Journal of
Social and Economic Statistics

Vol. 14, No. 2, Winter 2025
***DOI**: 10.2478/jses-2025-0008*

ZIPF'S LAW AND THE EVOLUTION OF MUSICAL LYRICS: A COMPARATIVE ANALYSIS OF THE 1960s, 1990s AND THE PRESENT

Camelia-Maria DAN[a*], Cristina-Rodica BOBOC[b]

[a,b] *The Bucharest University of Economic Studies, Romania*

Abstract

Song lyrics offer valuable insights into the social and cultural contexts of each period, reflecting the themes and concerns of the time. This study applies Zipf's law to analyse the evolution of lyrical complexity using data from the Billboard Year-End Hot 100 singles during three distinct periods: 1959–1969, 1991–1999, and 2016–2024. To assess repetitiveness trends, the analysis employs the Zipf coefficient ($\hat{\alpha}$) and the coefficient of determination (R^2) derived from log-log models. The findings reveal lower repetitiveness in both the 1960s and the contemporary period, according to the value of the coefficient α, with higher degree observed in the 1990s. Contrary to the initial hypothesis, the results suggest that the digital age has not contributed to the simplification of lyrics. This research underscores the influence of technology on music production and opens avenues for further exploration into its impact on other forms of art and culture.

Keywords:Music, Zipf's Law, TikTok, Lyrics, Complexity

JEL Classification:Z11, Z13, C46, L82

***Acknowledgements**: This research was conducted within the Applied Statistics and Data Science master's programme at the Bucharest University of Economic Studies. The authors would like to express their sincere gratitude to *Proffesor Claudiu Herteliu* for providing the academic context and support for this work.

1. Introduction

Zipf's work, carried out over a period of twenty years, began with his doctoral thesis and articles "Relative Frequency as a Determinant of Phonetic Change" (Zipf, 1929) and "Selected Studies on the Principle of Relative Frequency" (Zipf, 1932). His activity continued in two stages. In 1935 with "The Psycho-Biology of Language, an Introduction to Dynamic Philology" (Zipf, 1935) and in 1949 with "Human Behaviour and the Principal of Least Effort, an Introduction to Human Ecology" (Zipf, 1949).

In the early 1940s, George Kingsley Zipf noticed that the frequency of words in a text is inversely proportional to their rank, so that the most used words appear significantly more frequently than the least used. This regularity, later referred to as Zipf's law, contributed to the development of quantitative

* Corresponding author, Camelia-Maria Dan – dancamelia20@stud.ase.ro

linguistics by providing statistical insight into lexical distribution in both written texts and speech streams (Zipf, 1949). Zipf's law is an example of a law of power behavior (Reed, 2001) and this observation was later extended by Benoit Mandelbrot to explain a wider range of natural phenomena, called fractal geometry in nature. This theory was based on the Principle of Minimum Effort, which argues that there is a balance between unification and diversity in a language, ensuring that the vocabulary is not overly repetitive, which would make communication too simplistic, nor excessively varied, which would result in too many possible combinations and make communication unclear and complicated (Mandelbrot, 1977).

One of the main applications of Zipf's law is in the social and economic sciences where the study of different forms of communication has always been of considerable interest for comparative research (Patel, 2003). Both music and natural language are structured means of expression, and their generation and perception are considered to use basic neural mechanisms. This connection has led to the application of linguistic concepts in music analysis, such as grammar, syntax, and the meaning of lyrics (Lerdahl & Jackendoff, 1996). Zipf also provided several examples of how his distribution appears in musical compositions. As these investigations were carried out before the availability of computational tools, the analyses were performed manually and focused on meaning the intervals between repeated notes and the distribution of melodic intervals. Among the analyzed works were "B flat" by Mozart, "Etude" in F minor, "Op. 25, No. 2" by Chopin, "Doing What Comes Naturally" by Irving Berlin and "Who?" by Jerome Kern (Zipf, 1949).

"Is it possible to quantify the degree of complexity in a system by measuring its distance from both randomness and order?". A preliminary empirical answer to this question was provided in the 1930s by the philologist and economist G. K. Zipf, who statistically studied the repetition of words in the transcription of long speeches and in various written texts. He discovered a strong regularity between the relative frequencies of word occurrences, now known as Zipf's law, which is widespread among different authors, styles, and languages. The law says that the number of repetitions of the n^{th} most used word is approximately proportional to the inverse of rank n (Zanette, 2008). According to the works of Boroda & Polikarov (1988) and Manaris et al. (2003), a regularity similar to Zipf's law can also be observed in musical compositions.

Previous studies have applied Zipf's law to analyze various aspects of music, such as pitch, duration, and melodic and harmonic intervals, and have developed Zipf-based metrics to evaluate the enjoyability of music. Higher-order and fractal aspects of musical balance have been explored, demonstrating that Zipf distributions on certain dimensions are necessary conditions for considering a piece of music enjoyable. These metrics have been used to identify musical genres and even composers, and some studies have applied techniques such as neural network classification. However, although progress has been made in understanding musical structure through the lens of Zipf's law, we have not identified any research that analyzes the evolution of lyrics over time.

This paper explores the possibility of extending to the field of music a quantitative feature of language related to the frequency of word use, namely Zipf's law. In this context, Zanette (2006) showed that the statistical distribution of note usage in musical compositions follows a Zipfian pattern and can be explained by the predictions of Simon’s model. The study is based on the automatic analysis of digital MIDI files covering compositions from different musical styles and periods. This approach suggests that linguistic models applied in text analysis can be extended to music as well, demonstrating similarities between the cognitive processes involved in language and music (Zanette, 2008).

One research hypothesis is based on the fact that during the last 15 years, repetition has become an essential element in the composition of modern songs, and as artists want popularity on social media and easy-to-memorize songs, song lyrics will also become more repetitive.

The primary objective of this paper is to examine how the complexity of popular music lyrics in the United States has evolved over time by applying Zipf's Law as an analytical tool. Rather than aiming to validate the law itself, the study uses its rank-frequency framework to assess patterns of repetitiveness and lexical diversity across different historical periods. In this context, three distinct eras were compared in order to explore potential changes in lyrical structure and to consider how broader technological and cultural developments, including the diffusion of the internet and later of social media, may relate to shifts in linguistic expression within song lyrics.

2. Literature review

The analysis of the evolution of lyrics in music and how they reflect social and cultural changes is an important contribution to the social sciences. One of the pioneering large-scale investigations in this area was conducted by Voss & Clarke (1975). They collected data continuously for 24 hours from radio stations broadcasting classical, jazz, blues, and rock music and measured dynamic physical variables, including the output voltage of an audio amplifier, fluctuations in loudness, and fluctuations in pitch. The study revealed that the fluctuations in pitch and loudness in the music follow Zipf's cast. Also in 1975, the two authors developed a computer program to generate music using a noise source based on Zipf's distribution (also known as noise $1 / f$ or pink noise). The results were remarkable: the music obtained by this method was considered by most listeners to be much more enjoyable than that generated using a white noise source (which produced music that was too random) or a $1 / f^2$noise source, which produced music that was too similar. In fact, the complexity of this "$1 / f$ music" goes far beyond what one would expect from such a simple algorithm, suggesting that "$1 / f$ noise" could play an essential role in the creative process (Manaris et al., 2003).

The analysis of musical traits used in content-based search systems was studied by Chi Lap Yip & Ben Kao (1999) using Zipf's law, with the objective of assessing whether these traits comply with the statistical patterns observed in language. The law was applied to a collection of popular songs from Hong Kong, Taiwan and Japan, represented as MIDI files, analyzing six musical features: interval sequences, note duration ratios, interval profiles and their variants. The study addressed four major topics: the number of musical alphabets and possible words, the conformity of musical features to Zipf's law, the identification of possible musical stop words and the determination of the optimal value for musical indices. The results showed that sequences of short features, such as bigamous and trigamous, do not perfectly fit a straight line in the Zipfian chart, exhibiting convex behavior. However, at a general level, all six traits follow Zipf's law, which suggests that they can be modeled as stochastic processes, allowing for the application of techniques such as co-occurrence analysis for indexing and classification. The study identified musical stop words, such as small and frequent intervals, that can be excluded to reduce the size of the musical cues, and variations in note duration provide distinct information about the musical pieces.

Manaris et al. (2003) applied the Zipf-Mandelbrot law in the context of music in order to identify and describe the conditions under which a musical piece is perceived as pleasant by listeners. The authors analyzed a corpus of 220 musical pieces spanning multiple genres, including baroque, classical, romantic, rock, and jazz, and developed a set of Zipf-based metrics derived from fundamental musical characteristics. These include pitch, defined as the relative balance of pitch values within a piece,

duration, capturing the distribution of event lengths independently of pitch, as well as melodic intervals, which describe the distribution of pitch changes over time. The study highlights that Zipf distributions are a necessary, but not sufficient, condition for identifying music as enjoyable, suggesting that a combination of these metrics can be used to identify musical genres and even composers. The results show that the parts considered pleasing comply with Zipf distributions across multiple dimensions, and the classification of the parts based on these metrics was validated using an Artificial Neural Network. The authors conclude that these metrics can be used as fitness features, guiding the generation of songs that reflect the structure of highly listened to songs, i.e., hit songs. The study offers a new perspective on how statistical analysis can bring new insights into the music industry.

The analysis of the statistical properties of the distributions of timbral codes extracted from different audio recordings, such as free speech, Western and non-Western music, but also sounds of nature, was studied by Haro et al. (2012). The authors found that these distributions follow a Zipfian law with exponents close to one, regardless of the audio source. The authors showed that the most common timbral codes exhibit a more homogeneous structure. The study was conducted on 740 hours of recordings, using the Bark-band encoding method, which encodes short-time spectral envelopes into perceptually motivated frequency bands that approximate the frequency resolution of the human auditory system, yielding normalized timbral representations. The results showed that the frequency distribution is independent of the encoding method, the length of the analysis window, or the type of audio source. In addition, the analysis identified timbral codes specific to language and music that are absent in natural sounds, suggesting the existence of a common mechanism for generating these distributions. Finally, the authors showed that a Yule-Simon process with memory provides a reasonable quantitative approximation of the observed data, suggesting the presence of a simple generative mechanism underlying the temporal recurrence of timbral patters (Haro et al., 2012).

Tonal and atonal music were analyzed using Zipf's law by Martinson & Zieliński (2019) to compare the statistical features of the pieces "Allemande" from "Partita in A minor" by Johann Sebastian Bach and "Syrinx" by Claude Debussy. The study used specialized software to calculate the occurrence frequencies of ascending and descending intervals, pitches, and rhythmic durations, with the results plotted on a log-log scale. The authors found that the frequency histograms for these features showed significant deviations from linearity, and the slope coefficient (α) was often less than -1, indicating that Zipf's law was not followed for the intervals in both musical pieces. The results showed that "Syrinx" approaches a (pink) noise, marked by $1 / f$, and "Partita in A minor" a Brownian noise, marked by $1 / f^2$, which emphasizes the differences between tonal and atonal pieces. Although the slope for tonal pitches approached the expected value of -1 in Bach's case, the coefficients of determination (R^2) were lower, indicating a poor fit. In the case of rhythms, only "Syrinx" partially complied with Zipf's law ($\hat{\alpha} = -1.14$), while Bach's exhibited a much steeper slope ($\hat{\alpha} = -4.19$).

Another insight into the application of Zipf's law in music was provided by Perotti & Billoni (2020), who investigated the statistical distributions of notes and chords in musical scores. The study, based on works by composers such as Bach, Vivaldi, Piazzolla, and Stravinsky, highlighted that individual notes better follow an extended exponential distribution, while generalized chords, defined as sets of simultaneous notes, follow the Zipf-Mandelbrot law with exponents close to one. The results suggest that generalized chords can be considered natural Zipfian units in music, indicating structural similarities between music and language.

3. Methodology

In this study, the applicability of Zipf's law was analyzed across three key periods in music history, using three-year intervals for a more precise examination. Specifically, the music from 1959 to 1969 was analyzed in three segments: 1959-1961, 1964-1966, and 1967-1969. This was followed by the period 1991-1999, divided into 1991-1993, 1994-1996, and 1997-1999. These periods are compared with the period following the intensification of social media use, which is believed to have influenced the repetitiveness and reduced the complexity of music. This includes the years just before the emergence of TikTok (2016-2018), and the periods when the app gained widespread popularity: 2019-2021 and 2022-2024.

Billboard is a renowned American magazine, founded in 1894, specializing in the music industry and known for publishing music charts that reflect the popularity of songs and albums. The most important of these is the Billboard Year-End Hot 100, introduced in 1958, which ranks the 100 most successful singles in the United States at the end of the year, based on CD and concert ticket sales, especially 60 years ago, radio or TV broadcasts, and more recently, social media appearances. Over time, the methodology for calculating the ranking has evolved to adapt to changes in the way music is listened to. (Lech, et al., 2024)

The source of the data used in this study consists of the annual Billboard Year-End Hot 100 charts, available on the Wikipedia platform1, from which song titles and artist names were extracted for each year of interest, the years being 1959, 1960, 1961, 1964, 1965, 1966, 1967, 1968, 1969, being described by the period 1959-1969, then 1991, 1992, 1993, 1994, 1995, and 1996, 1997, 1998, 1999 resulting from the period 1990-1999, and the most recent period, 2016-2024, consists of the analysis of the lyrics from the years 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023 and 2024. Thus, the sample contains 100 songs for each year of interest, resulting in 2700 songs. The process of collecting the lyrics was an iterative one, using a Python script that queried the API of the Genius platform, recognized for its complete collection of musical lyrics. The data has been saved in .txt format to facilitate further analysis.

The analyses were performed using the Python work environment, using classic libraries such as Pandas and NumPy, which allow calculations and table creation, matplotlib for graphs, and the WordCloud library for visual analysis of word frequencies. For reading and processing the data, the docx library was used, which allows the loading of Word documents, and the SciPy and Sci-kit learn libraries were used for the regression model and statistical processing. The BeautifulSoup library was used for the web scrapping technique, which helped to extract the lyrics.

Data preprocessing was an important step in calculating the frequencies of words in song lyrics. Changes have also been made to the usual abbreviations in English, transforming them into its extended forms to preserve the full meaning of the sentences and avoid erroneously increasing the frequency of the letters "t" or "s" coming from forms such as "don't" and "it's". In practice, this meant replacing expressions like ‘till, don’t, won’t, can’t, I’ll, I’m, you’re, they’re, who’s, what’s, it’s, and isn’t with their extended versions: until, do not, will not, cannot, I will, I am, you are, they are, who is, what is, it is, and is not.

An important technical aspect of the preprocessing stage was the decision to treat uppercase and lowercase letters as equivalent and to remove punctuation marks such as “…”, “-“, “ “, “!” or “?”, as these do not contribute to the analytical objective. Beyond this, several additional procedures were applied to

[1] https://en.wikipedia.org/w/index.php?search=Billboard+YearEnd+Hot+100+singles&title=Special%3ASearch&ns0=1 accessed on 22 December 2025

ensure that the text was processed in a consistent and analytically sound manner. First, all instances of multiple consecutive spaces were reduced to a single space to maintain uniformity and facilitate correct word separation. Subsequently, all letters were converted to lowercase, with the exception of the pronoun "I", which was preserved in its capitalized form to comply with standard English grammar. Furthermore, all punctuation symbols, including ".", ",", ":", "/", "?", "(", ")", "#", "'", "$", "%" and "@" were removed, as they do not hold semantic value for the analysis. Finally, all numerical characters were eliminated, given that they are not relevant to the computation of word frequencies.

The processing of the lyrics also involved removing the most frequent articles, conjunctions, and prepositions in English. Specifically, the articles "a", "an", and "the" were excluded, as their high frequency does not provide meaningful information for word-frequency analysis. Similarly, the conjunctions "and", "but", and "or", as well as the prepositions "in", "on", "at", "with", and "to", were removed for the same reason.

The periods analyzed in this study were selected to reflect significant moments in musical history, marked by changes in genres, new technologies and social and political trends. The period 1959-1961 was chosen because it represents the first three years of the Billboard Year-End Hot 100 chart, with dominant genres such as rock 'n' roll, which defined the beginnings of modern pop music (Lech et al., 2024). Between 1964 and 1966, the British invasion transformed the American music scene with the emergence of The Beatles, marking a global cultural revolution (Charlton, 2003). The following years, 1967-1969, were characterized by psychedelic music and the influence of social movements, reflected in albums such as Sgt. Pepper's Lonely Hearts Club Band by The Beatles (Macdonald, 2008).

Between 1990 and 1993, the American music industry underwent a major transition, driven by the popularization of MTV, the mainstream breakthrough of grunge and alternative rock, symbolized by Nirvana's 1991 album *Nevermind*, and the growing commercial prominence of rap music (Reynolds, 2011). The period 1994-1996 brought the diversification of genres, with rap dominating the charts and the formation of boy bands such as the Backstreet Boys (Cepeda, 2004). In 1997-1999, pop music became the most listened music genre, with artists such as Britney Spears redefining the musical landscape and bringing it closer to contemporary music genres (Wald, 2011).

The last periods, 2016-2018 and 2019-2021, have the role of reflecting the influence of technology on musical evolution. Before the launch of TikTok in 2016, songs were traditionally promoted, but with the emergence of social media in 2018, algorithms and viral trends have completely transformed the music industry. The period 2022-2024 marks the effect propagated by the music industry since the beginning of 2016, characterized by the blending of music genres, the central role of digitization, and the dominance of virality driven by social media in the music industry (Mathews, 2024).

Zipf's law (Zipf, 1949) describes a statistical regularity in language, according to which the frequency of occurrence of a word is inversely proportional to its rank in the frequency list. This means that the r-most common word occurs with a frequency approximately equal to the frequency of the most frequent word, divided by r. Mathematically, the law is expressed by the formula:

$$P(r) \sim \frac{1}{r^{\hat{\alpha}}} \quad (1)$$

where $P(r)$ represents the frequency of occurrence of the word with the rank r; r is the rank of the word in descending order of frequency; α is a positive exponent, usually close to 1.

This relationship can be converted into a logarithm regression model, to facilitate statistical analysis:

$$ln\big(P(r)\big) = -\hat{\alpha} \cdot ln(r) + C + \mu \quad (2)$$

where $P(r)$ represents the frequency of occurrence of each word; C and $\hat{\alpha}$ are constants estimated by the

Ordinary Least Squares Method (OLS), and $\hat{\alpha}$ is known as the Zipf coefficient; r is the rank of the word, with 1 indicating the word with the highest frequency of occurrence and n the word with the lowest frequency of occurrence. For words with the same frequency of occurrence, the average rank was calculated as the arithmetic average of the individual positions; μ is the random component, independent of the other variables in the regression equation.

This relationship shows that in most natural languages, a relatively small number of words appear very frequently, while the majority occur only rarely. Zipf's Law reflects the principle of least effort, suggesting that language evolves toward an equilibrium between communicative efficiency and lexical diversity.

The parameter α is estimated from the slope of the regression line in the log-log plot of word frequencies and typically takes a value close to 1. A steeper slope, corresponding to $\hat{\alpha} > 1$, indicates a more restricted and repetitive vocabulary, whereas a flatter slope, with $\hat{\alpha} < 1$, is associated with a more diverse lexical structure (Ausloos et al., 2015). Zipf's law holds when the slope $\hat{\alpha}$ on the double logarithmic scale is equal to -1.

Departures from this idealized straight-line relationship are captured by the coefficient of determination R^2, which ranges from 0 (poor fit) to 1 (perfect fit) and evaluates the degree to which the model fits the data (Martinson, 2019). The coefficient of determination R^2 represents the proportion of the variance explained by the fitted trend (Gao, 2023), using the equation:

$$R^2 = 1 - \frac{SSR}{SST} \quad (3)$$

where $SSR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is called residual sum of squares, and $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is called total sum of squares.

To assess whether the estimated Zipf coefficient differs statistically from the theoretical value of 1, a two-sided Student's test was applied (Lumley et al., 2002), under the following hypotheses:

H_0: $\hat{\alpha} = 1$ (Zipf's law is respected.)

H_1: $\hat{\alpha} \neq 1$ (Zipf's Law is not respected)

where $se(\hat{\alpha})$ denotes the standard error of the estimated coefficient $\hat{\alpha}$.

The statistic of the test is given by the relationship:

$$t_c = \frac{\hat{\alpha}-1}{se(\hat{\alpha})} \quad (4)$$

If $t_c > +t_{\frac{\hat{\alpha}}{2},n-k}$ or $t_c < -t_{\frac{\hat{\alpha}}{2},n-k}$, or p-value is lower than the chosen significance level, then the null hypothesis will be rejected. We conclude that there is a significant difference between the value of the coefficient α and 1, indicating a significant deviation from Zipf's Law.

4. Results

Table no. 1 and Figure no. 1 present the key descriptive statistics used in the lexical analysis, highlighting the dynamics of lyrical complexity and the data source.

Table no. 1: Descriptive statistics

Period	Total word count	Number of distinct words	The most common word	Source
1959-1961	46.795	3.602	you (1.919)	Billboard 1959
				Billboard 1960
				Billboard 1961
1964-1966	53.223	3.444	you (2.320)	Billboard 1964
				Billboard 1965
				Billboard 1966
1967-1969	60.660	4.382	you (2.777)	Billboard 1967
				Billboard 1968
				Billboard 1969
1959-1969	**160.678**	**7.274**	**you (7.016)**	
1991-1993	99.596	5.522	you (7.016)	Billboard 1991
				Billboard 1992
				Billboard 1993
1994-1996	104.530	5.837	you (5.334)	Billboard 1994
				Billboard 1995
				Billboard 1996
1997-1999	116.554	6.089	you (6.172)	Billboard 1997
				Billboard 1998
				Billboard 1999
1991-1999	**320.680**	**10.673**	**you (15.970)**	
2016-2018	25.493	6.872	I (5.934)	Billboard 2016
				Billboard 2017
				Billboard 2018
2019-2020	117.107	6.782	I (5.820)	Billboard 2019
				Billboard 2020
				Billboard 2021
2022-2024	106.137	7.148	I (5.094)	Billboard 2022
				Billboard 2023
				Billboard 2024
2016-2024	**348.737**	**12.798**	**I (16.848)**	

Source: Own processing using Python

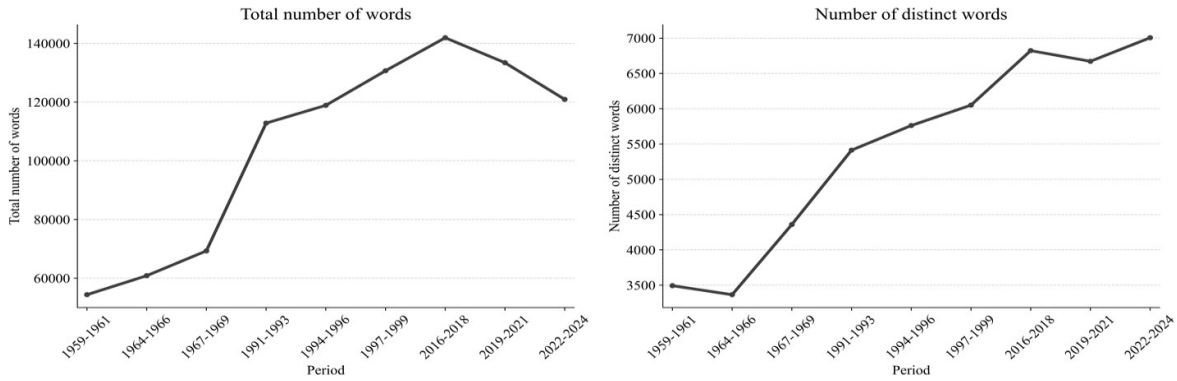

Figure no. 1: Evolution of lyrics between 1959 and present

Source: Own processing using Python

Table no. 2 shows the top five words with the highest frequency, and the last five words with the lowest frequency, serving as an exploratory analysis between periods, on the complexity of the song's lyrics.

Table no. 2: Top 5 common and rare words for each period

Period	Top 5 words	Last 5 words	Period	Top 5 words	Last 5 words	Period	Top 5 words	Last 5 words
1959-1961	you	idle	1991-1993	you	dunkin	2016-2018	I	Biscayne
	I	rambled		I	donuts		you	guillotine
	me	apron		me	pops		me	promethazine
	love	clicking		my	muggs		it	lone
	my	canyon		love	unreal		my	position
1964-1966	you	handout	1994-1996	you	insecure	2019-2021	I	shooter
	I	names		I	driver		you	taming
	me	engineer		me	bangers		me	equally
	my	dolls		my	bungee		my	silly
	love	stakes		love	sedative		it	pimple
1967-1968	you	federal	1997-1999	you	invade	2022-2024	I	curtain
	I	reliable		I	passions		you	fertile
	me	deadly		me	requested		me	kidnapped
	my	wiser		my	certainties		it	pine
	love	jail		that	resilient		my	planet

Source: Own processing using Python, based on data extracted from the Billboard Year-End Hot 100 Singles

In data mining, the frequency of words is often measured visually, using a Word Cloud chart, as shown in Figure no. 2. Such visualizations allow us to identify the most common word and compare frequencies according to the size of the text. Similarly, Zipf's law proves this theory, but not visually, using a log-log graph (Figure no. 3).

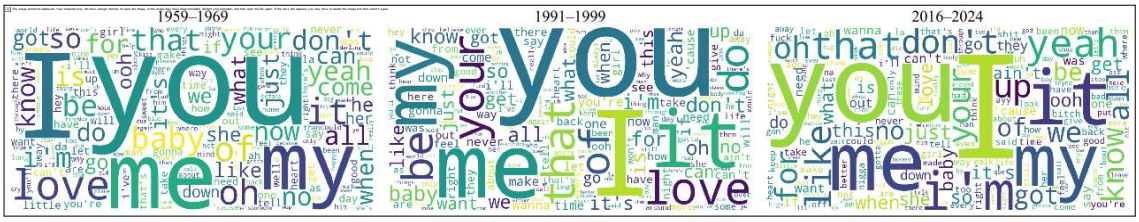

Figure no. 2: Word cloud for the analysed periods

Source: Own processing using Python

The values of Zipf's coefficients were included in Table no. 3, but also the performance of the models for all periods, both individual and cumulative over periods, measured by the coefficient of determination R^2.

The value of the Zipf coefficient has been highlighted over time with the help of Figure no. 3, allowing the identification of the periods with the highest and lowest coefficients.

Table no. 3: Zipf coefficient and its statistical significance

Period	Zipf coefficient (α)	Student test statistics	p-value	Coefficient of determination R^2
1959-1961	0.92	-0.43	0.664	0.960
1964-1966	0.96	-0.19	0.846	0.961
1967-1969	0.97	-0.16	0.874	0.967
1959-1969	**1.04**	**0.21**	**0.832**	**0.966**
1991-1993	1.04	0.20	0.840	0.966
1994-1996	1.04	0.25	0.806	0.971
1997-1999	1.06	0.33	0.744	0.969
1991-1999	**1.12**	**0.66**	**0.513**	**0.973**
2016-2018	1.03	0.21	0.833	0.977
2019-2021	1.02	0.14	0.892	0.975
2022-2024	0.99	-0.06	0.977	0.974
2016-2024	**1.09**	**0.53**	**0.596**	**0.978**

Source: Own processing using Python

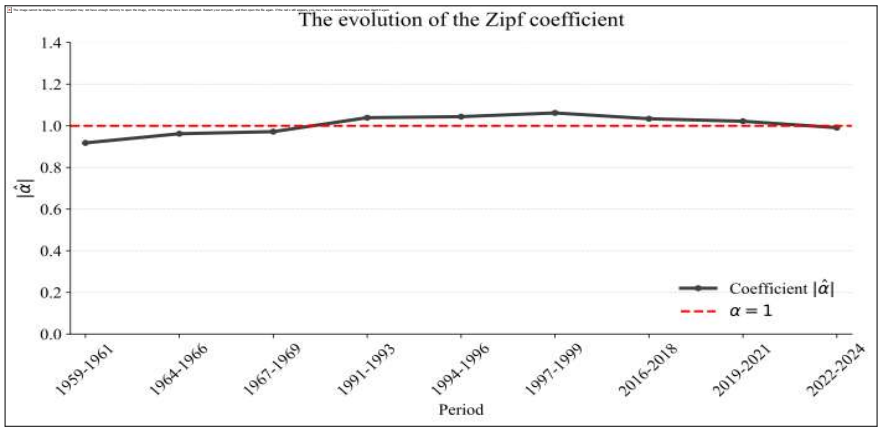

Figure no. 3: Zipf coefficient (alpha)

Source: Own processing using Python

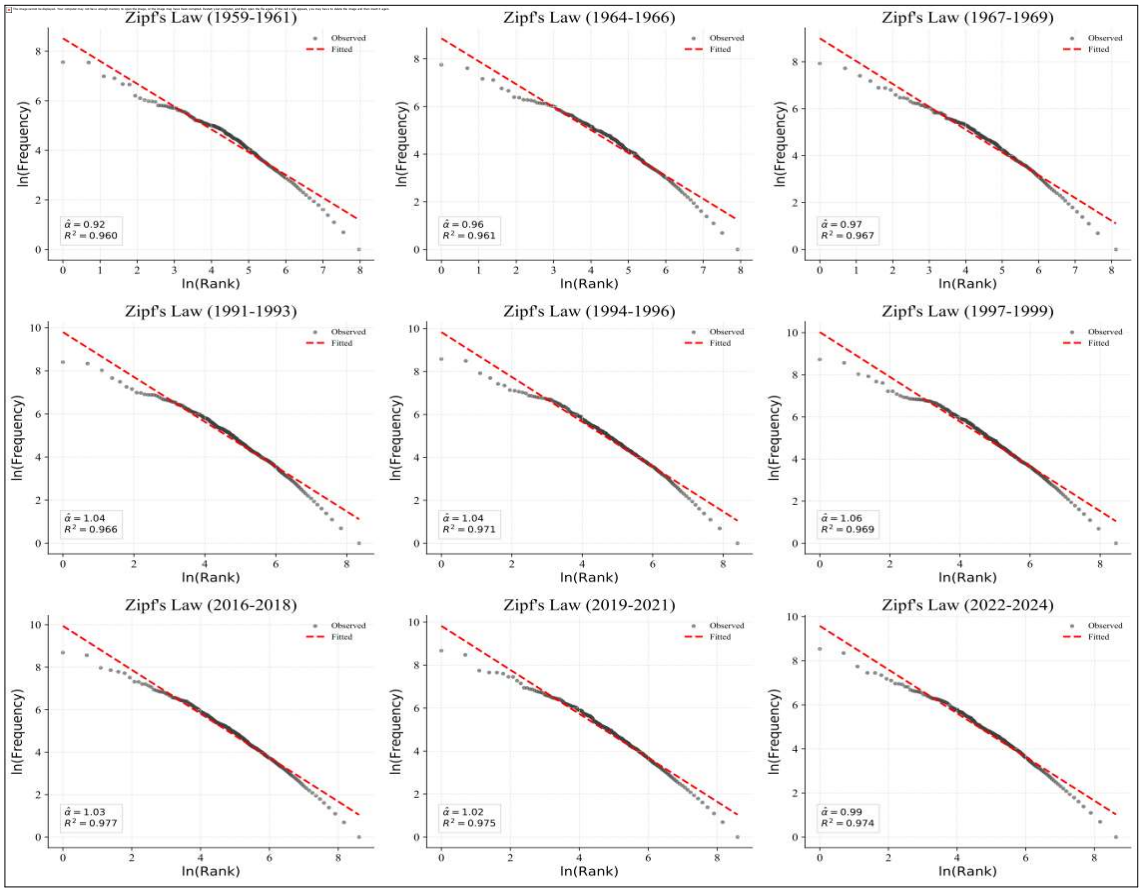

Figure no. 4: Range-frequency distribution according to Zipf's Law

Source: Own processing using Python

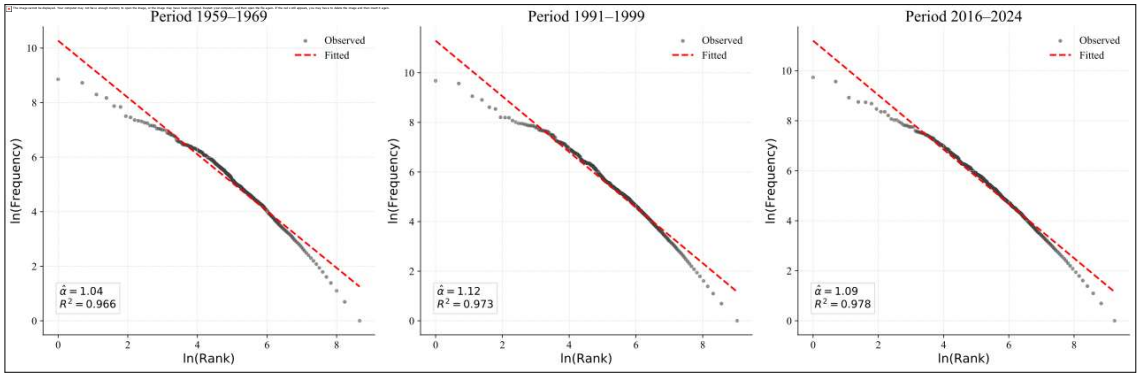

Figure no. 5: Zipf's law rank-frequency distribution for the periods 1959-1969, 1991-1999 and 2016-2024

Source: Own processing using Python

Figure no. 4 and Figure no. 5 show the deviations from the Zipf-Mandelbrot model for each period. The results will be interpreted in an effective manner in the next section dedicated to discussions.

5. Discussions

Table no. 1 and Figure no. 1 highlight the upward trend in the complexity of the musical lyrics of the most listened to songs in the United States, reflected by the gradual increase in the total number of words and the number of distinct words in the ranges analyzed. Between 1991 and 1999, the ratio of the total number of words to the number of distinct words is the highest, with each distinct word being repeated, on average, about 30 times. So, the lyrics from this period are the most repetitive, influenced by dominant genres of that period, such as rap, hip-hop and rock.

Between 2016 and 2024, the ratio decreases to approximately 27, suggesting a diversification of the vocabulary used and contradicting the initial assumption that contemporary lyrics are the most repetitive. Moreover, the period 1959-1969 shows a ratio of around 22, indicating a more diversified vocabulary compared to 1991-1999. Based on these results, the period 1959-1969 can be considered the most complex in terms of lyrical structure.

An interesting shift in the distribution of frequently used words can be observed across the analyzed periods. Between 1959 and 1999, the most frequent word was "you", a pattern that may reflect a lyrical focus on interpersonal relationships. In contrast, during the period 2016-2024, the word "I" became the most frequent, suggesting a greater emphasis on subjective, self-oriented expression in contemporary lyrics.

The results of the Zipf coefficient (Table no. 3) highlight the differences in the repetitiveness and diversity of vocabulary in the music industry over the periods analyzed. For individual periods in the 1960s, the estimated coefficient $\hat{\alpha}$ varies between 0.92 and 0.97, reflecting a complex and diversified vocabulary. The t-test shows that for all periods, for a significance level of 5%, we do not have enough evidence to reject the null hypothesis, so Zipf's law is respected. In the period 1967-1969, $\hat{\alpha} = 0.97$ indicates a greater diversity of vocabulary, and the t-test does not show a significant difference from the theoretical value of 1, with a p-value=0.874.

In the 1990s, the coefficient $\hat{\alpha}$ remains relatively stable, with values between 1.04 and 1.06, indicating a more restricted and repetitive vocabulary compared to the theoretical threshold of 1 (Ausloos et al., 2015). The t-tests for these years reveal no statistically significant deviation from 1, suggesting that Zipf's law is upheld during this period. Moreover, the coefficient of determination of 0.97, which exceeds the

commonly accepted threshold of 0.95, indicates a near-perfect fit to the model and strong consistency with Zipf's law.

For recent periods, the Zipf coefficient ranges between 0.99 and 1.03, suggesting a less diversified vocabulary than expected, but the t-tests show no significant deviations from 1, emphasizing the repetitive nature of the lyrics. For the entire period, the years 2016-2024 register $\hat{\alpha} = 1.09$, and the t-test does not show a significant difference from 1. Although the coefficient suggests a slight diversification of vocabulary from individual sub-periods, the slope being higher than in the period 1959-1969.

However, according to the coefficient of determination R^2, Zipf's law is respected for all periods, given that the values exceed the threshold of 0.95 (95% of the model variability is explained using the included explanatory variables). The period 1959-1969 remains the most diversified in terms of language, while songs from the 90s are associated with the most repetitive lyrics.

Figure no. 4 highlights the fact that for all periods, the Zipf coefficient is close to the value 1, considered in the literature to be the corresponding value confirming Zipf's law. However, the metrics can help with comparisons between periods regarding the complexity of the lyrics. Given that the values $\hat{\alpha} > 1$, associated with a smaller and more repetitive vocabulary, we can see that the period 1991-1999 is characterized by the most repetitive vocabulary, as well as the period 2016-2021. However, the most recent period shows that the lyrics no longer have the same repetitiveness, compared to the 90s.

Figure no. 4 displays the log-log plot of word frequency versus word rank in song lyrics across the selected time periods. Deviations from the regression line are noticeable, potentially reflecting the impact of various musical trends. These deviations are particularly pronounced during specific periods, such as 1991-1993, 1994-1996, and 1997-1999, suggesting that cultural shifts can significantly influence the frequency distribution. For completeness, the same relationship is also presented in Appendix using the original variables without logarithmic transformation. However, due to the strongly skewed and heavy-tailed nature of power-law distributions, this representation emphasizes the dominance of a small number of high-frequency words, while the majority of observations remain compressed near zero.

The indirect relationship between the frequency of the words and their rank confirms the applicability of Zipf's Law to the lyrics of the songs of these periods. This proves that, overall, the most frequently used words follow a similar distribution pattern, with word frequency decreasing as rank increases, an observation originally made by Zipf when analyzing the word usage in his students' written works (Zipf, 1949).

Similar deviations are evident in the combined periods (Figure no. 5), where departures from the regression line can be observed across all time intervals. However, the estimated values remain close to the theoretical expectations. These visual representations support the results obtained from the logarithmic regression models (Table no. 3) and do not support the initial hypothesis that songs in the digital age exhibit lower lyrical complexity than those composed in the 1960s or 1990s.

6. Conclusion

The findings of the study indicate that applying Zipf's Law provides useful insights into the evolution of lyrical complexity in songs from the Billboard Year-End Hot 100 across the analyzed periods. The results show a higher degree of repetitiveness in the 1991-1999 interval, reflected in a higher Zipf coefficient compared with the other periods, which may be associated with the prominence of genres such as rap,

hip-hop, and rock. These genres often employ stylistic patterns that rely on repeated lexical elements, frequently as part of rhythm-focused lyrical structures.

In contrast, the 2016-2024 period shows a slight increase in vocabulary diversity, reflected by a lower Zipf coefficient compared with the 1991-1999 interval, although it remains higher than the value observed for 1959-1969. This pattern may be related to the diversification of contemporary musical styles and the rise of short-form digital media platforms, which have influenced the structure and production of mainstream music. Furthermore, the 1959-1969 period displays the most diverse vocabulary among the analyzed intervals, suggesting higher lyrical complexity and a reduced degree of repetitiveness.

The analysis of the coefficient of determination indicates that Zipf's Law holds across all examined periods. For each interval, the R^2 value exceeded 0.95, suggesting that the Zipfian model accounts for more than 95% of the variability in the word-frequency distribution of song lyrics. This represents a key indicator of model fit and implies that, overall, the lyrics conform to the expected inverse relationship between word frequency and rank.

The t-test was applied to the estimated Zipf coefficients to assess whether they differ significantly from the theoretical value of 1. The results indicated no statistically significant deviations for any of the analyzed periods, suggesting that Zipf's Law holds in all cases. The 2016-2024 interval shows a moderate increase in vocabulary diversity, a pattern that may be associated with the broader globalization of musical production and the growing influence of diverse stylistic and cultural sources.

However, this study presents several limitations. First, the selected time intervals may not fully capture the diversity of the wider music landscape. Moreover, the analysis is based exclusively on songs from the Billboard Year-End Hot 100 charts, which may not be representative of the broader spectrum of musical production. In addition, the methodological approach assumes that repetitiveness and vocabulary diversity are sufficient indicators of lyrical complexity, without incorporating other potentially relevant dimensions such as cultural context or the influence of production practices on lyricism. Furthermore, the statistical properties of the residuals in the log-log regression framework are sensitive to the choice of the cutoff parameter used to define the range of included observations. In particular, the inclusion of a large number of low-frequency items from the tail of the distribution may affect the error structure and the validity of standard inference, potentially introducing heteroscedasticity and dependence patters.

As possible directions for future research, a more detailed examination of the evolution of the evolution of language in music, through advanced lexical analyses, could provide a clearer understanding of how social and cultural dynamics shape contemporary lyrical patterns. Future analyses could extend beyond word frequency to include semantic and connotative dimensions, highlighting period-specific linguistic features. In addition, future work could systematically investigate the sensitivity of Zipfian estimates to the choice of the cutoff parameter defining the lower bound of the distribution, with particular attention to the role of low-frequency items in shaping the residual structure and inferential properties. From a methodological perspective, alternative estimation approaches to ordinary least squares, such as maximum likelihood estimation, could be employed to address potential issues related to heteroscedasticity and dependence structures commonly observed in Zipfian data. Another relevant avenue would be to investigate the influence of digital platforms, such as Spotify or YouTube, on patterns of musical consumption and the dynamics of song popularity.

References

Ausloos, M., Nedic, O., Fronczak, A.,&Fronczak, P. (2015). Quantifying the quality of peer reviewers through Zipf's law. Scientometrics, 106(1), 347–368. https://doi.org/10.48550/arXiv.1508.05624

Boroda, M.G.,&Polikarpov, A.A. (1988). The Zipf-Mandelbrot law and units of different text levels. Musikometrika, 1, 127–158. https://doi.org/10.48550/arXiv.cs/0406015

Cepeda, R. (Ed.) (2004). *And it don't stop: the best American hip-hop journalism of the last 25 years*. Macmillan. Available at:https://archive.org/details/anditdontstopbes00ceped

Charlton, K. (2003). *Rock Music Styles*. McGraw-Hill Humanities, Social Sciences & World Languages. Available at: https://archive.org/details/rockmusicstylesh0000char_04ed

Gao, J. (2023). R-Squared – How much variation is explained? Research Methods in Medicine & Health Sciences, 5(4). https://doi.org/10.1177/26320843231186398

Haro, M., Serrà, J., Herrera, P.,& Corral, Á. (2012). Zipf's Law in Short-Time Timbral Codings of Speech, Music, and Environmental Sound Signals. PLoS ONE, 7(3), e33993–e33993. https://doi.org/10.1371/journal.pone.0033993

Lap, Y.C.,& Kao, B. (1999). A study on musical features for melody databases. International Conference on Database and Expert Systems Applications, 724–733. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-48309-8_67

Lech, M.E., Lehmann, S.,&Juul, J.L. (2024). Is it getting harder to make a hit? Evidence from 65 years of US music chart history. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2405.07574

Lerdahl, F.,& Jackendoff, R. (1996). A generati*ve theory of tonal music*. Cambridge, Mass.: Mit Press.

Lumley, T., Diehr, P., Emerson, S.,& Chen, L. (2002). The Importance of the Normality Assumption in Large Public Health Data Sets. Annual Review of Public Health, 23(1), 151–169. https://doi.org/10.1146/annurev.publhealth.23.100901.140546

Macdonald, I. (2008). *Revolution in the head: the Beatles' records and the sixties*. London: Random House.

Manaris, B., Vaughan, D., Wagner, C.C., Antonio, J.,& Davis, R.H. (2003). Evolutionary Music and the Zipf-Mandelbrot Law: Developing Fitness Functions for Pleasant Music. Lecture Notes in Computer Science, 522–534. https://doi.org/10.1007/3-540-36605-9_48

Mandelbrot, B.B. (1977). *The fractal geometry of nature.* W.H. Freeman and Company.

Martinson, K.,& P. Zieliński (2019). Statistical Properties of Tonal and Atonal Pieces of Music. Acta Physica Polonica B Proceedings Supplement, 13(4), p.815. https://doi.org/10.5506/aphyspolbsupp.13.823

Mathews, A. (2024). The Future of Music is in AI; Thanks To Spotify. [online] AIM Research | Artificial Intelligence Market Insights. Available at: https://aimresearch.co/market-industry/the-future-of-musicis-in-ai-thanks-to-spotify

Patel, A.D. (2003). Language, music, syntax and the brain. Nature Neuroscience, 6(7), 674–681.

Perotti, J.I.,&Billoni, O.V. (2020). On the emergence of Zipf's law in music. Physica A: Statistical Mechanics and its Applications, 549, 124309.

Reed, W.J. (2001). The Pareto, Zipf and other power laws. Economics Letters, 74(1), 15–19. https://doi.org/10.1016/s0165-1765(01)00524-9

Reynolds, S. (2011). *Retromania: Pop culture's addiction to its own past.* Macmillan. Available at: https://archive.org/details/retromaniapopcul0000reyn

Voss, R.F.,& Clarke, C.J. (1975). 1/f noise in music and speech. *Nature,* 258, 317–318.

Wald, E. (2011). How the Beatles destroyed rock 'n' roll: an alternative history of American popular music. New York; Oxford: Oxford University Press.

Zanette, D.H. (2006). Zipf's law and the creation of musical context. Musicae Scientiae, 10(1), 3–18. https://doi.org/10.1177/102986490601000101

Zanette, D.H. (2008). Music, Complexity, Information. arXiv (Cornell University). https://doi.org/10.48550/arxiv.0807.0565

Zipf, G.K. (1929). Relative frequency as a determinant of phonetic change. Harvard Studies in Classical Philology, 40, 1–95. Available at: https://scispace.com/pdf/relative-frequency-as-a-determinant-of-phonetic-change-1isfubt996.pdf

Zipf, G.K. (1932). Selected studies of the principle of relative frequency in language. Harvard University Press. Available at:https://pure.mpg.de/rest/items/item_2407800_3/component/file_2459540/content

Zipf, G.K. (1935). The Psycho-Biology of Language: An Introduction to Dynamic Philology. Houghton Mifflin Company. Available at: https://iqla.org/wp-content/uploads/2025/05/Zipf_1935_1965_Psycho-Biology_Language.pdf

Zipf, G.K. (1949). Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology. Addison-Wesley Press. Available at: https://archive.org/details/in.ernet.dli.2015.90211/page/n7/mode/2up

Appendix

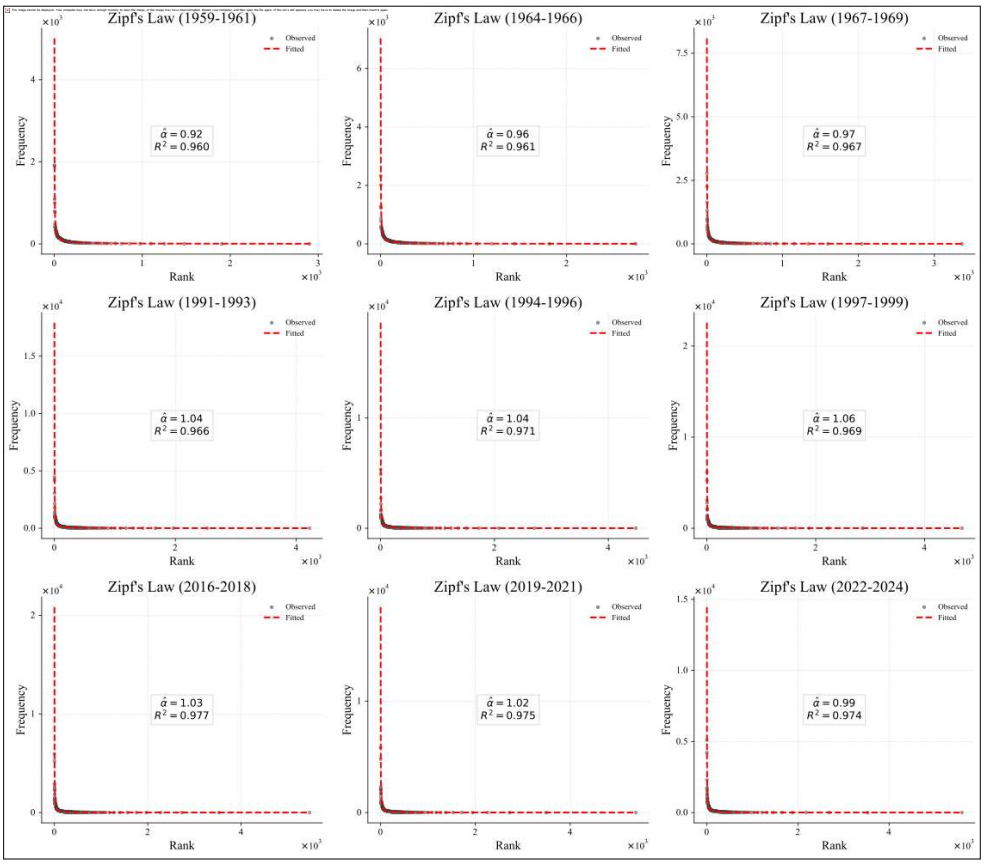

Figure no. A1: Rank–frequency distributions of song lyrics under Zipf's Law based on original variables

Source: Own processing using Python