format of the open source Project Jupyter software. A MacBook with operating system Catalina was used. Word frequency calculation scripts utilized built in features of the Python NLTK package, including a word tokenizer and part-of-speech-tagger. Word frequency per song is defined as the ratio of the number of instances of the given word to the total number of words in the text. For details, see the .ipynb Jupyter Notebook files in the GitHub repository for the full code with annotations.

2.3 Analysis Procedure

Word Frequency Variable:

To test the proposed hypothesis, word frequency was calculated for first-person pronouns (e.g. “I”), prepositions, and nouns. Additional types of frequency variables of potential interest were found and are shown below in Table 1, but these results are not discussed in the scope of the report.

Data Analysis:

Running the word frequency calculation files automatically exported .csv format data table for each dataset with songs as the rows and associated variables as the columns. These are also available for viewing in the GitHub repository.

The linear correlation coefficient between artist age at time of song release and each of the word frequency variables was calculated in Microsoft Excel with the CORREL() function. The significance level was pre-set at $p < 0.05$. P-values were calculated with the online calculator recommended by Professor Cushman: https://www.socscistatistics.com/pvalues/pearsondistribution.aspx.

3 Results

Table 1. Linear Correlation Measurements Between Word Frequency Variables and Artist Age

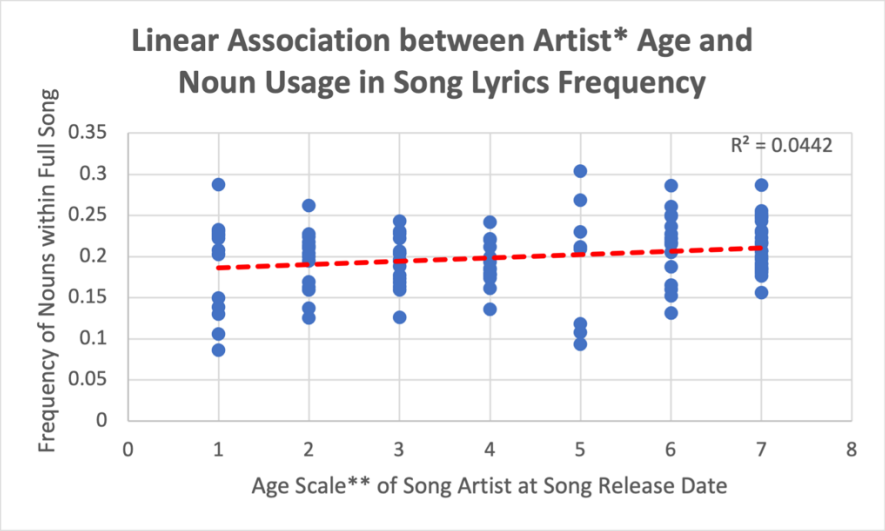
Word Frequency Variable	Examples	Correlation Coefficient *r*	P-value ($p < 0.05$)
I-freq ("I" Usage)	"I"	0.0494	0.6001
1stsg-freq (First person singular pronouns)	"I", "me", "my", "mine", "myself"	0.1484	0.1135
1stpl-freq (First person plural pronouns)	"we", "us", "our", "ours", "ourselves"	0.0251	0.7900
2ndsg-freq (Second person singular pronouns)	"you", "your", "yours", "yourself"	-0.0915	0.3308
preposition_freq (prepositions, as tagged by NLTK part-of-speech (pos) tagger)	"of", "to", "for", etc.	0.1780	0.0570
prp_ALL_freq (all personal pronouns, as tagged by NLTK pos tagger, including possessive pronouns)	"I", "we", "your", etc.	0.0991	0.2920
noun_freq (nouns, as tagged by NLTK pos tagger)	"girl", "city", "peace", etc.	**0.2040***	**0.0288**

Note: N = 115

* Value is significant with $p < 0.05$.

The linear correlation coefficients between artist age at time of song release and each of the word frequency variables, along with Pearson's p-values (significance level $p < 0.05$), are shown in Table 1. The relationship between age and noun frequency was statistically significant with correlation coefficient $r = 0.2040$, p-value $0.0288 < 0.05$, $n = 115$. Figure 1 displays this slight positive relationship graphically.

Figure 1. Slight Positive Relationship Between Age and Noun Usage Frequency



** The analyzed dataset only included works by one artist, Taylor Swift.*
*** See 2.3 Procedure for details on how age scale values were determined.*

4 Discussion, Limitations, and Future Work

Most results were not statistically significant, which is to be expected given the relatively small sample size of songs: $n = 115$ for the Extended-Full Lyrics. It could be the case that the null hypothesis is true and there is no actual correlation between these variables with age, but further study is required to eliminate the possibility of sample size as a primary factor, as only the works and ages of one musical artist were analyzed in this project. A direct extension of this study could collect and analyze a larger corpus of songs from multiple singer-songwriters who satisfy the necessary condition having led a sufficiently long musical career for which comparative age analysis is feasible.

The statistically significant p-value for the correlation between noun frequency and age supports part 3) of the proposed hypothesis, that there is a positive relationship between artist age and noun usage. Though statistically significant, this result is not sufficient to determine with certainty the existence of such a trend in the larger popular of singer-songwriters, as only one artist was studied here. Nevertheless, it is an additional incentive for future work in applying this methodology to a larger dataset of artists and songs.

Other limitations include potential confounding factors that were difficult to account for in this preliminary study, including those caused by the intrinsic framework of music, such as rhythm, song structure, word rhyming schemes, and music genre. Further studies should be conducted to better determine how to account for such variables and whether they present additional variables to analyze for possible insightful patterns.

Further ideas for future studies include: studying the relationship between age and other kinds of linguistic variables, e.g. positive or negative affect words, emotion words, different tenses of verbs, etc.; studying language usage patterns across generations of artists; conducting such analyses by genre and then comparing.

5 References

1. Pennebaker, J. (2011). *The Secret Life of Pronouns*. Bloomsbury Publishing.
2. Pennebaker, J. W., & Stone, L. D. (2003). Words of wisdom: Language use over the life span. *Journal of Personality and Social Psychology*, *85*(2), 291–301. https://doi.org/10.1037/0022-3514.85.2.291

Note from the author: This project was a fascinating interdisciplinary opportunity to study the sociolinguistic and psychological phenomena from James Pennebaker's foundational work and

apply such findings in a new way to study musical artists. I was certainly engrossed and plan to continue developing this project in the future.