A Trend Analysis on Concreteness of Popular Song Lyrics

Kahyun Choi
Indiana University Bloomington
choika@iu.edu

J. Stephen Downie
University of Illinois
jdownie@illinois.edu

ABSTRACT

Recently, music complexity has drawn attention from researchers in Music Digital Libraries area. In particular, computational methods to measure music complexity have been studied to provide better music services in large-scale music digital libraries. However, the majority of music complexity research has focused on audio-related facets of music, while song lyrics have been rarely considered. Based on the observation that most popular songs contain lyrics, whose different levels of complexity contribute to the overall music complexity, this paper investigates song lyric complexity and how it might be measured computationally. In particular, this paper examines the concreteness of song lyrics using trend analysis. Our analysis on the popular songs indicates that concreteness of popular song lyrics fell from the middle of the 1960s until the 1990s and rose after that. The advent of Hip-Hop/Rap and the number of words in song lyrics are highly correlated with the rise in concreteness after the early 1990s.

CCS CONCEPTS

• **Information systems → Content analysis and feature selection**.

KEYWORDS

trend analysis, song lyrics, concreteness of words, text complexity, readability

ACM Reference Format:
Kahyun Choi and J. Stephen Downie. 2019. A Trend Analysis on Concreteness of Popular Song Lyrics. In *6th International Conference on Digital Libraries for Musicology (DLfM '19), November 9, 2019, The Hague, Netherlands.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3358664.3358673

1 INTRODUCTION

Automatically annotating digital music with appropriate metadata has been a significant topic in Music Digital Libraries research [27]. Research on automatic music annotation has aimed at extracting various types of music descriptors, such as topic, language, mood, genre, and complexity. Among various metadata types, this paper particularly pays attention to lyric complexity.

Research on music complexity of digital music collections began in the mid 2000s, and has focused on computation methods of measuring complexity of audio music[15, 18, 21, 22, 30, 39, 43–45, 52, 53, 58, 59, 62, 63]. Complexity measures were developed as features or metadata used to better describe music, resulting in better music digital libraries. Rather than taking a holistic approach, researchers broke down music into major facets, and focused on each facet's complexity. In particular, tonal, rhythmic, and timbral facets have been explored in terms of audio music complexity. However, lyrics have been mostly excluded in the research, despite the facts that most popular songs have lyrics and their complexity influences overall music complexity. This paper aims at filling this gap by focusing on the complexity of popular song lyrics.

Although the claim may be somewhat controversial, song lyrics can be considered literature [50]. The fact that Bob Dylan was awarded the Nobel prize in literature for his lyrics in 2016[1] and both Leonard Cohen and Chuck Barry won the PEN New England Literary Excellence Award for their lyrics in 2014 support the claim [3]. Among many different forms of literature, song lyrics are usually considered similar to poems [50] because various poetic devices such as rhyme, repetition, metaphor, and imagery also occur in song lyrics. These unique genres of literature are usually written in verse rather than in prose, and are much shorter than the other genres, including short stories or fables.

When exploring quantitative methods to measure the complexity of song lyrics, it is natural to apply methods that measure literary complexity. When it comes to computationally measure literary complexity, readability formulas (or text complexity metrics) are generally used. Readability formulas have been developed and explored for almost a century; over 200 have been developed, and more than a thousand papers have been published about them [17]. Metrics have developed for various groups of people: children and adults, military personnel and civilians, readers and writers, etc. School teachers can use such readability tools to provide appropriate reading materials to their students, so that students are not frustrated by overly complex books or bored with simple texts [28, 51]. Moreover, adult literacy studies cover the readability of written material in various situations, such as the ability of military personnel to read and understand critical military manuals [28], or the comprehension of medical patient education materials [14]. Writers can also use websites that provide the readability formulas, such as *https://readable.io/* and *http://www.readabilityformulas.com/* to write clear documents and create readable websites. For example, according to *similarweb.com*, as of August, 2018, *https://readable.io/* and *http://www.readabilityformulas.com/* have 12,789,000 and 74,000 hits per month, respectively.

Since 2010, the Common Core State Standards (CCSS), which defines what K-12 students need to study in language arts and mathematics, has embraced readability studies [20]. Given that more than 40 states in the U.S. have adopted the standard[2], it is the

DLfM '19, November 9, 2019, The Hague, Netherlands

ACM ISBN 978-1-4503-7239-8/19/11…$15.00
https://doi.org/10.1145/3358664.3358673

[1] *http://www.nobelprize.org*
[2] http://www.corestandards.org/standards-in-your-state/

most widely used guidelines on how to measure text complexity. According to CCSS, text complexity of K-12 textbooks declined over the last half century, although students will be required to read much more complex text after graduation. To address the gap, CCSS emphasizes text complexity and provides guidelines for choosing appropriate textbooks.

CCSS also provides information on the latest text complexity metrics.[3] It includes one public domain readability metric, Flesch-Kincaid [33], and six representative text complexity tools: ATOS™ by Renaissance Learning [49], Degrees of Reading Power® by Questar Assessment, Inc.[4], The Lexile® Framework for Reading by MetaMetrics [40], Reading Maturity by Pearson Education[5], SourceRater by Educational Testing Service [56], and easability indicator by Coh-Metrix [25]. Each of these uses a wide range of features to measure text difficulty, and their word-level variables, described below, can be directly extracted from song lyrics in the form of verse.

- **Word Frequency:** Word frequency has long been considered as an indicator of word difficulty. It has been proven that readers tend to comprehend frequently used words quickly and easily, and therefore texts with many frequently used words tend to be easier [9, 35]. The earliest word-frequency list for English teachers, created by Edward Thorndike in 1922, covers only 10,000 words as he had to manually count the frequency of words [60]. However, recent text complexity metrics have much larger vocabulary-frequency lists. For instance, Lexile® employs about 600 million words. As of 2014, ATOS™ includes more than 2.5 billion words from more than 170,000 books [40, 49]. LNS, the lyric text complexity measure proposed by Ellis et al. is related to this variable, however it is based on inverse document frequency instead of term frequency [18].
- **Word Length:** Word length includes basic statistics such as average and standard deviation of not only the number of characters but also the number of syllables in a word. The number of syllables has been an important variable of readability formulas, including the Flesch Reading Ease and the Gunning Fog formula, since the beginning of readability studies [19, 23, 26, 46]. One of ways to calculate the number of syllables of words is using the Carnegie Mellon University pronouncing dictionary. This open-source pronunciation dictionary contains over 134,000 North American English words and is currently available online (http://www.speech.cs.cmu.edu/cgi-bin/cmudict). Since each vowel can be identified with a numeric stress marker (0, 1, or 2), the number of vowels equals to the number of syllables.
- **Word Familiarity:** Word familiarity has been an important variable of many readability metrics since the beginning of readability studies. For example, Thorndike's 1921 word list includes thousands of graded words based on their frequency, and it claimed that "by its use teachers [could] tell how familiar words are likely to be to children" [12, 60]. Recent studies still use word frequency to derive word familiarity [41], however word familiarity can be also obtained directly through user studies [12]. For instance, Coh-Metrix includes a psycholinguistic database that includes word familiarity scores of thousands of words from adult subjects [24].
- **Word Grade Level** The word grade level feature used by ATOS™is called the Graded Vocabulary List. It is an extensive word list that incorporates previously developed graded word lists, word lists of standard school exams, and others [49]. When a discrepancy is identified while merging the existing lists, the latest source takes priority. This list assumes that each word belongs to a certain grade level, and it is validated by comparing sample words to words used on five major standardized tests. Although they assigned different grade levels to different derivative forms of words, they assigned the same grade levels to homographs (different meanings of the same word). For example, *hear* is defined as a 1st grade word while *hearings* is defined as a 6th grade word. However, *wick* is listed as a 3rd grade level word, although 3rd grade students cannot understand some of its meanings.
- **Pearson Word Maturity Metric** The Pearson Word Maturity Metric takes a drastically different approach to calculate word difficulty. Unlike ATOS™'s Graded Vocabulary List, it uses a degree grade instead of a scalar grade of word understanding [34]. Also, it assigns different degrees of word knowledge to homographs as it is based on semantic analysis. This metric totally relies on how to select training sets for each grade level, and how to compare word vectors from training sets and reference models. Compared to manually generated graded vocabulary lists, this machine learning based approach is scalable and automatic. However, more research is needed before this model can replace the manual vocabulary lists [34].
- **Concreteness:** Concreteness ratings are used by SourceRator and Coh-Metrix [24, 56]. Concreteness of a word refers to whether the word is concrete or abstract. Concrete words denote objects one can experience directly through your senses or actions, while abstract words describe ideas and other non-physical concepts. For example, *couch* is a concrete word that refers to an object that you can see and touch, while *justice* is an abstract word [8]. It has been found that readability and word concreteness correlate with each other [51] when their relationship was tested to prose.

The Word-level variables are of interest to this paper, since it is not clear where sentences in music lyrics begin and end. Among the most popular word-level variables used by the complexity metrics chosen by CCSS, we focus on concreteness to examine text difficulty of popular song lyrics for the following reasons. First, no previous study explored concreteness of song lyrics. Second, concreteness is closely related not only to text difficulty, but also imageability and memorability. Given that song lyrics are often full of images and usually memorized by listeners, the findings in this research can be re-purposed to explore imageability and memorability of song lyrics in the future. Third, concreteness ratings are publicly available. Conversely, most of the other variables, such as word familiarity

[3] http://www.corestandards.org/wp-content/uploads/Appendix-A-New-Research-on-Text-Complexity.pdf
[4] http://www.questarai.com/assessments/district-literacy-assessments/degrees-of-reading-power/
[5] http://www.readingmaturity.com/

and Pearson Word Maturity Metric, do not have associated, non-proprietary data. Finally, this research will also expand research on literature concreteness by incorporating song lyrics, which was excluded in the recent research on the concreteness of large-scale book corpora [29].

Before this paper, Ellis et al. [18] took the text complexity approach to lyric complexity, and introduced the Lexical Novelty Score (LNS) as a measure of difficulty of song lyrics, based on word frequency, one of the word-level variables of traditional readability measures. LNS assumes words appearing infrequently in a large text corpus tend to make song lyrics more complex to understand. The main advantages to using LNS over word frequency scores from readability formulas is that LNS is solely derived from a corpus of spoken language, as lyrics are closer to spoken language than written language. In particular, they use the SUBTLEXus corpus, a collection of subtitle transcripts of movie and TV programs [7]. Word frequency information from both modern and traditional readability formulas are primarily derived from written corpora [9], although Coh-Metrix, one of the modern readability metrics, also exploited spoken sources, including the BBC World Service and taped telephone conversations [25].

Dodds et al.[16] demonstrated quantitative trends of song lyrics by investigating historical changes in how song lyrics portray happiness from 1960 through 2007. Their large-scale study analyzed the lyrics of 232,574 songs composed by 20,025 artists, although many songs were excluded if there were not enough matching words to ANEW, which is a list of affective scores of English norms [6]. The study revealed a clear downward trend of the happiness over time. Further analyses disclosed how frequencies of positive words have decreased while those of negative words have increased. In addition, trend analyses of individual genres showed that the valence scores of each genre is mostly stable over time and genres with low valence values, such as metal, punk, and rap, appear later.

2 CONCRETENESS OF SONG LYRICS

This research explores concreteness of song lyrics. Concrete words are those that refer to specific objects or remind of a particular situation. Abstract words, on the other hand, need other words to generate meaning. For instance, *table* is a highly concrete term because people know what a table looks like and can be reminded of a certain image. *Justice* is a highly abstract word because one cannot feel it with any of the five senses, but can understand through examples of situations. Texts composed of more concrete than abstract words have a variety of cognitive benefits: they tend to be more easily comprehended and retrieved; they tend to be more interesting than texts with more abstract words; and they tend to be imaginable [1, 29, 47, 54, 55]. For these characteristics, word concreteness has been one of the most important criteria for analyzing text difficulty [47, 51, 56].

When exploring concreteness of song lyrics, this paper also analyzes historical trends in concreteness of song lyrics, and it is the first attempt on song lyrics while concreteness of books and how it has changed over time have been both explored in order to determine whether concreteness of the English language has increased over time. Hills et al.[29] conducted trend analysis of concreteness of four collections of English books and speeches (e.g. the Google Ngrams corpus of American English [48]; the Corpus of Historical American English [13]; and inaugural addresses by American presidents). Like this paper, concreteness ratings of English word norms are obtained from the collection generated by Brysbaert et al. [8] and the concreteness values for each year were calculated by averaging concreteness values of all words appeared in books released on that year with frequencies of the words also considered. They reported that English has been getting more concrete in the datasets over the last 200 years, from 1800 to 2000, which implies that books are getting easier to read and learn from. This is partially because the proportion of closed word classes, such as articles and determiners, which have lower concreteness values than open word classes, have increased. However, concreteness scores of words within open word classes, including nouns and verbs, have increased, contributing to the upward trend of English words concreteness.

3 RESEARCH QUESTIONS

This paper analyzes the concreteness of 5,100 popular song lyrics to seek to answer research question: "How has text complexity of popular song lyrics changed over time in terms of concreteness?" To better understand the trends, we also aim to answer another research questions: "What is the relationship between the concreteness trends and genres?" and "What is the relationship between the concreteness trends and word statistics in song lyrics?"

4 EXPERIMENT DESIGN

4.1 Data

4.1.1 Music Collection. We analyze the lyrics of 5,100 songs from Billboard Year End hot 100 songs from between 1965 and 2015, which is publicly available from *billboard.com*. In the past, the Billboard Year End chart was calculated based on sales and radio airplay information. Recently, streaming information is also taken into account. The songs in the chart represent the most popular songs over 51 years, as Billboard chart is one of the most reliable sources for popular music in the U.S. For the same reason, previous popular music studies have used Billboard charts to identify trends on popular music [10, 36, 37].

4.1.2 Lyrics. We obtained a reliable lyric corpus from LyricFind[6] which is a world-wide lyric licensing company, via a signed research agreement. Utilizing this lyrics dataset has many advantages over other ones. Compared to crawled lyrics from websites, lyrics in this corpus are clean because they are for commercial services. Unlike Ellis et al.'s bag-of-words corpus [18], these are also intact, so grammatical information of each word is available. So far, three studies in MIR used this corpus: Ellis et. al measured lexical novelty of lyrics from the bag-of-words representation [18]; Atherton and Kaneshiro analyzed lyrical influence networks by using intact lyrics [2]; and Tsaptsinos proposed an automatic music genre classification system by applying recurrent neural network models to song lyrics [61].

4.1.3 Metadata. We examine relationships between concreteness of song lyrics and three types of metadata, including year, artist, and

[6]The author thanks Roy Hennig, Director of Sales at LyricFind, for kindly granting the access to their lyric database for our academic research.

genre. The year and artist metadata were taken from the Billboard Year End chart. Genre metadata was collected by using iTunes Search API[7], which returns a JSON file with a variety of metadata. Among them, *PrimaryGenreName* value was taken as a genre value. 150 songs have *unknown* genre, and the rest of the songs have one primary genre value each. The most popular genres containing at least 10 songs in the dataset are:

- **Major genres:** *Pop, Rock, R&B/Soul, Hip-Hop/Rap, Country, Dance, Alternative, Soundtrack, Electronic, Singer/Songwriter, Reggae, Jazz, Christian & Gospel, and Vocal*

The rest of genres on the long tail are:

- **Minor genres:** *House, Classical, Pop Latino, Hip-Hop, Rap, Blues, Disco, Easy Listening, Funk, Alternative Folk, Folk-Rock, Latin, Latino, New Age, Urbano Latino, World, Adult Alternative, American Trad Rock, Americana, Blues-Rock, Brazilian, British Invasion, Childrens Music, Crossover Jazz, Folk, Gangsta Rap, German Folk, Halloween, Heavy Metal, Lounge, Metal, Pop/Rock, Psychedelic, Punk, and Soul*

4.1.4 Concreteness Ratings. Brysbaert et al.[8] collected and published large-scale, crowdsourced concreteness scores of English norms. The initial word list with 60,099 English words and 2,940 two-word expressions was built mainly based on the SUBTLEX-US corpus [7] and augmented by various widely known corpora, such as the English Lexicon Project [4] and the British Lexicon Project [31]. Since song lyrics are usually closer to the spoken language than written language, it is advantageous to use this corpus, whose majority of words come from sources with spoken language. Survey participants on Amazon Mechanical Turk rated concreteness value of a word with a 1-5 point scale, and they also reported whether they knew the word well. After removing words that many people checked *word not known*, 37,058 words and 2,896 two-word expressions remained. This concreteness rating list is big enough to cover 83 % of the unique words in the song lyrics used in this work.

4.2 Lyric Preprocessing

After retrieving song lyrics from the LyricFind corpus using titles and artists, the state-of-the-art technology, Stanford CoreNLP [42] was used to tokenize them. The tool was also used to lemmatize them because the words in concreteness ratings are English lemmas. Part-of-speech tagging was also done to further analyze the concreteness trend in terms of each part-of-speech tag. As a result, 37,856 unique words were extracted from the 5,100 songs in the dataset.

4.3 Analysis Methods

4.3.1 Overall Concreteness Score. The concreteness score of individual song lyrics, denoted by v_{text}, is the weighted average of the concreteness of each word in each song lyrics where v_k is the concreteness of k-th word and f_k is its frequency.

$$v_{text} = \frac{\sum_k v_k f_k}{\sum_k f_k}. \quad (1)$$

The concreteness score for each year is the average concreteness scores of lyrics appeared in the chart of the year. Figure 1 shows how the overall concreteness score is calculated, using lyrics from an *Eminem* song as example.

[7] http://apple.co/1qHOryr

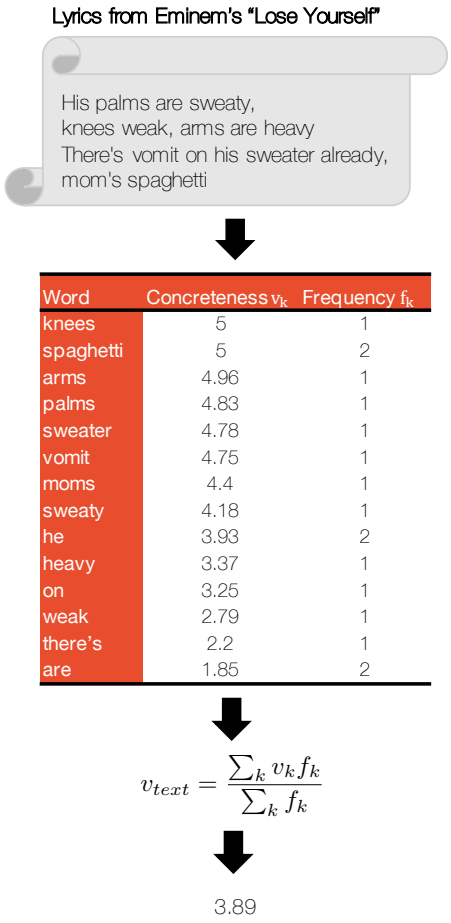

Figure 1: A pictorial example of how the overall concreteness is calculated, using an excerpt of the lyrics from Eminem's "*Lose Yourself*"

4.3.2 Trend Analysis Methods. In order to identify any trends in concreteness scores over time, this research uses scatter plots, change point analyses, and Cox-Stuart sign test [11]. Scatter plots are used to identify rough trends. To provide better visualization for long-term analysis, smoothed lines obtained from a moving average filter with a five-year span are also reported. Although a scatter plot is a helpful tool to analyze a general trend with our naked eyes, more systemic methods are required to identify the level and significance of changes. For this reason, an algorithm that detects change points finds those that divide sections with different degrees and directions of slope. Subsequently, Cox-Stuart sign test is applied to determine whether trends are statistically significant.

To identify change points, *findchangepts* is employed, which is implemented in *Matlab* [38]. As we are interested in the slope of the data, linear regression has been chosen as a statistical property for the detection algorithm. The search method of *findchangepts* is binary segmentation, which is the most established one [32].

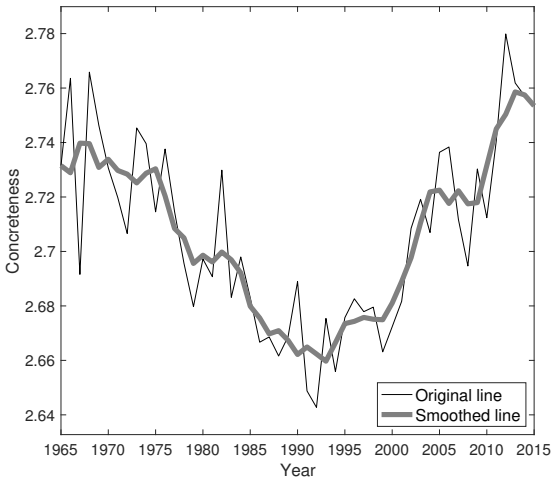

Figure 2: Concreteness time series for song lyrics

For each point, it divides data into two sections and calculates the residual errors. A change point minimizes the total residual error.

In order to determine the significance of each trend, we used Cox-Stuart sign test with 95% confidence level. This simple test has been widely used to see various trends (e.g., the topics of developers' interest [5], vitamin and mineral intake from fortified food[57], etc.). This trend test divides the observation vector into two vectors, and counts the numbers of positive and negative differences between the two vectors. More positive differences than negative ones means an increasing trend, and the opposite means a decreasing. P-value is measured based on the binomial distribution.

5 RESULTS

5.1 General Trend

Figure 2 shows how concreteness scores of pop song lyrics have changed over the last 50 years. There is a clear downward trend until the early 1990s and an upward trend afterward. The change point is 1991, and both of the trends are statistically significant. The thin line indicates the averaged annual concreteness scores, and the thick line shows its smoothed version by passing a 5-point moving average filter. The highest concreteness value of 2.78 is observed in 2012, and the lowest concreteness value of 2.64 is observed in 1992. The difference between the two points is 0.14. Given that the gaps between the maximum and minimum concreteness scores of books over the last 200 years range from 0.1 and 0.2 [29], 0.14 is quite a big difference in a much shorter period of time. Various factors may have influenced the trends of concreteness scores of song lyrics. Among the many different factors that could influence concreteness, we focused on three: proportion of genres, proportion of open/closed class words, and length of music lyrics.

5.2 Genre and Trend

Lyrics in conjunction with audio are widely used to automatically classify genres of popular songs because each genre has relatively unique lyrical characteristics. To examine how concreteness of each genre is different from each other and how the difference may influence overall concreteness trends, we calculated counts and average concreteness scores of individual genre categories. Table 1 shows the frequencies of the four main genres; *Pop* accounts for 35% of the Billboard collection, followed by *Rock* at 19%, *R&B/Soul* at 16%, and *Hip-Hop/Rap* at 9%. The average concreteness of *Pop*, 2.68, is the lowest among the major genres, while that of *Hip-Hop/Rap*, 2.79, is the highest, which is higher than the average concreteness scores of the collection. Lyrics of *Rock* and *R&B/Soul* turned out to be slightly abstract because their average concreteness scores, 2.70, are lower than the overall average, 2.72. Table 2 shows the genre distribution of songs in minor genres and their concreteness scores. The average concreteness score of *Reggae*, 2.83, is the highest among all minor genres, followed by *Jazz*, 2.77. However, they hardly contribute to the overall trend because they only account for 0.4% and 0.3%, respectively. *Country* song lyrics also recorded a relatively high average concreteness score, 2.76, and they account for 5%. The two lowest average concreteness scores are *Alternative* (2.66, 3% of the collection) and *Vocal* (2.67, 0.2%). *Dance* accounts for 3% and its concreteness value is the same as *Pop*, 2.68.

Genre	Group Count	Average Concreteness
Hip-Hop/Rap	457	2.79
Others	912	2.72
R&B/Soul	797	2.70
Rock	939	2.70
Pop	1745	2.68
All	4850	2.72

Table 1: Average concreteness scores of major music genres along with group counts

Genre	Group Count	Average Concreteness
Reggae	17	2.83
Jazz	16	2.77
Country	255	2.76
Christian& Gospel	14	2.72
Singer/Songwriter	19	2.70
Electronic	20	2.69
Dance	163	2.68
Vocal	11	2.67
Alternative	119	2.66

Table 2: Average concreteness scores of minor music genres along with group counts

For further analysis of the relationship between genres and concreteness over time, we show the time series of concreteness scores of some dominant genres (see in Figure 3). Overall, all genres except *Rock* follow the similar trends to the entire collection, showing a V-shaped curve. On the other hand, concreteness scores of *Rock* has a statistically significant downward trend ($p < 0.001$). To measure impact of each genre to the overall collection over time, the historic proportion distribution of the major genres is also calculated, as shown in Figure 4. The noticeable change is the advent of *Hip-Hop/Rap* in the late 1980s, which *Hip-Hop/Rap* eventually overtook

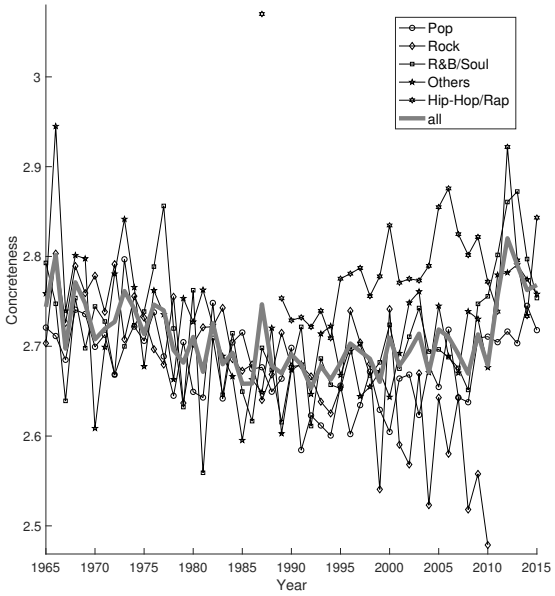

Figure 3: Concreteness time series for song lyrics broken down by major genres

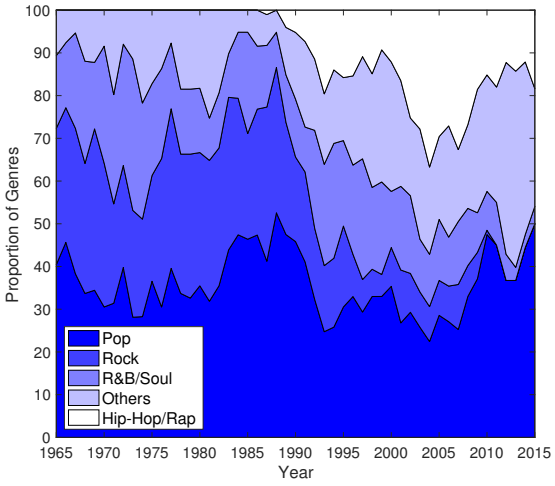

Figure 4: A stacked area plot of proportion of major genres

R&B/Soul and *Rock*. *Rock*'s share, in fact, decreased continuously until it disappeared from the chart in the early 2010s. However, when each concreteness trend without each major genre is computed (see Figure 5), the trends show the same pattern, which indicates that other factors in addition to the emergence of *Hip-Hop/Rap* influenced the upward concreteness trend in the last two decades.

5.3 Closed/Open Word Classes

Changes of shares of closed/open word classes over time may also influence the concreteness trend. In Table 3, Hills et al. [29] reported the average concreteness values and standard deviations of 11 word

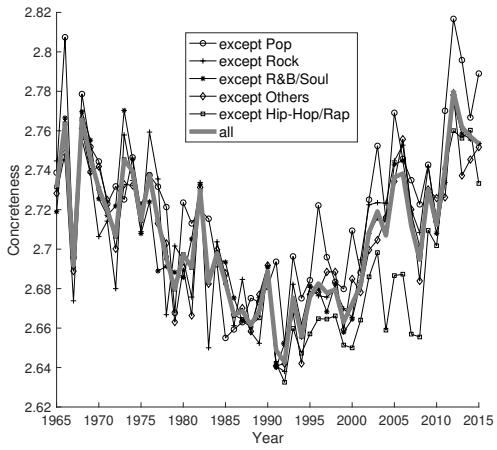

Figure 5: Concreteness time series for song lyrics except each major genre

Word class	Mean of Concreteness	Standard Deviation of Concreteness
Names	3.73	0.86
Nouns	3.53	1.02
Numbers	3.49	0.38
Verbs	2.92	0.76
Pronouns	2.76	0.71
Adjectives	2.50	0.72
Prepositions	2.29	0.64
Determiners	2.11	0.55
Adverbs	2.06	0.53
Articles	1.66	0.54
Conjunctions	1.64	0.54

Table 3: Concreteness for a selection of word classes in the Brysbaert et al. (Hills et al. [29])

classes in the concreteness ratings from the data collection publicized by Brysbaert et al. [8]. Words in Name category, such as *mayo* and *coffeecake*, have the highest concreteness value, 3.73, while conjunctions such as *before* and *if* have the lowest concreteness value, 1.64. Among 11 word classes, six of them belong to open word classes: names, nouns, numbers, verbs, adjectives, and adverbs. They have relatively higher concreteness scores than closed class words, such as conjunctions, articles, determiners, prepositions, and pronouns. If artists use more words from the open rather than the closed word classes when writing song lyrics and choose more concrete words within some word classes, it can lead to a rise in overall concreteness scores.

Figure 6 shows that the proportion of the open word classes had decreased until mid 1990s, it went upward afterwards and downward in 2010. Although the change points are 1996 and 2013, the trend between 2013 and 2015 is not statistically significant as the other two. The trend of overall concreteness values of song

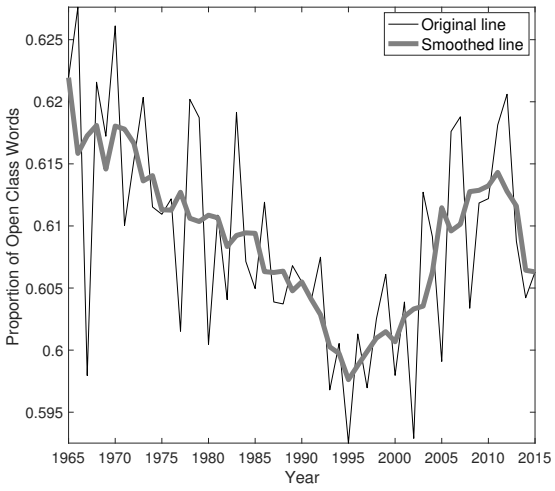

Figure 6: A portion of open word classes

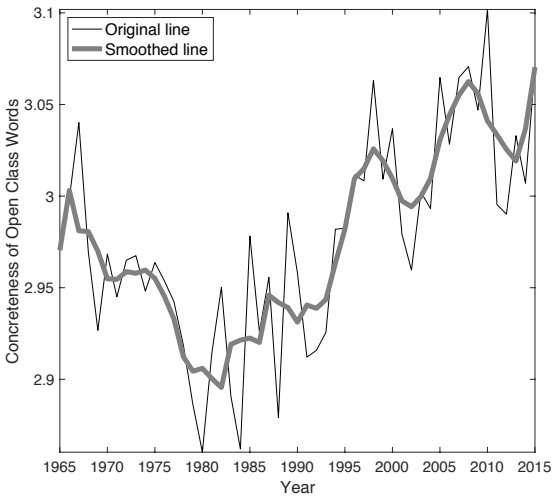

Figure 7: Concreteness time series for open word classes in song lyrics

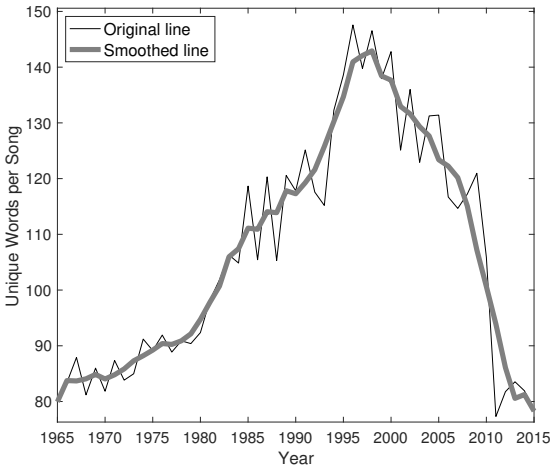

Figure 8: Average number of unique words in song lyrics

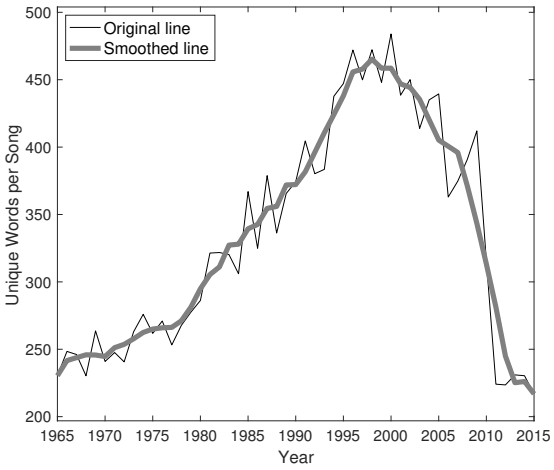

Figure 9: Average number of words in song lyrics

lyrics is highly correlated with the trend of proportions of the open word classes ($r = 0.5141$, $p = 0.001$). The concreteness scores over time within the open word classes also tightly correlate with the overall concreteness trend shown in Figure 7 ($r = 0.6105$, $p < 0.001$). The difference between the minimum and maximum concreteness values within the open word classes is even higher than the overall difference. If only the open word classes are considered, the change point is 1979, and the upward trend starts in the early 1980s instead of the early 1990s.

5.4 Word Length

We also examine how basic linguistic characteristics, such as numbers of words in song lyrics, may relate to the overall concreteness trend. As shown in Figure 8 and Figure 9, the average length of song lyrics has a negative correlation with the average annual concreteness scores. Lyrics had become longer until the late 1990s and shorter afterwards, regardless of whether word repetition was counted or not. Although the lowest point of the concreteness trend and that of the length of lyrics trend are about five years apart, they highly correlate with each other (when frequency of words are ignored: $r = -0.6096$, $p < 0.001$; when frequency of words are counted: $r = -0.5971$, $p < 0.001$).

6 DISCUSSION

This paper identified concreteness trends in pop song lyrics between 1965 and 2015. The trends went down until the early 1990s and went up after that. The high correlation with "Hip-Hop/Rap" may explain the rise after the early 1990s. The proportion of open class words and the length of song lyrics turned out to be highly correlated with the general concreteness trends, although the same trends were observed when only open class words were considered.

Among the various quantitative dimensions of text complexity, we paid particular attention to concreteness. Because concrete

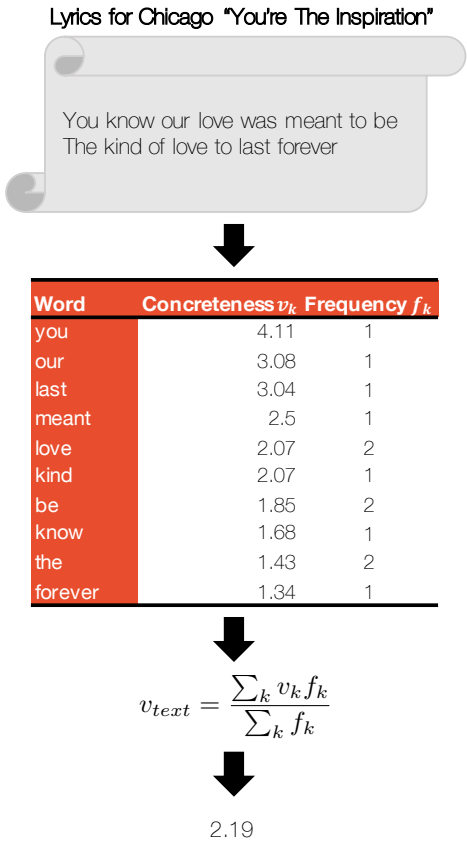

Figure 10: A pictorial example of how the overall concreteness is calculated by using the example of an excerpt of the lyrics of "You're the Inspiration" written by Chicago

words tend to be easy to remember and process, high average concreteness scores indicate lower levels of lyric complexity. We can, however, find difficult lyrics with high concreteness values or easy lyrics with low concreteness values. Common love songs with simple words and songs with metaphoric expressions are examples of such a case.

If the main abstract concept that the lyrics of a song are trying to convey is relationships or love, and artists say it directly instead of using figurative expressions, their concreteness scores tend to be low; although they tend to be easy to understand. For instance, the sentence from Chicago's "You're the Inspiration," in Figure 10, refers to relationships and love with ordinary words, such as "love", "know," and "forever." While their concreteness scores are relatively low (despite their word classes), the sentence is very easy to understand. Indeed, the topics of relationships and/or love are the most popular ones in popular music. Christenson et al. [10] explored the change in the distribution of major themes of song lyrics from Billboard biannual top 40 songs in the last 50 years. The conclusion is that the most popular theme of popular song lyrics is "Relationships/Love," ranging between 65 and 70%. Love songs are common, and ordinary words in love songs are abstract. This calls for theme-specific usage of concreteness as a readability measure.

Bohemian Rhapsody - Queen

Mama, just killed a man
Put a gun against his head
Pulled my trigger, now he's dead

Figure 11: An excerpt from song lyrics of Queen's "Bohemian Rhapsody"

Another counter-example is that of song lyrics with metaphoric expressions. Concreteness may lower the readability if metaphoric expressions add more layers to song lyrics and they are not banal. Some metaphoric expressions are challenging to understand for both humans and the state-of-the-art machine-learning technologies. Sometimes we might need high level linguistic ability to decipher metaphors. In Figure 11, the first three lines of Queen's "Bohemian Rhapsody" seem to be about murder. However, many people including the ones who knew the songwriter think that the song addresses his sexual identity instead of murdering someone[8]. Without such extra information, it is very difficult to truly understand the meanings of the lyrics, even if its concreteness is low.

These two counter-examples clearly reveal limitations of concreteness, and stress the need for additional qualitative analyses. Since it is challenging to conduct user studies, particularly on a large scale, it is important to explore how quantitative text complexity analyses can go beyond their limitations and embrace some qualitative dimensions of text complexity.

7 CONCLUSION

We explored concreteness, a quantitative dimension of text complexity, of popular song lyrics. A change point analysis and a Cox-Stuart sign test confirmed that concreteness of song lyrics showed a downward trend until 1991, followed by an upward trend. Analysis of the relationships to genres indicated that the growing popularity of hip-hop and rap may have contributed to the upward trend after 1991 because the genre's average concreteness scores are the highest among the major music genres and its prevalence coincides with the rise of song lyrics concreteness. As for word classes, although the proportion of the open word classes do correlate with the concreteness trend, the similar V-shaped trend was observed when only open word classes were considered. Therefore, the change of concreteness over time does not simply reflect the proportion of open word classes. The number of words in song lyrics may explain the concreteness, given the high correlation. Ultimately, possible limitations of concreteness as a text complexity metric were explored by analyzing counter examples. The study revealed that concreteness tends to be low for love songs and songs with figurative expressions. Considering all findings, it is important to use this metric with additional metadata, such as genres, topics, and lyric length. Moreover, like other quantitative text complexity metrics, further qualitative analyses are required to complement this metric.

[8] https://www.songfacts.com/facts/queen/bohemian-rhapsody

REFERENCES

[1] Jeanette Altarriba, Lisa M Bauer, and Claudia Benvenuto. 1999. Concreteness, context availability, and imageability ratings and word associations for abstract, concrete, and emotion words. *Behavior Research Methods* 31, 4 (1999), 578–602.

[2] Jack Atherton and Blair Kaneshiro. 2016. I said it first: Topological analysis Of lyrical influence networks. In *Proceedings of the 17th International Conference on Music Information Retrieval*. 654–660.

[3] PEN Lyrics Award. Retrieved January 30, 2017. 2012 PEN Lyrics Award. http://www.pen-ne.org/lyrics-award/2016/7/13/2012-winners.

[4] David A Balota, Melvin J Yap, Keith A Hutchison, Michael J Cortese, Brett Kessler, Bjorn Loftis, James H Neely, Douglas L Nelson, Greg B Simpson, and Rebecca Treiman. 2007. The English lexicon project. *Behavior Research Methods* 39, 3 (2007), 445–459.

[5] Anton Barua, Stephen W Thomas, and Ahmed E Hassan. 2014. What are developers talking about? An analysis of topics and trends in stack overflow. *Empirical Software Engineering* 19, 3 (2014), 619–654.

[6] Margaret M Bradley and Peter J Lang. 1999. *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Technical Report. Citeseer.

[7] Marc Brysbaert and Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods* 41, 4 (2009), 977–990.

[8] Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods* 46, 3 (2014), 904–911.

[9] Xiaobin Chen and Detmar Meurers. 2016. Characterizing Text Difficulty with Word Frequencies. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*. 84–94.

[10] Peter G Christenson, Silvia de Haan-Rietdijk, Donald F Roberts, and Tom FM ter Bogt. 2018. What has America been singing about? Trends in themes in the US top-40 songs: 1960–2010. *Psychology of Music* (2018). https://doi.org/10.1177/0305735617748205

[11] David Roxbee Cox and Alan Stuart. 1955. Some quick sign tests for trend in location and dispersion. *Biometrika* 42, 1/2 (1955), 80–95.

[12] Edgar Dale. 1931. Evaluating Thorndike's Word List. *Educational Research Bulletin* (1931), 451–457.

[13] Mark Davies. 2009. The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics* 14, 2 (2009), 159–190.

[14] Terry C Davis, Michael A Crouch, Georgia Wills, Sarah Miller, and David M Abdehou. 1990. The gap between patient reading comprehension and the readability of patient education materials. *The Journal of Family Practice* (1990).

[15] Bruno Di Giorgi, Simon Dixon, Massimiliano Zanoni, and Augusto Sarti. 2017. A data-driven model of tonal chord sequence complexity. *IEEE/ACM Transactions on Audio, Speech and Language Processing* 25, 11 (2017), 2237–2250.

[16] Peter Sheridan Dodds and Christopher M Danforth. 2010. Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies* 11, 4 (2010), 441–456.

[17] William H DuBay. 2004. The Principles of Readability. *Impact Information* (2004).

[18] Robert J Ellis, Zhe Xing, Jiakun Fang, and Ye Wang. 2015. Quantifying Lexical Novelty in Song Lyrics. In *Proceedings of the 15th International Conference on Music Information Retrieval*. 694–700.

[19] Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology* 32, 3 (1948), 221.

[20] National Governors Association Center for Best Practices and Council of Chief State School Officers. 2010. *Common core state standards*. Authors Washington, DC.

[21] Francesco Foscarin. 2017. *Chord sequences: Evaluating the effect of complexity on preference*. Ph.D. Dissertation. POLITECNICO DI MILANO.

[22] Peter Foster, Matthias Mauch, and Simon Dixon. 2014. Sequential complexity as a descriptor for musical similarity. *IEEE/ACM Transactions on Audio, Speech and Language Processing* 22, 12 (2014), 1965–1977.

[23] Edward Fry. 1968. A readability formula that saves time. *Journal of Reading* 11, 7 (1968), 513–578.

[24] Arthur C Graesser, Danielle S McNamara, and Jonna M Kulikowich. 2011. Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher* 40, 5 (2011), 223–234.

[25] Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers* 36, 2 (2004), 193–202.

[26] Robert Gunning. 1952. *The Technique of Clear Writing*. McGraw-Hill, New York.

[27] Perfecto Herrera, Juan Bello, Gerhard Widmer, Mark Sandler, Òscar Celma, Fabio Vignoli, Elias Pampalk, Pedro Cano, Steffen Pauws, and Xavier Serra. 2005. Simac: Semantic interaction with music audio contents. In *Proceedings of 2nd European Workshop on Integration of Knowledge, Semantic and Digital Media Technologies*. IET.

[28] Elfrieda H Hiebert. 2012. *Readability and the Common CoreâĂŹs staircase of text complexity*. Santa Cruz, CA: TextProject Inc.

[29] Thomas T Hills and James S Adelman. 2015. Recent evolution of learnability in American English from 1800 to 2000. *Cognition* 143 (2015), 87–92.

[30] Aline K Honingh and Rens Bod. 2010. Pitch Class Set Categories as Analysis Tools for Degrees of Tonality. In *Proceedings of the 11th International Conference on Music Information Retrieval*. 459–464.

[31] Emmanuel Keuleers, Paula Lacey, Kathleen Rastle, and Marc Brysbaert. 2012. The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods* 44, 1 (2012), 287–304.

[32] Rebecca Killick, Paul Fearnhead, and Idris A Eckley. 2012. Optimal detection of changepoints with a linear computational cost. *J. Amer. Statist. Assoc.* 107, 500 (2012), 1590–1598.

[33] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Technical Report. Institute for Simulation and Training, University of Central Florida.

[34] Kirill Kireyev and Thomas K Landauer. 2011. Word maturity: Computational modeling of word knowledge. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 299–308.

[35] George R Klare. 1968. The role of word frequency in readability. *Elementary English* 45, 1 (1968), 12–22.

[36] Marc Lafrance, Casey Scheibling, Lori Burns, and Jean Durr. 2017. Race, gender, and the Billboard Top 40 charts between 1997 and 2007. *Popular Music and Society* (2017), 1–17.

[37] Marc Lafrance, Lara Worcester, and Lori Burns. 2011. Gender and the Billboard Top 40 Charts between 1997 and 2007. *Popular Music and Society* 34, 5 (2011), 557–570.

[38] Marc Lavielle. 2005. Using penalized contrasts for the change-point problem. *Signal Processing* 85, 8 (2005), 1501–1510.

[39] Junghyuk Lee and Jong-Seok Lee. 2015. Predicting music popularity patterns based on musical complexity and early stage popularity. In *Proceedings of the 3rd Edition Workshop on Speech, Language & Audio in Multimedia*. ACM, 3–6.

[40] Colleen Lennon and Hal Burdick. 2014. The lexile framework as an approach for reading measurement and success. http://www.lexile.com/research/1/.

[41] Gondy Leroy and David Kauchak. 2013. The effect of word familiarity on actual and perceived text difficulty. *Journal of the American Medical Informatics Association* 21, e1 (2013), e169–e172.

[42] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics System Demonstrations*. 55–60. http://www.aclweb.org/anthology/P/P14/P14-5010

[43] Manuela M Marin and Helmut Leder. 2013. Examining complexity across domains: relating subjective and objective measures of affective environmental scenes, paintings and music. *PLoS One* 8, 8 (2013), e72412.

[44] Ladislav Maršík, J Pokornyy, and Martin Ilcík. 2014. Improving music classification using harmonic complexity. In *Proceedings of the 14th conference Information Technologies - Applications and Theory*. 13–17.

[45] Matthias Mauch and Mark Levy. 2011. Structural Change on Multiple Time Scales as a Correlate of Musical Complexity. In *Proceedings of the 12th International Conference on Music Information Retrieval*. 489–494.

[46] G Harry Mc Laughlin. 1969. SMOG grading-a new readability formula. *Journal of Reading* 12, 8 (1969), 639–646.

[47] Danielle S McNamara, Arthur C Graesser, Zhiqiang Cai, and Jonna M Kulikowich. 2011. Coh-Metrix easability components: Aligning text difficulty with theories of text comprehension. In *Annual Meeting of the American Educational Research Association, New Orleans, LA*.

[48] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331, 6014 (2011), 176–182.

[49] Michael Milone. 2014. *Development of the ATOS™: Readability Formula*. Renaissance Learning, Incorporated.

[50] Claudia Monica Ferradas Moi. 1994. Rock Poetry: The Literature Our Students Listen To. *Journal of the Imagination in Language Learning* (1994), 56–59.

[51] Jessica Nelson, Charles Perfetti, David Liben, and Meredith Liben. 2012. Measures of text difficulty: Testing their predictive value for grade levels and student performance. *Council of Chief State School Officers, Washington, DC* (2012).

[52] Robert Mitchell Parry. 2004. *Musical complexity and top 40 chart performance*. Technical Report. Georgia Institute of Technology.

[53] Guillaume Robal and Tim Blackwell. 2014. *Live algorithms with complexity matching*. Technical Report. University of London.

[54] Mark Sadoski. 2001. Resolving the effects of concreteness on interest, comprehension, and learning important ideas from text. *Educational Psychology Review* 13, 3 (2001), 263–281.

[55] Mark Sadoski, Ernest T Goetz, and Joyce B Fritz. 1993. Impact of concreteness on comprehensibility, interest, and memory for text: Implications for dual coding

theory and text design. *Journal of Educational Psychology* 85, 2 (1993), 291.
[56] Kathleen M Sheehan, Irene Kostin, Diane Napolitano, and Michael Flor. 2014. The TextEvaluator tool: Helping teachers and test developers select texts for use in instruction and assessment. *The Elementary School Journal* 115, 2 (2014), 184–209.
[57] W Sichert-Hellert, M Kersting, U Alexy, and F Manz. 2000. Ten-year trends in vitamin and mineral intake from fortified food in German children and adolescents. *European Journal of Clinical Nutrition* 54, 1 (2000), 81.
[58] Sebastian Streich. 2006. *Music complexity: A multi-faceted description of audio content.* Ph.D. Dissertation. Universitat Pompeu Fabra.
[59] Sebastian Streich and Perfecto Herrera. 2005. Detrended fluctuation analysis of music signals: Danceability estimation and further semantic characterization. In *Proceedings of the 118th Audio Engineering Society Convention.*
[60] Edward L Thorndike. 1921. *The TeacherâĂŹs Word Book.* Teacher's College, Columbia University.
[61] Alexandros Tsaptsinos. 2017. Lyrics-Based Music Genre Classification Using a Hierarchical Attention Network. In *Proceedings of the 18th International Conference on Music Information Retrieval.*
[62] Christof Weiß and Meinard Müller. 2014. Quantifying and visualizing tonal complexity. In *Proceedings of the Conference on Interdisciplinary Musicology.*
[63] Christof Weiss and Meinard Müller. 2015. Tonal complexity features for style classification of classical music. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing.* 688–692.