Genre Differences of Song Lyrics and Artist Wikis: An Analysis of Popularity, Length, Repetitiveness, and Readability

Markus Schedl
Johannes Kepler University Linz
Institute of Computational Perception
Linz, Austria
markus.schedl@jku.at

ABSTRACT

Music is known to exhibit different characteristics, depending on genre and style. While most research that studies such differences takes a musicological perspective and analyzes acoustic properties of individual pieces or artists, we conduct a large-scale analysis using various web resources. Exploiting content information from *song lyrics*, contextual information reflected in *music artists' Wikipedia articles*, and listening information, we particularly study the aspects of *popularity*, *length*, *repetitiveness*, and *readability* of lyrics and Wikipedia articles. We measure popularity in terms of song play count (PC) and listener count (LC), length in terms of character and word count, repetitiveness in terms of text compression ratio, and readability in terms of the Simple Measure of Gobbledygook (SMOG). Extending datasets of music listening histories and genre annotations from Last.fm, we extract and analyze 424,476 song lyrics by 18,724 artists from LyricWiki.

We set out to answer whether there exist significant genre differences in song lyrics (RQ1) and artist Wikipedia articles (RQ2) in terms of repetitiveness and readability. We also assess whether we can find evidence to support the cliché that lyrics of very popular artists are particularly simple and repetitive (RQ3). We further investigate whether the characteristics of popularity, length, repetitiveness, and readability correlate within and between lyrics and Wikipedia articles (RQ4).

We identify substantial differences in repetitiveness and readability of lyrics between music genres. In contrast, no significant differences between genres are found for artists' Wikipedia pages. Also, we find that lyrics of highly popular artists are repetitive but not necessarily simple in terms of readability. Furthermore, we uncover weak correlations between length of lyrics and of Wikipedia pages of the same artist, weak correlations between lyrics' reading difficulty and their length, and moderate correlations between artists' popularity and length of their lyrics.

KEYWORDS

music listening behavior; song lyrics; repetitiveness; readability; Last.fm; Wikipedia; analysis

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313604

ACM Reference Format:
Markus Schedl. 2019. Genre Differences of Song Lyrics and Artist Wikis: An Analysis of Popularity, Length, Repetitiveness, and Readability. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308558.3313604

1 INTRODUCTION

The way music is created and consumed has substantially changed over time. Various music genres and styles have evolved, prospered, and disappeared again from public knowledge and appreciation. Existing research studying differences in genre and style commonly takes a musicological perspective and analyzes acoustic properties of individual pieces, such as rhythm, dynamics, melody, harmony, and timbre. While being a valid approach, it foregoes exploiting the abundance of web-based data sources that provide information on music pieces and artists.

In contrast, we conduct a large-scale analysis of almost half a million songs by artists of a variety of genres (given by a commercial genre taxonomy and a user-generated genre folksonomy). Extracting content information from lyrics, contextual information provided in artists' Wikipedia articles, and combining these with listening information, we set out to study the characteristics and interplay of *popularity*, *length*, *repetitiveness*, and *readability* of lyrics and Wikipedia articles.

More precisely, we formulate and investigate the following research questions:
RQ1: Can we observe (and if so describe) significant genre differences in song lyrics with regard to repetitiveness and readability?
RQ2: Can we observe (and if so describe) significant genre differences in artist Wikipedia pages with regard to repetitiveness and readability?
RQ3: Is there evidence to support the cliché that lyrics of very popular artists are particularly simple and repetitive?
RQ4: Can we observe (and if so describe) significant correlations between characteristics of lyrics and Wikipedia articles, notably popularity, length, repetitiveness, and readability?

The answers to these RQs will help gaining a better understanding of the role of genre and popularity when characterizing music by content (lyrics) and context (Wikis) information. The results will further allow to ameliorate user preference models by incorporating respective characteristics, and in turn improve personalized music access systems such as information retrieval or recommendation systems. For instance, together with sociodemographic information of music listeners, e.g., provided in the LFM-1b datasets [9, 10, 12], the findings of the study at hand could

	Mean	Std.	Min.	Median	Max.
LC	1,990	4,021	5	594	48,640
PC	44,449	149,531	6	6,840	3,838,604
L_characters	1,151	535	108	1,035	4,339
L_words	229	106	23	206	869
L_CR	58.44	7.24	15.04	58.88	90.27
L_SMOG	12.59	3.04	3.00	12.47	31.93
W_characters	9,945	11,621	161	5993	111,835
W_words	1,650	1,931	27	994	18,347
W_CR	50.32	4.25	16.58	51.03	66.89
W_SMOG	11.35	1.31	5.24	11.37	27.19

Table 1: Statistical summaries of properties of song lyrics (L) and Wikipedia pages of music artist (W).

be used to create country-specific user models of lyrics preferences. Similar to previous work that showed superior performance of culture-aware music recommender systems that integrate descriptors of country-specific music mainstream [11], integrating user models of lyrics preferences could outperform methods that are unaware of this kind of information.

In the remainder of this article, we discuss related literature (Section 2), detail the data sources and data acquisition procedure we base our study upon (Section 3), describe the methodology we use to characterize lyrics and Wikipedia articles (Section 4), elaborate on our study to approach the RQs and discuss its results (Section 5), and conclude including further perspectives (Section 6).

2 RELATED WORK

Previous work on dynamics of music characteristics on a scale larger than individual pieces or artists has predominantly focused on the temporal evolution of music properties, investigating acoustic content features (e.g., rhythm, melody, harmony, and timbre). The analysis is commonly performed on datasets of several hundreds to thousands of music pieces only, and is often biased towards highly popular songs that appeared in the charts.

For instance, Schellenberg and von Scheve [13] investigate the evolution of tempo and key (major or minor) for about 1,000 top-40 songs in the charts, covering 25 years. They find a trend of increasing use of minor key (commonly perceived sadder than major key) and decrease in tempo. They also discover an increase in song duration and proportion of female artists over the years.

Mauch et al. [6] investigate 17,094 songs that appeared in the Billboard Hot 100 charts, covering the years 1960 to 2010. They compute harmonic and timbre features from the audio and apply a topic modeling approach (latent Dirichlet allocation), as a result of which each song is described as a mixture over eight chord progression clusters (topics) and eight timbral clusters. The authors report a strong and continuous decline over time of the topic reflecting dominant 7^{th} chords (often found in Jazz and Blues) and a tremendous increase of music with no clear chord structure in the early 1990s (corresponding to the rise of Hip-Hop and Rap music). This rise is also reflected in a timbral topic that characterizes music as "energetic", "speech", and "bright". Another timbral topic described as "guitar", "loud", and "energetic" reflects the predominance of Arena Rock bands in the 1980s, for instance, Kiss and Queen.

Conducting a large-scale analysis, Serrà et al. [14] examine pitch, timbre, and loudness of 464,411 music pieces recorded between 1955 and 2010. The authors uncover a tendency of simpler pitch sequences (less complex melodies), increasing loudness levels during mastering (lower volume dynamics), and homogenization of timbre (fewer variety in sound color or texture) over the course of time.

Also song lyrics have been an object of study. They are investigated, for instance, by DeWall et al. [1]. The authors consider lyrics of the 10 most popular songs on the Billboard Hot 100 year-end charts, released between 1980 and 2007. They perform several linguistic analyses and find that the use of first person singular pronouns ("I", "me", etc.) increased over time, whereas the use of first person plural pronouns ("we", "our", etc.) decreased; concluding an increase in narcissism. Furthermore, the authors identify a decrease of words related to social interactions ("sharing", "talking", etc.), an increase of words expressing anger and antisocial behavior ("hate", "kill", etc.), and a decrease of words expressing positive emotions ("love", "happy", etc.) over time.

Morris [7] investigates lyrics of 15,000 songs from the Billboard Hot 100 between 1958 and 2017. Similar to our approach, the author measures repetitiveness in terms of how well lyrics can be compressed. He identifies Rihanna, Britney Spears, and Beyoncé as the artists with the most repetitive lyrics (they can be compressed by $\geq 60\%$), whereas lyrics by Frank Sinatra, Elvis Presley, and Brad Paisley can be compressed by the least amount ($\leq 38\%$).

A study of album reviews by Amazon customers on the large scale is conducted by Oramas et al. [8], who analyze 263,525 user reviews of 65,566 albums. They investigate the evolution of sentiments expressed in reviews between 2000 and 2014, using as features degree of affectivity, fraction of positive words among all words, and emotion strength. The authors do not find a clear temporal trend or pattern in reviews with regard to sentiment, except for a peak in 2008 which they try to explain with the election of Barack Obama as US president. Categorizing the reviews according to album release year (1960 to 2014) instead of review publication year, they further discuss the evolution of the two genres Pop and Reggae. While for the former review sentiments have been very stable over the years, positive sentiments in Reggae album reviews peaked between the late 1970s and the mid 1980s.

In contrast to existing work, we take a larger variety and amount of music into account by not only focusing on popular music or chart songs, but considering songs listened to by a large web community of music aficionados, i.e., users of Last.fm. Furthermore, we jointly investigate characteristics of musical content (lyrics) and contextual information (Wikipedia articles). We also refrain from investigating the temporal evolution of these characteristics, instead we provide an analysis of their differences between music genres according to genre definitions of coarse and of fine granularity.

3 DATA ACQUISITION

We base our study on the LFM-1b dataset [9, 10], which contains roughly 1.09 billion listening events scrobbled[1] between 2005 and 2014 by more than 120,000 Last.fm users. Last.fm is one of the largest online music streaming platforms and provides various connectors

[1] Among Last.fm users it is common to use the term "scrobble" to indicate that a music piece has been listened to and to share this information with other users.

to other services, including Spotify, Pandora, iTunes, and Youtube, through which users can share what they are listening to.[2]

To analyze genre differences, we obtain **genre** information about artists using the *LFM-1b User Genre Profile* (LFM-1b UGP) dataset [12], which describes each artist by one or several genre and style annotations. In this dataset, two different dictionaries of genres/styles are considered: 20 broad (expert-defined) genres from the commercial music guide Allmusic[3] and 1,998 fine-grained (user-defined) musical style descriptors from Freebase.[4]

3.1 Acquiring Song Lyrics

For the 585,095 artists in the LFM-1b dataset, we gather the top tracks (most frequently listened to by Last.fm users as of January 2017), using the Last.fm API.[5] We then fetch the corresponding song lyrics from LyricWiki.[6] To obtain reliable results for the subsequent readability scoring, which is tailored to the English language, we next perform automatic language detection of the lyrics, using Google's language detection library,[7] which is reported to reach almost 100% precision for English using a naïve Bayes classifier trained on Wikipedia abstracts. Furthermore, to obtain significant results on the artist level, we only consider artists for which we could acquire at least 10 different song lyrics. This filtering eventually yields a dataset of 424,476 song lyrics by 18,724 artists.

3.2 Acquiring Artist Wikipedia Articles

For the 18,724 artists for which we could acquire (English) lyrics of at least 10 songs, we fetch the textual content of English Wikipedia articles (as of October 2018), using the Python wrapper Wikipedia-API.[8] To alleviate disambiguation issues and increase precision when retrieving Wikipedia articles, we use several keyword-based heuristics, e.g., identifying disambiguation pages or requiring that the first sentence contains at least one music-related term.[9] Following this approach, we eventually obtain 11,363 Wikipedia articles.

To ensure reproducibility of our study, the used datasets can be shared upon request. Please contact the author on this matter.

4 CHARACTERIZING LYRICS AND WIKIPEDIA ARTICLES

We adopt the following scoring approaches to quantify popularity, length, repetitiveness, and readability of lyrics and Wikipedia articles.

4.1 Measuring Popularity

To quantify artist popularity, we aggregate all listening events in the LFM-1b dataset for each artist and compute **play count (PC)** and **listener count (LC)** values, the former referring to the total number of listening events the respective artist attracted, the latter to the number of unique users who listened to the respective artist (at least once).

4.2 Measuring Length

To characterize an artist a's lyrics in terms of length, we compute the arithmetic mean of the number of **characters** and **words** in all song lyrics gathered for a. To describe a's Wikipedia article, we use the same two measures.

4.3 Measuring Repetitiveness

In order to approximate how repetitive a song's lyrics are, we adopt the idea that repetitiveness can be approximated by how much one can compress a text, applied to song lyrics by Morris [7]. More precisely, we use the deflation variant of the LZ77 algorithm [16] provided by the zlib library.[10] For each song, we compute the compressed length in characters of its lyrics and relate it to the uncompressed length, i.e., we compute the lyrics' **compression ratio (CR)**. We then define the repetitiveness score of an artist as the arithmetic mean of the CR of the artist's songs' lyrics in the dataset.

4.4 Measuring Readability

To assess the readability of lyrics, we use the **Simple Measure of Gobbledygook (SMOG)** [5]. It estimates the years of education required to comprehend the text under consideration and has frequently been validated empirically for English documents [2, 15]. Similar to the CR score, we define readability on the artist level as arithmetic mean of the SMOG scores of the artist's lyrics.

5 ANALYSIS AND RESULTS

A general statistical summary of popularity, length, repetitiveness, and readability scores, for both lyrics and Wikipedia articles, computed over all artists, is provided in Table 1. We observe, not surprisingly, that repetitiveness of lyrics (L_CR) is considerably higher on average than that of Wikipedia articles (W_CR), respectively, 58% vs. 50%. However, lyrics are less stable in this regard (standard deviation of 7% vs. 4%). In terms of readability, lyrics (L_SMOG) are slightly more difficult to understand than Wikipedia articles (W_SMOG), cf. SMOG score of 13 vs. 11, observing again less stability for lyrics (standard deviation of 3 vs. 1).

5.1 Genre Differences

Detailing the approach and results of our investigations to answer the RQs, we first focus on the question of genre differences in lyrics (**RQ1**) and Wikipedia articles (**RQ2**), with respect to the repetitiveness and readability scores. To this end, we illustrate in Tables 2 and 3 the means and standard deviations of CR and SMOG scores, for lyrics and Wikipedia articles, respectively, using the coarse genre taxonomy of Allmusic and the fine-grained Freebase folksonomy.

Considering the *Allmusic* genres, Table 2 reveals that RnB lyrics show the highest repetitiveness (CR of 62.83%), and they do so consistently (one of the lowest standard deviations of 5.57%). On the other hand, Heavy Metal lyrics are found to be least repetitive (CR of 57.13%), however, not consistently between this genre's artists

[2] https://www.last.fm/about/trackmymusic
[3] https://www.allmusic.com
[4] https://developers.google.com/freebase
[5] https://www.last.fm/api/show/artist.gettoptracks
[6] http://lyrics.wikia.com/Lyrics_Wiki
[7] https://code.google.com/archive/p/language-detection
[8] https://pypi.org/project/Wikipedia-API
[9] After manual inspection of hundreds of Wikipedia pages about music artists, we identified a set of terms at least one of which is used in (almost) every music artist-related page. These words include *music*, *band*, *singer*, *songwriter*, *musician*, *entertainer*, *rapper*, and *DJ*.
[10] http://www.zlib.net

		Lyrics				Wiki			
		CR		SMOG		CR		SMOG	
Genre	**Artists**	**Mean**	**Std.**	**Mean**	**Std.**	**Mean**	**Std.**	**Mean**	**Std.**
Rock	7160	58.92	6.75	12.37	2.80	50.78	3.92	11.41	1.27
Alternative	5880	59.06	6.84	12.43	2.78	50.57	3.90	11.43	1.26
Pop	5322	60.40	6.18	12.48	2.71	50.77	3.90	11.36	1.20
Electronic	2623	59.98	7.81	12.47	3.04	50.44	4.27	11.51	1.32
Folk	2372	57.89	5.94	12.33	2.56	50.56	3.58	11.57	1.20
Punk	2321	58.22	7.48	12.21	2.85	50.91	3.72	11.30	1.28
Jazz	1336	59.09	6.48	12.50	2.92	50.90	3.42	11.59	1.17
Blues	1318	59.16	5.24	12.11	2.56	51.33	3.26	11.60	1.11
RnB	1169	62.83	5.57	13.01	2.61	51.42	3.78	11.31	1.04
Rap	1057	58.36	6.76	13.65	3.20	50.61	4.36	11.27	1.23
Heavy Metal	968	57.13	7.37	12.86	3.31	51.02	4.17	11.59	1.26
Country	866	58.72	5.01	12.44	2.45	50.80	3.52	11.53	1.10
Easy Listening	769	60.59	5.01	12.53	2.50	51.21	3.42	11.44	1.04
Vocal	637	59.85	5.78	12.38	2.69	50.64	3.81	11.35	1.19
Mean		**59.30**		**12.56**		**50.85**		**11.45**	

Table 2: Compression ratio (CR) and readability score (SMOG) of lyrics and Wikipedia pages per genre, using the Allmusic genre taxonomy. Higher values of means are depicted in darker shades of green; higher values of standard deviations in darker shades of blue.

(high standard deviation of 7.37%). As for readability, understanding Rap lyrics by far requires the highest reading capabilities (SMOG of 13.65), whereas Punk and Blues lyrics are the easiest to comprehend (SMOG of ≈12). For Wikipedia articles, no substantial differences between music genres in terms of CR and SMOG are observable when using the Allmusic taxonomy.

Investigating the results obtained with the fine-grained *Freebase* genres (Table 3), we observe highest repetitiveness for lyrics in the genres Dance and Electropop (both CR of ≈63%). In contrast, lowest CR is observed for Death metal and Hardcore punk (both CR of ≈53%). The most consistent genres in terms of CR are Soft rock, Americana, and Rock and Roll (all standard deviation of CR <5%). The genres whose lyrics require the highest reading grades are Hip-Hop, Death metal, and Progressive metal (all SMOG >13). Easiest to understand are lyrics of the genres Post-rock, Psychedelic rock, and Dream pop (all SMOG <11.5). As for artist Wikipedia pages, analogous to the Allmusic taxonomy, both repetitiveness and readability scores occupy a quite narrow range: CR between 49.45 for Indie folk and 52.02 for Classic rock; SMOG between 10.99 for Pop punk and 12.25 for Avant-garde.

5.2 Differences for Popular Artists

To investigate the cliché that song lyrics of very popular artists are particularly simple and repetitive (**RQ3**), we compare the top 20 artists in the LFM-1b dataset (popularity measured in terms of LC) with all other artists in the set. Table 4 depicts the corresponding popularity, length, repetitiveness, and readability scores. It further shows the means of these scores computed over all artists in the dataset (in row "Overall mean").

We observe several pronounced differences between the top artists as well as between the top artists and all artists (overall

		Lyrics				Wiki			
		CR		SMOG		CR		SMOG	
Genre	**Artists**	**Mean**	**Std.**	**Mean**	**Std.**	**Mean**	**Std.**	**Mean**	**Std.**
Indie	4474	58.76	6.83	12.35	2.82	50.41	3.92	11.46	1.24
Alternative rock	3730	59.25	6.45	12.38	2.71	50.84	3.77	11.40	1.24
Singer-songwriter	3544	59.56	5.83	12.61	2.62	50.78	3.66	11.47	1.16
Indie rock	3184	58.30	6.65	12.18	2.73	50.28	3.88	11.43	1.25
Metal	2634	55.95	8.08	12.86	3.51	50.66	4.30	11.42	1.32
Pop rock	2263	61.83	5.55	12.84	2.57	51.20	3.84	11.30	1.20
Indie pop	2246	58.97	6.66	12.15	2.71	50.02	3.95	11.42	1.18
Experimental	1997	56.08	7.69	12.12	3.11	50.47	3.76	11.78	1.30
Classic rock	1968	60.17	5.45	12.24	2.58	52.02	2.94	11.55	1.11
Dance	1804	63.00	7.02	12.72	2.96	50.89	4.27	11.30	1.21
Hard rock	1689	59.73	6.36	12.37	2.72	51.52	3.66	11.43	1.22
Punk rock	1680	58.24	7.27	12.18	2.85	50.98	3.70	11.19	1.29
Electronica	1635	60.04	8.12	12.36	3.20	50.12	4.41	11.49	1.32
Soul	1623	61.37	5.98	12.65	2.73	50.91	3.78	11.36	1.11
Hip-Hop	1585	59.19	6.87	13.33	3.10	50.66	4.16	11.28	1.19
Emo	1546	58.93	7.33	12.28	2.84	51.04	3.92	11.23	1.26
Chill out	1504	59.13	6.91	12.51	2.87	50.39	3.93	11.53	1.23
UK 82	1440	60.31	6.97	12.31	2.85	51.38	3.47	11.63	1.18
Hardcore	1364	56.06	8.29	12.23	3.26	50.88	4.03	11.21	1.30
Folk rock	1278	57.67	5.21	12.23	2.57	50.74	3.43	11.72	1.13
Post-punk	1151	57.72	6.67	12.12	2.81	50.77	3.31	11.57	1.21
New Wave	1127	59.95	6.57	12.36	2.77	50.96	3.49	11.55	1.25
Funk	1123	60.73	6.43	12.71	2.93	51.13	3.52	11.40	1.15
Psychedelic	1108	56.74	6.87	11.56	2.69	51.16	3.40	11.96	1.24
Progressive rock	1051	57.34	6.63	12.19	2.94	51.31	3.58	11.78	1.22
Power pop	955	60.54	6.03	12.56	2.71	50.88	3.66	11.21	1.25
Synthpop	912	60.32	7.49	12.56	2.90	50.26	4.20	11.55	1.30
Electro	892	60.96	8.37	12.42	3.17	50.11	4.28	11.50	1.29
Soft rock	881	61.41	4.76	12.75	2.42	51.80	3.44	11.40	1.04
Pop punk	878	59.41	7.12	12.34	2.78	50.75	4.05	10.99	1.30
Britpop	872	60.92	6.09	12.48	2.56	51.35	3.55	11.42	1.13
Americana	853	57.53	4.75	12.36	2.37	50.69	3.44	11.58	1.10
Soundtrack	845	60.57	6.02	12.36	2.68	51.13	3.43	11.39	1.13
Progressive	815	54.76	7.60	12.15	3.35	50.84	4.01	11.86	1.33
Indie folk	750	56.70	5.70	11.93	2.51	49.45	3.75	11.58	1.16
Post-hardcore	734	55.63	8.35	11.52	3.02	50.63	4.08	11.17	1.35
Lo-fi	712	55.58	6.57	11.56	2.72	49.98	3.60	11.68	1.34
Death metal	705	53.01	8.27	13.59	4.03	50.41	4.29	11.51	1.49
Blues-rock	674	59.27	5.00	11.72	2.56	51.75	2.96	11.58	1.04
Electropop	661	62.96	7.60	12.70	2.87	50.47	4.29	11.44	1.26
Screamo	625	57.49	8.00	11.84	3.02	51.23	3.90	11.16	1.27
Christian	616	56.97	7.09	12.46	2.93	50.37	4.21	11.06	1.29
Post-rock	612	54.30	7.59	11.45	2.75	49.77	4.28	11.73	1.43
Alternative metal	607	59.38	7.34	12.67	2.96	50.93	4.48	11.37	1.36
Metalcore	604	56.10	8.30	12.24	3.54	50.75	4.27	11.13	1.33
Psychedelic rock	595	57.56	6.60	11.46	2.62	51.59	3.13	11.96	1.18
Trip hop	594	58.12	7.20	12.55	3.14	49.82	4.10	11.52	1.31
Rock and Roll	573	60.86	4.96	11.79	2.47	51.89	2.88	11.42	1.08
Dream pop	570	56.61	6.83	11.48	2.74	49.78	3.96	11.72	1.26
Grunge	565	59.42	6.24	12.28	2.74	51.35	3.22	11.37	1.15
Gothic	562	55.76	7.11	12.61	2.81	49.95	4.37	11.74	1.38
Garage rock	558	59.24	5.93	11.52	2.63	51.35	3.14	11.48	1.16
Industrial	540	57.13	7.74	12.61	3.11	50.11	4.18	11.68	1.39
Hardcore punk	521	52.81	7.95	11.66	3.24	50.44	3.92	11.08	1.39
Progressive metal	519	53.76	7.93	13.36	3.86	50.67	4.58	11.82	1.42
Avant-garde	514	53.78	7.64	12.19	3.32	50.42	3.87	12.25	1.35
Mean		**58.32**		**12.30**		**50.77**		**11.48**	

Table 3: CR and SMOG using the Freebase genre folksonomy.

means). To assess statistical significance of these differences, we conduct a t-test for equivalence of sample means, comparing the top artists with the other artists in terms of each property. The respective p-values are shown in the last row of Table 4, results significant ($p < 0.001$) highlighted. With respect to RQ3, indeed, the top artists' lyrics are significantly more repetitive (higher CR values) than others' lyrics. No significant difference can be observed, however, for readability nor length of lyrics. Differences for Wikipedia articles are highly significant for length, CR, and SMOG scores.

5.3 Correlations Between Characteristics

Concerning correlations between popularity, length, repetitiveness, and readability within and between lyrics and Wikipedia pages (**RQ4**), we compute over all artists in the dataset Spearman's rank-order correlation coefficient ρ to cope with the different value ranges of the respective measures. The correlation figures are reported in Table 5; statistically significant values at $p < 0.001$ are highlighted. We observe that, while significant, most non-obvious[11] correlations are only weak. In particular, we see that artist popularity does barely correlate with lyrics length ($0.02 < \rho < 0.09$), but does weakly to moderately correlate with Wikipedia article length ($0.33 < \rho < 0.37$). Supporting the findings from our analysis of top artists (RQ3; cf. Section 5.2 and Table 4), also over all artists in the dataset, popularity is (weakly, but significantly) positively correlated with repetitiveness; however, only when measuring popularity in terms of listener count ($\rho \approx 0.15$). Almost no correlation can be found for play count ($\rho \approx 0.05$). This might be because artists with more repetitive lyrics may appeal to a larger variety of listeners (reflected in the LC value) than those with less repetitive lyrics, whereas the latter are listened to more frequently by a smaller number of listeners (reflected in the PC value).

Readability difficulty of lyrics shows a tendency to slightly increase with their length ($0.29 < \rho < 0.32$); the same holds for Wikipedia readability and length ($0.34 < \rho < 0.36$). We exemplify this observation by considering the genres Rap and Hip-Hop, which are among those with highest SMOG scores (cf. Tables 2 and 3) and at the same time show highest lyrics length: the average length of a Rap song's lyrics is 2,099 characters or 414 words, that of a Hip-Hop song equals 1,921 characters or 381 words. In contrast, over all genres, the average length is 1,151 characters or 229 words.

Furthermore, weak cross-category correlations (between lyrics and Wikipedia pages) exist between length of lyrics and length of Wikipedia pages ($0.21 < \rho < 0.22$). It seems that artists whose songs have longer lyrics stimulate a slightly higher level of activity or dedication of Wikipedia authors than those with shorter lyrics.

6 CONCLUSIONS AND FUTURE WORK

We performed a large-scale analysis of more than 420,000 song lyrics and more than 10,000 Wikipedia pages about music artists, statistically investigating popularity, length, repetitiveness, readability, as well as their correlations. We identified substantial differences in repetitiveness and readability of lyrics between genres, most pronounced when using a fine-grained genre folksonomy. For artist Wikipedia pages, no substantial differences between genres were observable, regardless of the used genre taxonomy or folksonomy.

Evidence supporting the cliché that lyrics of very popular artists are particularly simple and repetitive could be found partially. On the one hand, compression rates of top artists' lyrics are indeed significantly higher than those of others, indicating a higher repetitiveness of those lyrics. On the other hand, this does not hold for readability. Therefore, our analysis showed that lyrics of top artists are repetitive but not necessarily simple in terms of readability.

As a matter of fact, there exist limitations of our study. In particular, community biases are likely to affect the distribution of data items (e.g., artist listening events on Last.fm or songs for which lyrics are available on LyricWiki). Last.fm users' artist and genre preferences which barely generalize to the population at large represent another form of community bias [4] our approach is prone to. Also the distribution of demographics of the involved web services' users is unlikely to correspond to that of the general population [3]. Being aware of these limitations (which we plan to reduce in future work), we nevertheless believe that our findings will enable a better understanding of the role of genre and popularity when characterizing music by content (lyrics) and context (Wikis) information. Integrated into user models, the investigated lyrics and Wiki characteristics may also improve personalized music retrieval and recommendation systems.

Future work will include investigating additional data sources such as artist fan pages or album reviews. Also, while we currently only use the (text) content of Wikipedia's artist pages, metadata about the pages could be included in the investigation, e.g., the number of revisions or contributors to the Wikipedia page. In addition, further features could be defined and analyzed, for instance related to emotion or sentiment expressed in lyrics, Wikis, reviews, etc. Another avenue for further studies is to consider user-specific characteristics, including age, gender, culture, education, knowledge, or personality, and to investigate to which extent the preference for certain categories of songs or artists (described by content and context features) is reflected in such user characteristics.

[11] Not surprisingly, popularity measures (PC and LC) as well as length measures (characters and words) strongly correlate. Medium correlations exist between compression ratio and length of lyrics; strong to very strong between CR and length of Wikipedia pages.

REFERENCES

[1] C. Nathan DeWall, Richard S. Pond Jr., W. Keith Campbell, and Jean M. Twenge. 2011. Tuning in to psychological change: Linguistic markers of psychological traits and emotions over time in popular U.S. song lyrics. Psychology of Aesthetics, Creativity, and the Arts 5, 3 (March 2011), 200–207. http://dx.doi.org/10.1037/a0023195

[2] Cecilia Conrath Doak, Leonard G. Doak, and Jane H. Root. 1996. Teaching Patients with Low Literacy Skills. Lippincott Philadelphia, USA. 36–59 pages.

[3] Thomas Krismayer, Markus Schedl, Peter Knees, and Rick Rabiser. 2018. Predicting User Demographics from Music Listening Information. Multimedia Tools and Applications (May 2018). https://doi.org/10.1007/s11042-018-5980-y

[4] Paul Lamere. 2008. Social Tagging and Music Information Retrieval. Journal of New Music Research: Special Issue: From Genres to Tags – Music Information Retrieval in the Age of Social Tagging 37, 2 (2008), 101–114.

[5] G. Harry Mc Laughlin. 1969. SMOG Grading—A New Readability Formula. Journal of Reading 12, 8 (1969), 639–646. http://www.jstor.org/stable/40011226

[6] Matthias Mauch, Robert M. MacCallum, Mark Levy, and Armand M. Leroi. 2015. The Evolution of Popular Music: USA 1960–2010. Royal Society Open Science 2, 5 (2015). https://doi.org/10.1098/rsos.150081

[7] Colin Morris. 2017. Are Pop Lyrics Getting More Repetitive? (2017). https://pudding.cool/2017/05/song-repetition (accessed: October 2018).

[8] Sergio Oramas, L. Espinosa-Anke, A. Lawlor, Xavier Serra, and H. Saggion. 2016. Exploring Customer Reviews for Music Genre Classification and Evolutionary Studies. In Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR 2016). New York, NY, USA.

Artist	LC	PC	L_chars	L_words	L_CR	L_SMOG	W_chars	W_words	W_CR	W_SMOG
Coldplay	48,640	2,576,390	922	187	62	11	41,942	7,006	56	12
Radiohead	44,707	3,437,326	786	154	59	13	44,541	7,146	56	14
Daft Punk	44,356	2,523,537	2,249	444	**81**	14	38,466	6,346	55	12
Nirvana	41,488	1,878,647	994	206	71	11	34,288	5,625	56	13
Red Hot Chili Peppers	40,421	2,221,660	1,686	336	67	14	79,842	13,460	56	12
Queen	39,874	1,614,548	1,298	265	62	11	68,061	11,463	56	13
Muse	39,726	2,460,597	841	163	65	16	27,692	4,493	55	12
The Rolling Stones	39,033	1,483,385	1,262	255	65	11	78,327	13,272	55	11
Rihanna	38,692	1,313,643	2,000	414	73	15	63,397	10,391	56	13
Adele	38,345	1,386,940	1,463	302	66	13	43,257	7,514	56	11
David Bowie	37,340	1,685,010	1,351	263	63	12	84,252	13,874	53	13
Foo Fighters	35,259	1,444,212	1,289	259	70	13	35,963	6,087	56	12
The Killers	35,199	1,367,039	1,236	252	63	13	30,002	5,044	56	12
Michael Jackson	34,817	958,130	2,390	486	72	14	**95,760**	**15,585**	55	13
The Beatles	34,778	**3,838,604**	1,104	223	67	13	92,065	14,747	54	**15**
Eminem	34,427	1,445,767	**4,339**	**869**	60	**17**	62,174	10,404	54	13
Florence + the Machine	34,208	1,729,489	1,449	299	69	12	23,333	3,950	56	12
Gorillaz	34,157	1,354,351	1,250	246	64	11	24,947	4,211	55	13
Pink Floyd	33,834	2,990,318	878	170	49	9	70,794	11,633	55	13
Linkin Park	33,672	2,296,327	1,568	317	71	**17**	37,141	6,160	**57**	12
Overall mean	2,000	44,961	1,151	229	58	13	9,954	1,652	50	11
***p*-value of t-test**	0*	0*	10^{-3}	10^{-3}	10^{-6}*	10^{-1}	10^{-62}*	10^{-62}*	10^{-7}*	10^{-4}*

Table 4: Properties of lyrics (L) and Wikipedia articles (W) for top artists in terms of listener count (LC). Numbers printed in bold are the maxima in each column; those in gray are the minima. The last but one row contains the overall means for *all* analyzed artists. The last row reports the (rounded) p-values of a t-test for equivalence of sample means between the top artists and others; * denotes significance at $p < 0.001$.

	LC	PC	L_chars	L_words	L_CR	L_SMOG	W_chars	W_words	W_CR	W_SMOG
LC	1.000*	0.955*	0.091*	0.093*	0.145*	-0.033*	0.369*	0.369*	0.302*	0.137*
PC		1.000*	0.030	0.023	0.054*	-0.026	0.334*	0.331*	0.300*	0.164*
L_chars			1.000*	0.993*	0.506*	0.315*	0.214*	0.220*	0.188*	-0.032*
L_words				1.000*	0.535*	0.293*	0.217*	0.223*	0.186*	-0.044*
L_CR					1.000*	0.102*	0.230*	0.238*	0.199*	-0.058*
L_SMOG						1.000*	0.000	-0.000	0.016	-0.001
W_chars							1.000*	0.999*	0.834*	0.357*
W_words								1.000*	0.833*	0.344*
W_CR									1.000*	0.266*
W_SMOG										1.000*

Table 5: Spearman's rank-order correlations between lyrics (L) and Wikipedia pages (W) characteristics; * denotes significance at $p < 0.001$.

[9] Markus Schedl. 2016. The LFM-1b Dataset for Music Retrieval and Recommendation. In Proceedings of the 6th ACM International Conference on Multimedia Retrieval (ICMR 2016). ACM, New York, NY, USA, 103–110. https://doi.org/10.1145/2911996.2912004

[10] Markus Schedl. 2017. Investigating country-specific music preferences and music recommendation algorithms with the LFM-1b dataset. International Journal of Multimedia Information Retrieval 6, 1 (2017), 71–84. https://doi.org/10.1007/s13735-017-0118-y

[11] Markus Schedl and Christine Bauer. 2017. Introducing Global and Regional Mainstreaminess for Improving Personalized Music Recommendation. In Proceedings of the 15th International Conference on Advances in Mobile Computing & Multimedia (MoMM 2017). ACM, Salzburg, Austria, 74–81. https://doi.org/10.1145/3151848.3151849

[12] Markus Schedl and Bruce Ferwerda. 2017. Large-Scale Analysis of Group-Specific Music Genre Taste from Collaborative Tags. In Proceedings of the 2017 IEEE International Symposium on Multimedia (ISM). IEEE, Taichung, Taiwan, 479–482. https://doi.org/10.1109/ISM.2017.95

[13] E. Glenn Schellenberg and Christian von Scheve. 2012. Emotional cues in American popular music: Five decades of the Top 40. Psychology of Aesthetics, Creativity, and the Arts 6, 3 (August 2012), 196–203. http://dx.doi.org/10.1037/a0028024

[14] Joan Serrà, A. Corral, M. Boguñá, M. Haro, and Josep Lluis Arcos. 2012. Measuring the Evolution of Contemporary Western Popular Music. Scientific Reports 2 (July 2012). https://doi.org/10.1038/srep00521

[15] Kevin Wong and Jessica R. Levi. 2017. Readability Trends of Online Information by the American Academy of Otolaryngology – Head and Neck Surgery Foundation. Journal of Otolaryngology – Head & Neck Surgery 156, 1 (2017), 96–102. https://doi.org/10.1177/0194599816674711

[16] Jacob Ziv and Abraham Lempel. 1977. A universal algorithm for sequential data compression. IEEE Transactions on Information Theory 23, 3 (May 1977), 337–343. https://doi.org/10.1109/TIT.1977.1055714