COMPUTATIONAL LYRICOLOGY: QUANTITATIVE APPROACHES TO UNDERSTANDING SONG LYRICS AND THEIR INTERPRETATIONS

by

KAHYUN CHOI

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Library and Information Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Doctoral Committee:

Professor J. Stephen Downie, University of Illinois, Chair
Professor Michael Twidale, University of Illinois
Professor Ted Underwood, University of Illinois
Associate Professor Sally Jo Cunningham, University of Waikato

Abstract

Recently, music complexity has drawn attention from researchers in the Music Information Retrieval (MIR) 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 dissertation research investigates song lyric complexity and how it might be measured computationally.

In a broad sense, lyric complexity comes from two aspects of text complexity–quantitative and qualitative dimensions–that have a complementary relationship. For a comprehensive understanding of lyric complexity, this study explores both dimensions. First, for the quantitative dimensions, such as word frequency and word length, refer to those that can be measured efficiently using computer programs. Among them, this study examines the concreteness of song lyrics using trend analysis. Second, on the contrary to the quantitative dimensions, the qualitative dimensions refer to a deeper level of lyric complexity that requires attentive readers' comprehension and external knowledge. However, it is challenging to collect a large-scale qualitative analysis of lyric complexity due to the resource constraints. To this end, this dissertation introduces user-generated interpretations of song lyrics that are abundant on the web as a proxy for assessing the qualitative dimensions of lyric complexity. To be specific, this study first examines whether the user-generated data provide quality topic information, and then

proposes a Lyric Topic Diversity Score (LTDS), a lyric complexity metric based on the diversity of the topics found in users' interpretations. The assumption behind this approach is that complex song lyrics tend to provoke diverse user interpretations due to their properties, such as ambiguous meanings, historical context, the author's intention, and so on.

The first findings of this study include 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. Second, interpretations are a good input source for automatic topic detection algorithms. Third, the interpretation-based lyric complexity metric looks promising because it is correlated with Lexical Novelty Scores (LNS), the only previously developed lyric complexity measure. Overall, this work expands the scope of music complexity by focusing on relatively unexplored data, song lyrics. Moreover, these findings suggest that any potential analysis and application on any objects can benefit from this kind of auxiliary data, which is in the form of user comments.

To Minje and Sage

Acknowledgments

First and foremost, I would like to express my gratitude to my advisor Prof. J. Stephen Downie for helping me feel a sense of belonging in the world of music information retrieval and digital libraries by introducing me to wonderful people, and providing fascinating research opportunities. I am also thankful for his endless constructive feedback during the long revision process of my dissertation.

I also owe a debt of gratitude to the rest of my wonderful committee: Prof. Michael Twidale, Prof. Ted Underwood, and Prof. Sally Jo Cunningham for their guidance and constructive feedback. Their words of encouragement empowered me through this intense time in my life.

I am also grateful to my academic sisters and role models, Prof. Jin Ha Lee and Prof. Xiao Hu. I am so thankful to have gotten to know them and have had chances to work with them. I learned a lot about how to do research and write papers from them.

This dissertation would not have been possible without the invaluable datasets. I thank Michael Schiano for providing the access to the precious user-generated comments on songmeaning.com. I also thank Roy Hennig for offering the gigantic repository of lyrics for research.

I extend my gratitude to the IMIRSEL labmates: Andreas Ehmann, Mert Bay, Craig Willis, and Yun Hao. I am grateful to Andreas and Mert for helping me to adjust to the new settings. I am deeply grateful to Craig Willis for his friendship and great work ethic. I extend my thanks to Yun Hao, my academic sister, for listening to me and helping me look on the bright side.

Last but not the least, I would like to thank my family. I am deeply grateful to my mom, my brother, my sister-in-law, and my parents-in-law for believing in me and supporting me emotionally and spiritually. Special appreciation goes to my husband Minje Kim and my daughter Sage Anne Kim. I cannot thank these two people enough for making me laugh and happy in any circumstance.

Contents

Listing of figures

List of Tables

Chapter 1

Introduction

1.1 Automatic Music Annotation

Automatically annotating digital music with appropriate metadata has been a significant topic in Music Information Retrieval (MIR) research [1]. When music was mostly stored and shared in analog forms, such as vinyl discs and tapes, most people were able to access relatively small amount of music through their own collections, radio, and TV. Popular songs were mostly organized and searched by their textual metadata, such as titles, artists, and genres. During the analog phase of music history, there were almost no automatic solutions to store and recommend music. However, now listeners can access millions of songs at any moment. For instance, Spotify, the most popular music streaming service, has over 30 million songs and more than 20,000 others are added everyday.[1]

This ever-growing amount of digital music has created challenges for MIR researchers, such as how to organize the large-scale music collections; how to provide better searching/browsing/recommendation interfaces to listeners so that they can navigate the sea of music easily; and how to solve the scalability issue of manual annotation. To answer these questions, MIR researchers have worked to enhance the process of describing music by developing new techniques to work with music metadata and music content.

Research on automatic music annotation has aimed at extracting various types of music descriptors by using a variety of data sources, such as audio, song lyrics, and user-generated data on the web. As the main source of music content analysis, various aspects of audio have been analyzed. For instance, one example of early automatic annotators, a part of the Semantic Interaction with Music Audio Contents (SIMAC) project, extracted lower-level descriptors from music audio: rhythm, harmony, timbre, and

[1] https://freeyourmusic.com/blog/best-music-streaming-platform/

instrumentation; intensity; structure; and complexity [1]. Other sets of music descriptors that interest MIR researchers can be found in the Music Information Retrieval Evaluation eXchange (MIREX), the major annual MIR system evaluation event held since 2005. So far, more than 2,000 algorithms were submitted to identify a variety of music descriptors: artist, chord, composer, genre, key, note, mood, onset, tag, tempo, etc., as shown in Table 1.1 [2]. MIREX submissions extract low-level features, such as note and onset, as well as high-level features, such as mood and genre [3].

Lyrics, the other main content of songs, have been used as an input to automatic music annotation systems. Various types of metadata, such as topic, language, mood, genre, and the level of readability, have been explored. The preliminary study done by Mahedero et al. used state-of-the-art computational natural language processing technologies to analyze song lyrics for multiple tasks, including language identification and thematic categorization [4]. MIR researchers have also worked on automatically identifying topics of song lyrics [4–12]. As for the readability measures of song lyrics, Ellis et al. drew people's attention to lyrics' readability by proposing the lyrics novelty measure [13]. Moreover, various MIR studies have come up with sophisticated systems that can extract mood and genre from song lyrics as either main data or auxiliary data [14–22].

As the web grows, not only audio and song lyrics but also user-generated data on the web have attracted the interest of MIR researchers [3, 23, 24]. There are some websites that provide organized music databases curated by a selective few, such as *allmusic.com* and *songfacts.com*. The taxonomies of music used there, such as genre, mood, and topic, have been used as a source of ground truth data for automatic music classification systems [9, 18, 23]. Another kind of web platform for music is websites where people share their opinions and explanations of music, such as *amazon.com*, *twitter.com*, *songmeanings.com*, *genius.com*, and *Last.fm*. Reviews, social tags, microblog postings, and discussion forum postings have also been used to improve automatic music annotation systems. The voluntarily user-created data provide rich qualitative information for free [7–10, 23, 25, 26]. However, as those data also include irrelevant and useless information, researchers need to process the data carefully [9, 24].

1.2 Automatic Lyric Complexity Annotation

Among various music descriptors, this dissertation research focuses on "complexity." Particularly, this dissertation investigates computational methods that can measure how difficult song lyrics are to understand. This research can be framed as music complexity

TASK NAME	2005	2006	2007	2008	2009	2010	2011	2012	2013
Audio Artist Identification	7		7	11					
Audio Beat Tracking		5		15(2)	22(2)	26(2)	24(2)	60(3)	54(3)
Audio Chord Detection				11	18(2)	15	18	22(2)	36(3)
Audio Classical Composer ID			7	8	30	27	16	15	14
Audio Cover Song Identification		8	8		6(2)	6(2)	4(2)		2
Audio Drum Detection	8								
Audio Genre Classification	15		7	26(2)	65(2)	48(2)	31(2)	31(2)	26(2)
Audio Key Detection	7					5	8	6	3
Audio Melody Extraction	10	10(2)		21(3)	12(6)	30(6)	60(6)	24(6)	24(6)
Audio Mood Classification			9	13	33	36	17	20	23
Audio Music Similarity		6	12		15	8	18	10	8
Audio Onset Detection	9	13	17		12	18	8	10	11
Audio Tag Classification				11	24(3)	26(2)	30(2)	18(2)	8(2)
Audio Tempo Extraction	13	7				7	6	4	11
Discovery of Repeated Themes & Sections									16
Multiple Fundamental Frequency Estimation & Tracking			27(2)	28(2)	26(3)	23(3)	16(2)	16(2)	6(2)
Query-by-Singing/Humming		23(2)	20(2)	16(2)	12(4)	20(4)	12(4)	24(4)	28(5)
Query-by-Tapping				5	9(3)	6(3)	3(3)	6(3)	6(3)
Real-time Audio to Score Alignment		2		4		5	2	3	2
Structural Segmentation					5	12(2)	12(2)	27(3)	26(3)
Symbolic Genre Classification	5								
Symbolic Key Finding	5								
Symbolic Melodic Similarity	7	18(3)	8			13	11	6	6
Total Number of Runs per Year	86	92	122	169	289	331	296	302	310
Total Number of Runs (2005-2013)	1997								

Table 1.1: MIREX Tasks and the Number of Runs between 2005 and 2013 (Downie et al. [2])

research and text complexity research: lyric complexity can be a part of music complexity because song lyrics are a part of most of popular songs. Lyric complexity can also be text complexity in that song lyrics are a text genre.

Music Complexity

Although music complexity has been an important topic for the last several decades in music studies, automatic annotation of music complexity became important in MIR recently, like other music metadata [27]. In the 1980s, Berlyne et al. discovered the close relationship between musical complexity and musical preference. The study revealed that listeners tend to like music with moderate complexity more than too simple or too complex music [28]. To capture this crucial descriptive information from songs automatically, Streich defined the complexity of music using four aspects and came up with systematic ways to extract them from music signals [27].

Lyrics have been mostly excluded in MIR research, despite the facts that most popular songs have lyrics and their complexity influences overall music complexity. Streich's research on musical complexity explored various aspects of music, but it focused only on the instrumental part of songs[27]. He also acknowledged that lyric complexity is a part of musical complexity, like other aspects including timbre, melody, and rhythm, but left the research gap for future researchers. This dissertation aims at filling this gap by focusing on the complexity of popular song lyrics.

Another reason why this topic needs to be investigated is that complexity of song lyrics has potential applications to various MIR operations, including visualizing, browsing, searching, and recommending music as identified by Streich [27]. Songs can be classified as low, moderate, or high depending on their lyrical complexity. As for the low or moderate lyrical complexity, "Easy listening" and "Focus" exemplify genres in which songs fit the low and moderate categories.These songs are appropriate for "Easy listening" because this music usually functions as the background sound to calm or uplift people's mood. Similarly, when users listen to music to focus on work, they might not want to be distracted by abstruse song lyrics. On the other hand, other users might be interested in complex music. For instance, if a user would like to listen to poetic songs, MIR systems can recommend songs with high lyrical complexity or written by poetic lyricists who often write profound song lyrics. Finally, users might want to discover controversial song lyrics to explore the world of music or just to pass the time. However, to date no tools or services that provide such a function exist.

Text Complexity

Given that song lyrics are text, the first reasonable approach to automatically annotate lyrical complexity is applying traditional text complexity metrics to them. Over the last 100 years, various readability measures have been developed and applied to several different environments, such as schools, military buildings, and hospitals to recommend proper levels of text to readers [29]. The basic features of these computational metrics take into consideration the difficulty level of words, number of syllables, and number of words in a sentence [30, 31]. However, as machine-learning technologies are evolving, researchers have been devising algorithms that can capture even higher levels of complexity, such as coherence and semantics [31].

Song lyrics are not usually in prose but verse, which brings up critical issues when using traditional text complexity metrics to measure song lyrics. Traditional text complexity metrics have been developed to measure mostly prose, such as textbooks, children's books, novels, manuals, etc. [29]. Unlike these types of text, song lyrics are composed of lines in verses instead of full sentences and paragraphs. Furthermore, lyrics are much shorter than other types of text and repetitive as well. Because of these differences, metrics based on higher units, such as sentence-level and paragraph-level metrics, are not directly applicable to song lyrics.

Despite the usefulness of text complexity metrics, they also have clear limitations. Regardless of types of text, some aspects of text complexity that can be captured by humans but not by traditional text complexity metrics [32–34]. The Common Core State Standards (CCSS), a set of high-quality academic standards in mathematics and English-language arts/literacy (ELA), introduced two dimensions of the inherent complexity of text: qualitative and quantitative dimensions of text complexity [2][35]. Text complexity metrics are only expected to measure quantitative aspects of text complexity efficiently, such as word length and text cohesion [34]. On the other hand, CCSS recommends an attentive reader to measure qualitative aspects of text complexity, such as levels of meanings, structure, knowledge demands, etc. [34].

In addition to these limitations, the unique nature of song lyrics poses additional challenges. Lyrics can have many layers of meanings coming from information beyond the text itself, such as the authors' biographic information and social context. Besides, the poetic nature of song lyrics often makes them complex by opening up many different interpretations. Humans might capture this higher level of complexity, but to date no

[2]http://www.corestandards.org/assets/Appendix_A.pdf

machine algorithm can collect relevant information and interpret song lyrics based on extra information as humans do. The better way to comprehensively measure text complexity is to rely on human evaluation [32, 34]. However, the manual evaluation has its own limitations: high cost and scalability.

To the best of my knowledge, there is no existing users' evaluation data regarding the text complexity of song lyrics. However, the enormous amount of user-generated interpretations of song lyrics on the web can be used to build a proxy to the missing data [7–9]. Often it is a difficult job for humans to fully understand the meaning of song lyrics, and when we listen to a song, we wonder what the song is about. To satisfy this curiosity, several websites, such as *songmeanings.com* [36], *genius.com* [37], and *lyricinterpretations.com* [38], have been created and running for a few years or even decades. According to the monthly traffic reports by *similarweb.com* [39], those websites have a big pool of users with 5M, 64M, and 0.6M total monthly visits, respectively, as of October 2016. People discuss the meaning of song lyrics by reading and posting their own interpretations, and rating them to show how much they agree or disagree with other users' interpretations.

These interpretations of song lyrics are invaluable because they allow us to observe how people comprehend song lyrics. Usually, songs that generate a great variety of different interpretations are complex, while those with converging interpretations are less complex. Computationally modeling this behavior of users will be quite a different approach to text complexity metrics because this approach is based on users' responses, while the traditional approach is based solely on lyric text. Like other social data, the interpretations of song lyrics are noisy with irrelevant information, they need to be assessed whether they are proper input to computational analysis systems.

1.3 Research Questions

This dissertation aims to investigate how to extract the complexity of song lyrics computationally to enhance tools for searching, browsing, recommending, and visualizing music. Among the multiple factors that make song lyrics complex, some come from linguistics, such as infrequently used phrases, while others derive from circumstances beyond the words, such as the songwriter's biography and the lyrics' social context. While the former can be captured by traditional text complexity metrics, the latter calls for human interpretations. Thus, this study aims to look at both sides and explore not only traditional complexity methods, but also user data from web discussions on the meanings

of song lyrics. To this end, the overarching research question of this dissertation is: How can the complexity of song lyrics be measured computationally? The following three studies answer specific research questions related to the overarching research question.

1.3.1 Study 1: Concreteness of Words as a Lyric Complexity Metric

The first part of this dissertation takes a traditional linguistic approach that measures quantitative dimensions of text complexity. Text complexity has been explored since the 1920s, and various quantitative variables ranging from word-level to sentence-level to paragraph-level have been identified. This study primarily focuses on word-level variables. One of the reasons why I have excluded variables of higher-level units is because song lyrics are not structured with sentences and paragraphs, but lines in verses. The other reason is that I consider more reasonable to start exploring the fundamental units and subsequently expand the scope of the research to embrace higher level components.

There are many word-level variables, such as word frequency, word length, word familiarity, word grade level, Pearson word maturity metric, and concreteness. I discuss each feature in more detail in chapter 2. This study focuses on concreteness, the concept that distinguishes between concrete and abstract words. Concrete words are those that we can experience using our senses, while abstract words can be understood only by other words [40]. For example, 'table' and 'chair' are concrete words because we can see and touch what they mean. On the other hand, 'loyalty' and 'justice' are abstract words because they can only be explained by other words and examples. Among all of the word-level variables, I have selected concreteness for the following reasons:

1. This is the first study that explores concreteness of song lyrics regarding readability.

2. Lyrics are the easiest of all types of text to memorize, and concreteness is also related to memorability. Although this study focuses only on complexity, the findings can be applied to research on memorability of song lyrics in the future.

3. Imageability is highly correlated to concreteness, and images are widely used in lyrics. The findings will be useful for research on lyrics imageability in the future.

4. The concreteness of large-scale book corpora has been explored recently [41]. However, song lyrics were not included in the research, although it is considered as genres of literature. This research will expand research on literature concreteness by incorporating song lyrics.

5. Word concreteness ratings of 40,000 words are available while the other word-level variables are not.

The primary objective of this study is to investigate concreteness as a complexity metric for song lyrics. This study examines how concreteness of song lyrics has changed over time. I use large-scale, crowd-sourced ratings to measure overall concreteness of each selected song lyrics (I provide additional details of the process in Chapter 3.2.3). I analyze correlations between song lyrics and factors, including Part-of-Speech (POS) tags, genres, and artists to deeper understand the concreteness trend of song lyrics. Because some of POS tags tend to have much lower concreteness scores than the others, if there are more words of song lyrics from the former, the overall concreteness score is lower [41]. For this reason, this study breaks down words into two groups, one for POS tags with high concreteness and another for POS tags with low concreteness. I can thus investigate how their proportion over time influence the concreteness trend. Furthermore, I explore the concreteness of genres to see if each genre has their concreteness score and how the rise and fall of each genre influence the concreteness trend. Finally, in order to further understand the nature of this measure, I examine manually the five most concrete and the five least concrete songs.

With the purpose of studying the measuring process of a word-level linguistic feature, mainly concreteness, of song lyrics as a text complexity metric, this dissertation proposes to answer the following research questions.

- Research Question 1: How has text complexity of popular song lyrics changed over time in terms of concreteness?
  - Research Question 1-1: What is the relationship between the concreteness trends and genres?
  - Research Question 1-2: What is the relationship between the concreteness trends and word statistics in song lyrics?

1.3.2 Study 2: Evaluating Usefulness of User-generated Interpretations

Because some aspects of complexity can be captured only by observing how people comprehend song lyrics, this dissertation also assesses users' interpretations as input for automatic annotation systems in Study 2. Although the user data contains useful information, it is also noisy with irrelevant information, such as expressing how much they

like or dislike a song or an artist [8]. For this reason, it is necessary to determine how much useful information they provide and to compare the data with song lyrics as an input to MIR systems.

This study compares the two sources in the lyrics topic classification task, the user-generated data and song lyrics. Because the user-generated data based approach to the complexity of song lyrics tries to measure the number of topics discussed among people, it is essential to determine whether the user-generated data contain enough topic-related information and whether automatic classification systems can accurately extract the information. To this end, this study not only compares the classification accuracy but also analyzes the most representative terms from each source to deeply understand the differences between them.

Although previous studies have tested the usefulness of user-interpretations for automatic topic classification/extraction tasks, these studies have represented words using the bag-of-words model, which is a primitive representation [7–9]. Word embedding has been devised recently, and it is considered more advanced than the bag-of-words model in that it is global and semantic [42]. This study aims to compare the usefulness of the two sources when words are represented as word embedding. By expanding the previous studies, this dissertation proposes a more comprehensive understanding of users' interpretations on the web as an input to an automatic annotation system.

Thus, the study of assessing the crowd-sourced, user-generated data as an input to MIR systems explores the research questions below.

- Research Question 2: Can an automatic algorithm successfully identify underlying topics of song lyrics from user-generated interpretations?
  - Research Question 2-1: Are users' interpretations of song lyrics more useful than song lyrics for the topic annotation task?
  - Research Question 2-2: How different are the most representative words of interpretations and lyrics?

1.3.3 Study 3: Topical Diversity of Interpretations as a Lyric Complexity Metric

Study 3 aims to introduce a text complexity measure based on users' interpretations of song lyrics that are shared online. My primary focus is on disputes among users over controversial song lyrics and agreements over simple or straightforward song lyrics. Such a

user-based measure is necessary because a quantitative approach can capture qualitative dimensions of text complexity. For instance, although it is indirect, this metric might be able to capture complexity coming from high-level linguistic features, such as cynicism and humor. Furthermore, it can also grasp how extra information, such as the author's intention and social context, opens up a variety of interpretations.

This dissertation analyzes song lyrics using a statistical topic modeling algorithm to measure the level of disagreement among users over a particular song lyric. According to Choi et al., topic modeling the users' interpretations was proved to capture topics of song lyrics reasonably well [8]. The assumption behind this approach is that complex songs tend to lead to different kinds of interpretations with many different topics, while simple or straightforward songs tend to lead to a small number of interpretations with only a few topics. This study explores how to measure the diversity of learned topics from the interpretations of song lyrics. Moreover, to better understand the proposed metric, this study examines the correlation of the metric with a traditional text complexity metric.

Concerning the user-generated data based complexity measure, this study addresses the following research questions.

- Research Question 3: Would the diversity of topics in interpretations of song lyrics be useful for measuring the complexity of song lyrics?
  - Research Question 3-1: Can the proposed measure in Chapter 5 capture differences in topical diversity between popular songs and less popular songs?
  - Research Question 3-2: Can we understand the measure better by analyzing the relationship between diversity of topics in song interpretations and lexical novelty of song lyrics?

1.4 Chapter Outline

This dissertation is organized into six chapters. Chapter 2 contextualizes this study in the literature on automatic music complexity annotation in MIR, text complexity metrics, and topic modeling. Chapter 3 focuses on concreteness, one of the variables used in the traditional text complexity metrics, and analyzes how concreteness of song lyrics has changed over time. Chapter 4 introduces user-generated interpretations that contain qualitative analysis of song lyrics and assesses their usefulness as an input to automatic topic annotation systems. Chapter 5 proposes a method that uses the user-generated

interpretations as a proxy for computing complexity of song lyrics. Chapter 6 concludes this dissertation with key findings, limitations, and possible future research.

1.5 Summary

This chapter contextualized automatic lyric complexity annotation research by introducing the importance of automatic music annotation in MIR in the digital music era, pointing out how automatic music complexity annotation research has focused primarily on instrumental parts of music instead of song lyrics.

The overarching research question was first proposed, and three studies to answer the question were introduced: investigating the traditional text complexity approach, examining user-generated interpretations as an input to an automatic topic annotation system, and exploring how to utilize the voluntarily created qualitative data as a proxy to assess lyric complexity.

Chapter 2

Literature Review

2.1 Music Complexity in MIR

Research on music complexity in MIR began in the mid 2000s, and has focused on computation methods of measuring complexity [13, 27, 43–55]. Complexity measures were developed as features or metadata used to better describe music, resulting in better MIR systems. Rather than taking a holistic approach, MIR researchers broke down music into major facets, and focused on each facet's complexity. In the early days of MIR, Downie suggested that the following seven facets played the most important roles in MIR: the pitch, temporal, harmonic, timbral, editorial, textual, and bibliographic facets [56]. To the best of my knowledge, MIR researchers in music complexity has focused on only five facets: the pitch, temporal, harmonic, timbral, and textual facets. As Streich pointed out, the first four facets are accessible in the domain of audio signal processing. Section 2.1.1 reviews research on computational measures of song audio complexity in MIR in terms of each facet. The last textual facet "[includes] the lyrics of songs, arias, chorales, hymns, symphonies, and so on" [56]. In particular, the lyrics of popular songs are the primary musical aspect of this dissertation study, and Section 2.1.2 reviews research on computational measures of song lyrics complexity in MIR.

2.1.1 Computational Measures of Song Audio Complexity in MIR

For the purposes of this study, the four audio-related facets noted above were renamed or merged based on how they were used in the literature of audio music complexity in MIR. The redefined facets–tonal, rhythmic, and timbral–are the set of facets that MIR researchers explored in terms of audio music complexity. The tonal facet encompasses both the pitch and harmonic facets because the current audio signal processing technologies still

cannot precisely detect pitches from polyphonic music, so analysis usually resorts to abstract higher level descriptions of harmonic information [27]. The two were separate in Downie's seven facets because they were drawn from symbolic music (i.e., printed music), where the pitch and harmonic information are readily available. The rhythmic facet corresponds to Downie's temporal facet, which "includes tempo indicators, meter, pitch duration, harmonic duration, and accents. Taken together these five elements make up the rhythmic component of a musical work" [56]. Finally, the "timbral facet comprises all aspects of tone color", and is strongly related to how instruments are used in music performance [56].

Except the most recent work in MIR, tonal complexity has been captured by an abstract description of the tonal information, such as chroma and the Harmonic Pitch Class Profile features (HPCP) [27, 46, 50, 51, 57]. Those features represent audio signal across the twelve pitch classes (C to B) by selecting components of corresponding frequencies in the spectrogram [58, 59]. A variety of statistical measures, such as entropy, flatness, and the Kullback-Leibler divergence, have been used to measure tonal complexity [27, 46, 51]. However, recent efforts using deep learning algorithms on a large-scale dataset with annotated chords have made chord identification more accurate. Therefore, Giorgi et al. and Foscarin et al. proposed tonal complexity measures based on identified chords using chord identification systems [54, 55]. Rhythmic complexity has been derived by measuring how strong periodic trends appear in a music audio signal [44, 46]. In particular, among various types of rhythmic complexity, Streich et al. focused on rhythmic complexity in relation to danceability, whether people can easily dance to the music [44]. They categorized music into three danceability categories, "extremely easy" music, such as techno and Brazilian, "moderately easy" music, such as jazz, rock and roll, and Brazilian popular music, and "very difficult" music, such as high art music and classical music. They considered danceability and rhythmic complexity as having an inverse correlation.

The essential factor in timbral complexity is the number of instruments used. However, current technology cannot detect instruments correctly from a randomly selected music sample. Therefore, a variety of algorithms were developed to estimate that information by using spectral features, such as the Mel Frequency Cepstral Coefficients, Spectral Roll-Off, and the Spectral Flatness measure [27, 46].

2.1.2 Computational Measures of Song Lyrics Complexity in MIR

Recently, Ellis et al. [13] 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 [60]. Word frequency information from both modern and traditional readability formulas are primarily derived from written corpora [61], although Coh-Metrix, one of the modern readability metrics, also exploited spoken sources, including the BBC World Service and taped telephone conversations [62]. Ellis et al. [13] made public the LNS values of 275,905 lyrics, which is used in this research to find a relationship to the proposed complexity metric.

2.2 Measuring Text Complexity

Text complexity is a broad concept that includes not only quantitative aspects measured by readability formulas but also qualitative aspects of text [63]. When it comes to computationally measure text 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 [29]. 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 [30, 31]. 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 [30], or the comprehension of medical patient education materials [64]. 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.

This section reviews the history of readability studies to provide an overview of quantitative approach to text complexity. Subsequently, word-level variables used in text complexity metrics and qualitative dimensions of text complexity are discussed.

2.2.1 Classic Readability Studies

According to DuBay, readability studies can be divided into classic readability studies and new readability studies [29]. The classic readability studies cover the formulas developed before the early 1950s, including the Flesch Reading Ease formulas [65], the Dale-Chall formula [66], and the Gunning Fog formula [67]. The main factors that these formulas take into account are the number of words per sentence, the number of syllables per sentence, and the percentage of difficult words [13, 34].

2.2.2 New Readability Studies

Since the early 1950s, new readability studies have explored a variety of aspects of text complexity in depth [29], improving the formulas by taking various approaches. For instance, they have examined their shortcomings to complement them [32, 33]. Moreover, they have adopted the latest knowledge related to readability from various disciplines, such as cognitive psychology, linguistics, and machine learning. Computer software containing readability formulas, such as The Writer's Workbench, was introduced in the early 1980s [68]. In addition, readability statistics have been included in Microsoft Word for at least 10 years. More computer programs that can measure text complexity have appeared, and educators have used them to measure the text complexity of textbooks [31] .

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 [35]. Given that more than 40 states in the U.S. have adopted the standard[1], it is the 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 determines the level of text difficulty based on a three-part model consisting of quantitative dimensions, qualitative dimensions, and reader and task considerations.[2] The

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

[2] http://www.corestandards.org/assets/Appendix_A.pdf

quantitative dimensions of text complexity are focused on using computer software to quickly measure the difficulty level of text. The qualitative dimensions of text complexity are factors that require human evaluation. Reader and task considerations emphasize that each student be given an appropriate text based on the student's "motivation, knowledge, and experience."[3] As this dissertation study focuses on general text complexity metrics instead of personalized ones, both quantitative and qualitative dimensions of text complexity have been reviewed in detail.

2.2.3 Quantitative Dimensions of Text Complexity

Quantitative dimensions of text complexity includes the number of words in a sentence, the number of syllables in a word, the degree of abstraction of a word [35]. These aspects of text complexity can be more efficiently measured by computer programs. There are a variety of proprietary and non-proprietary tools with readability metrics that educators can use. As the text complexity is measured by machines rather than humans, such tools can process long texts very rapidly.

CCSS also provides information on the latest text complexity metrics.[4] It includes one public domain readability metric, Flesch-Kincaid [69], and six representative text complexity tools: ATOS™ by Renaissance Learning [70], Degrees of Reading Power® by Questar Assessment, Inc.[5], The Lexile® Framework for Reading by MetaMetrics [71], Reading Maturity by Pearson Education[6], SourceRater by Educational Testing Service [72], and easability indicator by Coh-Metrix [62]. 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-level Complexity Variables

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 [61, 73]. 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

[3] http://www.corestandards.org/assets/Appendix_A.pdf

[4] http://www.corestandards.org/wp-content/uploads/Appendix-A-New-Research-on-Text-Complexity.pdf

[5] http://www.questarai.com/assessments/district-literacy-assessments/degrees-of-reading-power/

[6] http://www.readingmaturity.com/

words [74]. 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 [70, 71]. 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 [13].

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 [65, 67, 75, 76]. 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" [74, 77]. Recent studies still use word frequency to derive word familiarity [78], however word familiarity can be also obtained directly through user studies [77]. For instance, Coh-Metrix includes a psycholinguistic database that includes word familiarity scores of thousands of words from adult subjects [79].

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 [70]. 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 [80]. 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 [80].

Concreteness: Concreteness ratings are used by SourceRator and Coh-Metrix [72, 79]. 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 [40]. It has been found that readability and word concreteness correlate with each other [31] when their relationship was tested to prose.

2.2.4 Qualitative Dimension of Text Complexity

In CCSS, "qualitative dimensions and qualitative factors refer to those aspects of text complexity best measured or only measurable by an attentive human reader, such as levels of meaning or purpose; structure; language conventionality and clarity; and knowledge demands."[7] Experts are needed to perform qualitative analysis of text complexity, as the current quantitative tools cannot capture text complexity based on the elements below, which are selected based on previous readability studies. CCSS's English Language Arts Appendix A provides detailed explanations of the four important elements of qualitative dimensions of text complexity:

Levels of Meaning: Some texts have one level of meaning that is easily identified. More complex texts, with multiple levels of meaning, can be easily distinguished. Extremely complex texts, with multiple levels of meaning, are difficult to interpret due to their ambiguity. Current machine learning technologies are not evolved enough to distinguish multiple layers of meaning, so attentive readers need to identify and interpret complex texts. This dissertation study argues that topical diversity of user interpretations of song lyrics can be correlated with levels of meaning.

Structure: Easy texts tend to have simple, conventional structures, while complicated

[7] http://www.corestandards.org/assets/Appendix_A.pdf

texts tend to have complex, unconventional structures. Stories told using multiple points of view with many characters tend to be more difficult to understand. It is also difficult to follow stories that are not arranged in chronological order. The structure of song lyrics might be related to the level of their difficulty; however, it is out of scope of this dissertation study.

Language Conventionality and Clarity: Texts with low complexity tend to rely on literal, clear, and contemporary language. Figurative and ironic language make texts complex. In addition, some authors intentionally use ambiguous language to mislead readers, increasing the difficulty of the texts. Furthermore, archaic or unfamiliar language also leads to higher text complexity.

Knowledge Demands: People can easily understand texts that convey only a single familiar theme. However, texts with multiple complex or abstract themes often require uncommon life experiences to interpret them. In terms of intertextuality, texts with low complexity tend to have no or few references or allusions to other texts, while texts with high complexity tend to have many. The level of cultural knowledge required to understand texts also determines the level of text complexity.

2.3 Determining Topics of Song Lyrics

The task of determining text topics is highly related to measuring text complexity. Levels of meaning is a qualitative dimension of text complexity, as shown in Section 2.2.4. To be specific, while it is easy to identify topics of texts with one level of meaning, it is difficult to identify topics of complex texts with multiple levels of meaning. This section reviews literature on automatic topic identification research in MIR.

Section 2.3.1 explains a topic modeling algorithm with an illustrative example. Based on the generative model of the topic modeling, Section 2.3.2 describes how understanding lyrics of a song means having a clear interpretation. Section 2.3.3 and Section 2.3.4 review MIR research on automatic identification of topics of song lyrics by using song lyrics and their interpretations.

2.3.1 Topic Modeling

This study uses the Latent Dirichlet Allocation (LDA) [81] to learn topics. LDA is one of the most popular topic modeling methods to find topics from a corpus of natural language documents, such as conference papers, tweets, and online forum postings [82–84]. As shown

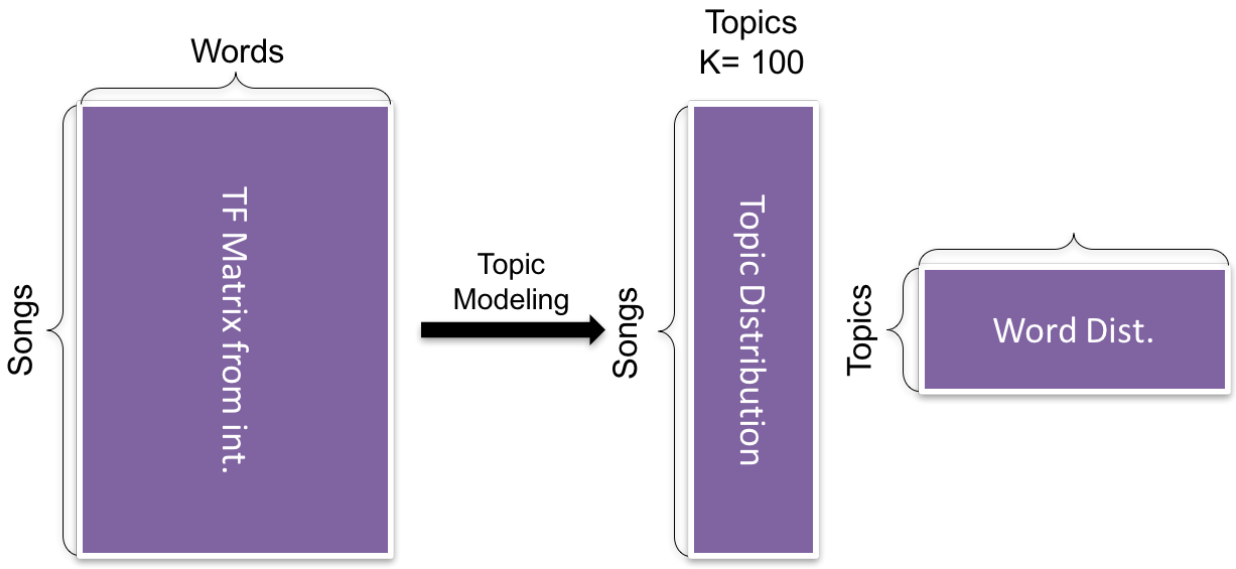

Figure 2.1: Applying topic modeling on a Term-Frequency (TF) matrix

in Figure 2.1, this generative model assumes that each document is represented as a multinomial distribution over topics, and each topic is represented as a multinomial distribution over words in the corpus.

Figure 2.2 shows a pictorial example of inference with LDA using an excerpt of the lyrics to Queen's "Bohemian Rhapsody." This example displays the two most active topics for the document: *Murder* and *Family*. Each topic is represented as a probability distribution of words; the five most active words of the *Murder* topic are *kill*, *gun*, *dead*, *killer*, and *crime*, and the top five words for the *Family* topic are *mother*, *father*, *brother*, *sibling*, and *mama*. LDA assumes that each word in the document is assigned to a specific topic. In this example, the red words in the lyrics are assigned to *Murder* and the blue word is assigned to *Family*.

Topic modeling has also been used to analyze topical trends of big data. For instance, there is a study that tracked the topical trend of computational linguistics conferences by looking at LDA topics found in the conference proceedings [82]. The topic distributions were grouped annually and the changes of the strength of specific topics of interest was measured. Similarly, there is a study that examined Stack Overflow, an online question and answering forum for developers, to analyze topics and trends [84]. Both studies show the effectiveness of topic modeling for discovering trends from big data corpora ranging from academic papers to online forum postings.

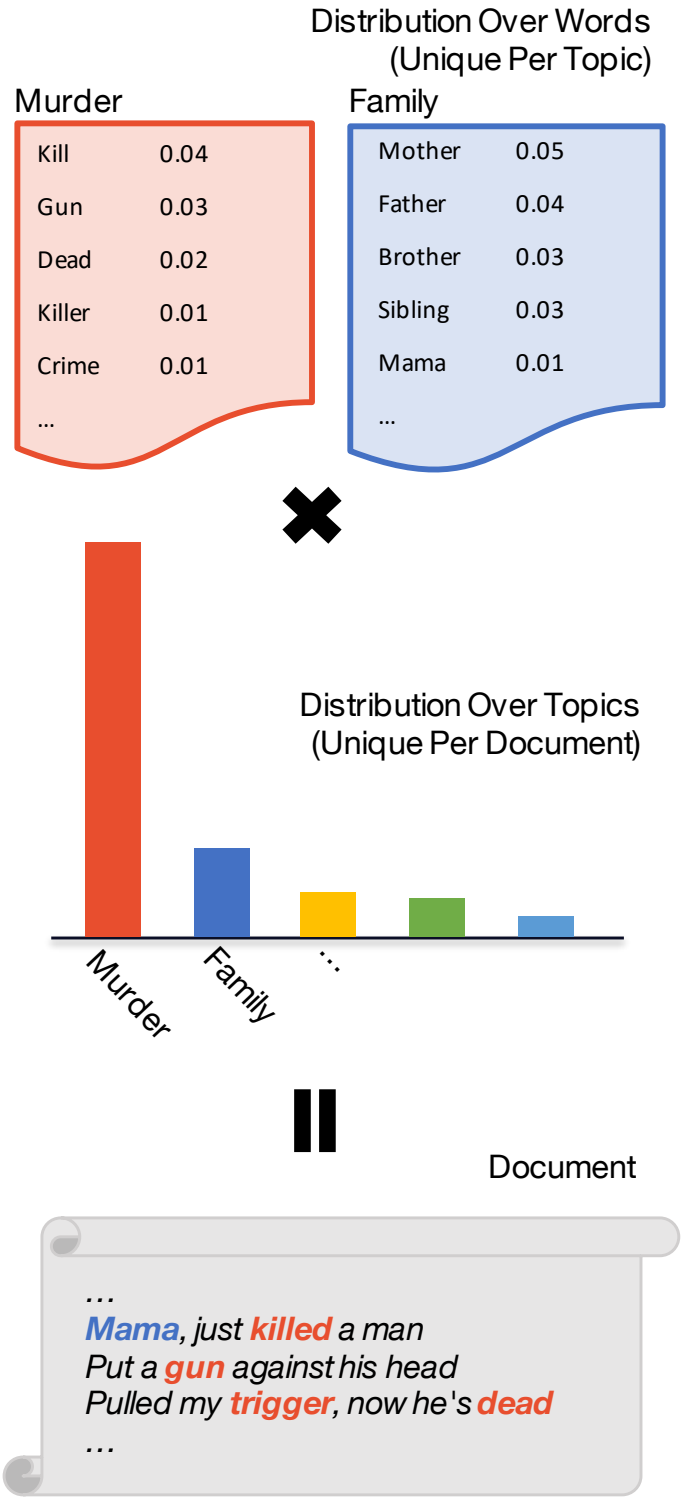

Figure 2.2: A pictorial example of inference with LDA. The document is an excerpt from the lyrics of Queen's "Bohemian Rhapsody"

2.3.2 Lyric Understanding and Topic Detection

Detecting major topics of song lyrics can be considered the first step of understanding them. If you understand the lyrics of a song, you can form an interpretation in your mind. If we look at the process of writing an interpretation from the perspective of topic modeling, which is a probabilistic generative model, the interpretation documents can be generated as follows:

1. Sample a word distribution per topic: $\phi_d \sim \text{Dir}(\beta)$, β is the Dirichlet parameter to control the sparsity of the word distribution
    (a) For the d-th interpretation document, sample a topic distribution $\theta_d \sim \text{Dir}(\alpha)$, where d indicates the documents and α is the Dirichlet parameter to control the sparsity of the topic distribution
        i. For the n-th word in the d-th document, sample a topic from the topic distribution: $z_{n,d} \sim \text{Mult}(\theta_d)$, where n indicates the word position
        ii. For the n-th word in the d-th document, sample a word from the word distribution associated with $z_{n,d}$-th topic: $w_{n,d} \sim \text{Mult}(\phi_{z_{n,d}})$

We repeat (i) and (ii) N_d times, which is the number of words in the d-th document.

Therefore, as shown in Figure 2.3, if an artist writes a song with a few main subjects LDA should, in theory, be able to find the topics that approximate the original ones. However, due to the artistic nature of the lyric generation process, inferring topics from the lyrics directly is often not obvious. On the other hand, if the user and the artist happen to share the same topics (i.e., the user has a good understanding about the lyrics), then applying LDA to the interpretation can be a better way to learn the topics.

2.3.3 Detecting Topics from Song Lyrics

There are few Music Information Retrieval (MIR) studies that used topic modeling algorithms to learn topics from song lyrics and other text in an unsupervised fashion. Kleedorfer et al. [6] introduced a method to index songs by their topics. They applied Non-negative Matrix Factorization (NMF) to 60,000 popular songs. NMF is another widely used topic modeling algorithm, which can be seen as a simpler version of LDA without an assumed Dirichlet priors about its parameters. In the study, six subjects were asked to assess the labels to the learned topics. A high degree of inter-rater agreement suggested that the learned topics were discernible. Later, Sasaki et al. [11] applied LDA to song lyrics

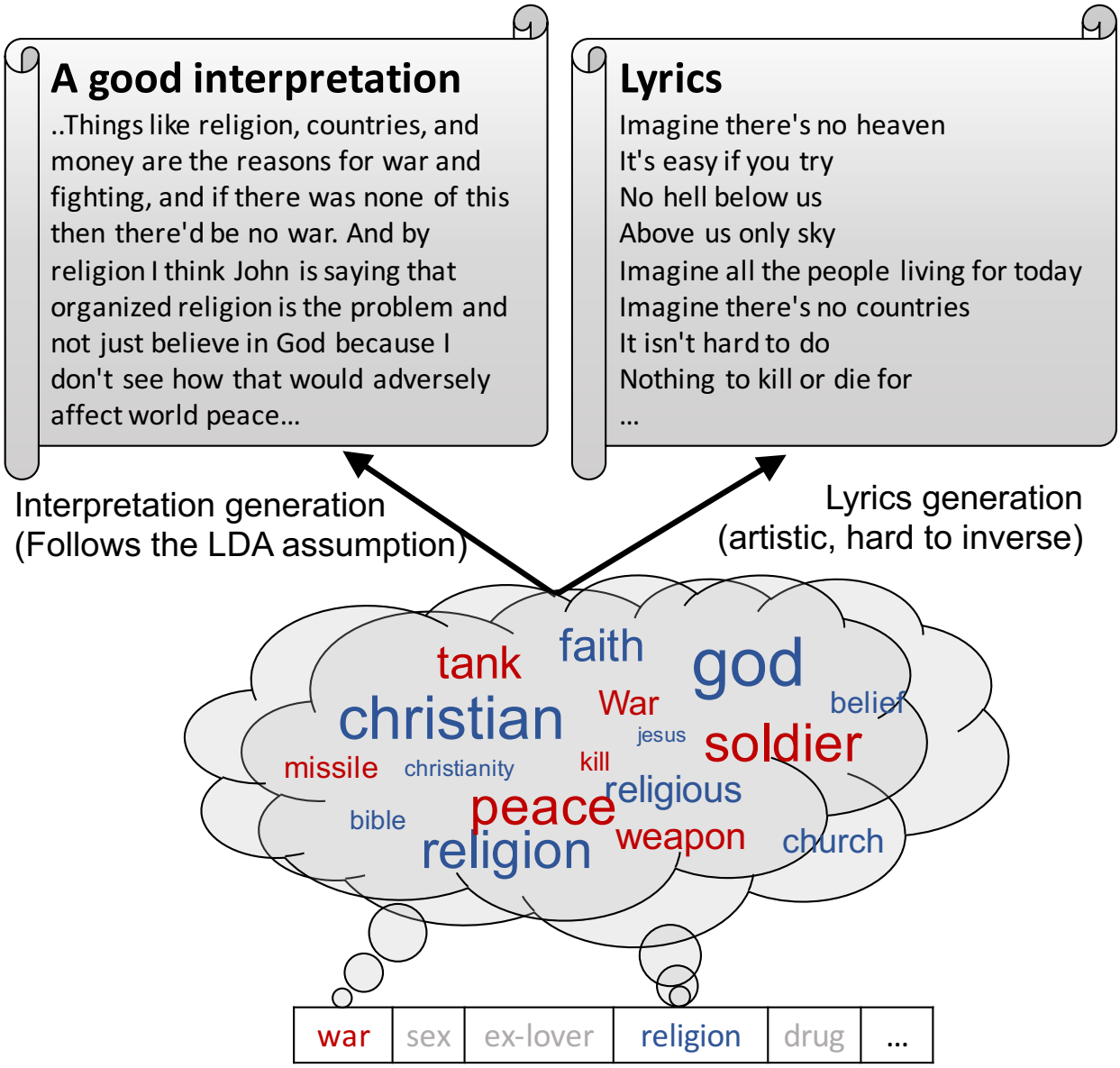

Figure 2.3: From topics to interpretations of song lyrics

to enable lyric browsing through a visual interface that displays the most salient topics. A study using 17 subjects reported that the interface was useful in finding similar song lyrics. Furthermore, Sterckx et al. [12] explored how to assign unsupervised LDA topics from song lyrics to topics acquired through a supervised way. Kurtosis was reported as the most effective aligning metric between the two types of topics. Finally, Choi et al. [8] applied topic modeling to song interpretations instead of song lyrics, since song interpretations tend to provide more straightforward topic-related information than ambiguous song lyrics. LDA's intrinsic and extrinsic evaluations suggested that it can learn meaningful topics from interpretations. The study also proposed an automatic topic filtering algorithm.

One of the first attempts to classify thematic categories of song lyrics by using a supervised method is the approach of Mahedero et al. [4]. The authors reported 82% accuracy when applying the Naïve Bayes classification algorithm to 125 songs with five categories: *Love*, *Violent*, *Protest* (*Anti-war*), *Christian*, and *Drugs*. Later, Choi et al.[7, 9] expanded this research to bigger datasets of 800 and 900 songs respectively, and included more subject categories, such as *Sex*, *Places*, *Ex-lover*, *Cheating*, and *Loneliness*. However, the focus of these papers was on determining which sources are more useful in classifying subject categories between lyrics and interpretations. Both studies used a bag-of-words representation, or Term Frequency matrix.

2.3.4 Detecting Topics from Interpretations

Lyrics are main sources of topic information, however, they tend to be ambiguous and difficult to understand, because they use metaphors and figurative languages in songs. Conversely, interpretations are written in prose rather than verse and often provide extra information using straightforward words that are helpful in understanding the song lyrics.

For this reason, in addition to lyric text, interpretations of lyrics were used to identify topics in previous studies. The experimental results in Table 2.1 shows that interpretations outperform lyrics in the task of song subject classification [9]. This study compared the most representative terms in song lyrics and their interpretations, and showed that interpretation terms tend to be richer. This might be because people tend to elaborate on a subject with their own words. In addition, terms that refer to abstract concepts such as *religion* and *relationship* are highly ranked only in interpretations. This might result from lyrics often using imagery to describe those concepts rather than mentioning them directly.

An unsupervised approach was taken in [8] to detect topics from song interpretations. Here, 100 LDA topics were learned from about 25,000 song interpretations collected from

Subjects	Lyrics	Interpre-tations	Concatenation	Late Fusion
Places	49.0 %	58.0 %	59.0 %	61.0%
Sex	65.0 %	70.0 %	75.0%	73.0 %
Ex-lover	36.0 %	67.0 %	67.0 %	68.0%
Drugs	36.0 %	69.0 %	71.0%	70.0 %
War	65.0 %	76.0 %	79.0%	79.0%
Parent	34.0 %	57.0 %	59.0 %	60.0%
Religion	35.0 %	70.0%	67.0 %	70.0%
Death	29.0 %	51.0%	51.0%	50.0 %
Average	43.6 %	64.8 %	66.0 %	66.4%

Table 2.1: Classification accuracy across categories (Choi et al. [9])

songmeaings.com. Meaningful topics were selected systematically based on topic weights and coherence of the top ten words of each topic (see Figure 2.4). The assumption behind this approach is that highly popular topics in this collection consist of collection-specific terms, while extremely rare topics are mostly outliers like topics with foreign words. Furthermore, since only discernible topics with coherent terms are useful, the Normalized Pointwise Mutual Information (NPMI) based on word co-occurrences in Wikipedia, was calculated to pick meaningful topics. It is reported that NPMI has been highly correlated with topic coherence assessed by human [85]. As shown in Table 2.3.4, eventually the algorithm was able to pick meaningful topics such as *M1:Relationship/Love* and *M4:Religion* and to exclude irrelevant topics such as those with collection-specific terms and incoherent topics.

2.4 Summary

Section 2.1 provided a general overview of music complexity research in MIR to explain how lyric complexity is a part of music complexity. Section 2.2 subsequently reviewed quantitative text complexity measures from classic readability studies and more recent readability studies to contextualize the research in Chapter 2. Not only quantitative dimensions of text complexity but also qualitative dimensions of text complexity were discussed in detail to emphasize the need for computational methods to utilize qualitative dimensions of text complexity. Section 2.3 summarized MIR research on automatic topic detection from song lyrics and their interpretations to contextualize the research in Chapters 4 and 5.

Selected Topics	Topic ID	Topic Weights	NPMI	Top Words (Top 10)
High Weight	H1	0.62	-0.07	awesome, yeah, cd, relate,kinda,lol, total, lot, rock, play
	H2	0.61	0.04	reference, refer, sense, verse, idea, obvious, probable, kind, lot, bit
	H3	0.6	0.05	interpretation,narrator, verse, experience, place, literal, sense, point, speaker, fact
Medium Weight	M1	0.32	0.06	relationship, break, feeling, work, long, leave, girl, hurt, situation, stay
	M2	0.17	0.16	child, father, mother, parent, family, son, dad, brother, daughter, kid
	M3	0.15	0.11	sex, sexual, girl, prostitute, lust, woman, dirty, sexy, whore, desire
	M4	0.15	0.24	god,christian, religion, faith, religious, church, belief, bible,christianity,jesus
	M5	0.13	0.11	drug, addiction, heroin, high, addict, smoke, cocaine, reference, refer, coke
	M6	0.12	0.1	war, fight, soldier, bush, bomb, country, battle, kill, army, military
Low Weight	L1	0.04	-0.09	green, river, rise,weezer, edge, lucky, pie, bye, buy, stuff
	L2	0.03	-0.19	la, ghost, holly, deaf, vulture, ear,bebot, yeah,gorillaz, bounce
	L3	0.01	0.27	la,spanish, el, en, lo, se,es, mi,una,por

Table 2.2: Selected topics from 100 learned topics (Choi et al. [8])

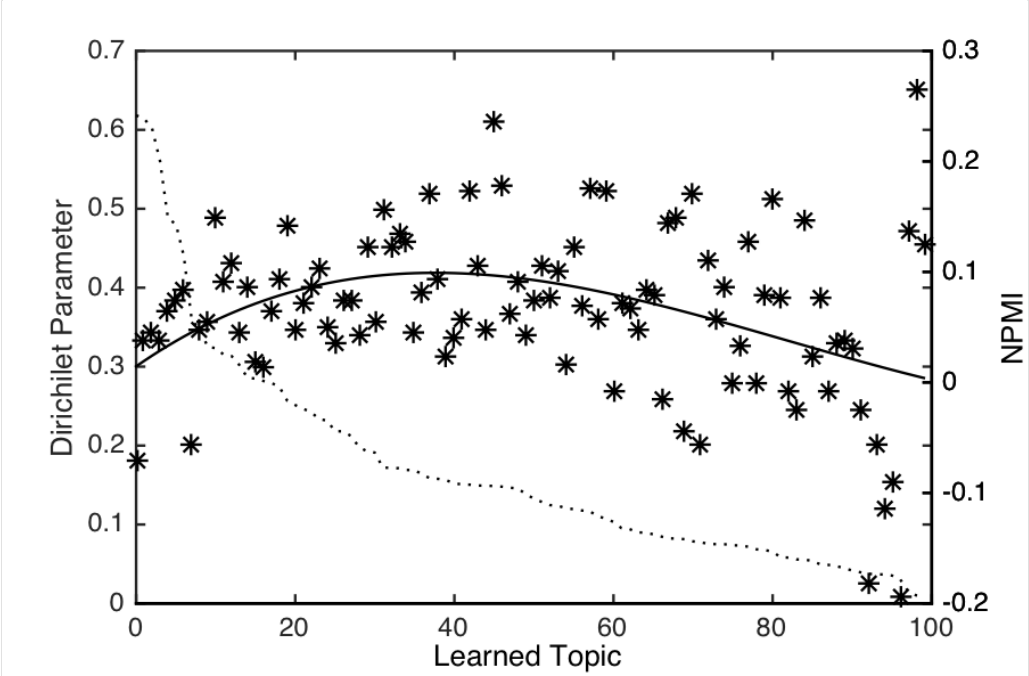

Figure 2.4: Prior topic weights (Dirichlet parameter) and NPMI values of LDA topics (k=100) (Choi et al. [8])

Chapter 3

Study 1: Concreteness of Words as a Lyric Complexity Metric

3.1 Introduction

Although the claim may be somewhat controversial, song lyrics can be considered literature [86]. 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 [87]. Among many different forms of literature, song lyrics are usually considered similar to poems [86] 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. As reviewed in Chapter 2, researchers have proposed various metrics based on a variety of linguistic variables that influence the level of text complexity. This dissertation is an early study of lyric complexity and focuses on variables, the individual elements of metrics. Particularly, word-level variables are of interest to this dissertation, since it is not clear where sentences in music lyrics begin and end. Chapter 2 reviewed the most popular word-level variables used by the complexity metrics chosen by CCSS: word frequency, word length, word familiarity, word grade level, Pearson Word Maturity Metric, and concreteness. Among these, this chapter pays particular attention to word concreteness.

Concrete words are those that refer to specific objects or remind of a particular

[1] *http://www.nobelprize.org*

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 [41, 88–91]. For these characteristics, word concreteness has been one of the most important criteria for analyzing text difficulty [31, 72, 91].

Among the many variables used in quantitative metrics to analyze text complexity, this chapter focuses on concreteness to examine text difficulty of popular song lyrics for the following reasons. First, no previous study explored concreteness of song lyrics, although word frequency was explored in MIR by Ellis et al. [13]. 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. Finally, concreteness ratings are publicly available and are important data for this chapter. Conversely, most of the other variables, such as word familiarity and Pearson Word Maturity Metric, do not have associated, non-proprietary data.

While this dissertation is the first to analyze historical trends in concreteness of song lyrics, 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.[41] conducted trend analysis of concreteness of four collections of English books and speeches (e.g. the Google Ngrams corpus of American English [92]; the Corpus of Historical American English [93]; and inaugural addresses by American presidents). Like this chapter, concreteness ratings of English word norms are obtained from the collection generated by Brysbaert et al. [40] 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.

Dodds et al.[94] 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 [95]. 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.

This chapter analyzes the concreteness of 5,500 popular song lyrics to seek to answer research question 1: "How has text complexity of popular song lyrics changed over time in terms of concreteness?" To better understand the trends, this chapter also aims to answer research question 1-1: "What is the relationship between the concreteness trends and genres?" and research question 1-2: "What is the relationship between the concreteness trends and word statistics in song lyrics?"

3.2 Experiment Design

3.2.1 Data

Music Collection

This chapter analyzed the lyrics of 5,100 songs from Billboard Year End hot 100 songs from between 1960 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 56 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 [96–98].

Lyrics

We obtained a reliable lyric corpus from LyricFind[2] which is a world-wide lyric licensing company, via a signed research agreement. Utilizing this lyrics dataset has many

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

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 [13], 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 [13]; Atherton and Kaneshiro analyzed lyrical influence networks by using intact lyrics [99]; and Tsaptsinos proposed an automatic music genre classification system by applying recurrent neural network models to song lyrics [100].

Metadata

This chapter examines relationships between concreteness of song lyrics and three types of metadata, including year, artist, and genre. The year and artist metadata were taken from the Billboard Year End chart. Genre metadata was collected by using iTunes Search API[3], 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*

Concreteness Ratings

Brysbaert et al.[40] 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 [60] and augmented by

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

various widely known corpora, such as the English Lexicon Project [101] and the British Lexicon Project [102]. 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 chapter.

3.2.2 Lyric Preprocessing

After retrieving song lyrics from the LyricFind corpus using titles and artists, the state-of-the-art technology, Stanford CoreNLP [103] 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.

3.2.3 Analysis Methods

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 kth word and f_k is its frequency.

$$v_{text} = \frac{\sum_k v_k f_k}{\sum_k f_k}. \tag{3.1}$$

The concreteness score for each year is the average concreteness scores of lyrics appeared in the chart of the year. Figure 3.1 shows how the overall concreteness score is calculated, using lyrics from an *Eminem* song as example.

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 [104]. 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.

Figure 3.1: A pictorial example of how the overall concreteness is calculated, using an excerpt of the lyrics from Eminem's "*Lose Yourself*"

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 are applied to determine whether trends are statistically significant.

To identify change points, *findchangepts* is employed, which is implemented in *Matlab* [105]. As this chapter is 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 [106]. 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, this chapter 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 [84], vitamin and mineral intake from fortified food[107], 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.

3.3 Results

3.3.1 General Trend

Figure 3.2 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 [41], 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, this chapter focused on three: proportion of genres, proportion of open/closed class words, and length of music lyrics.

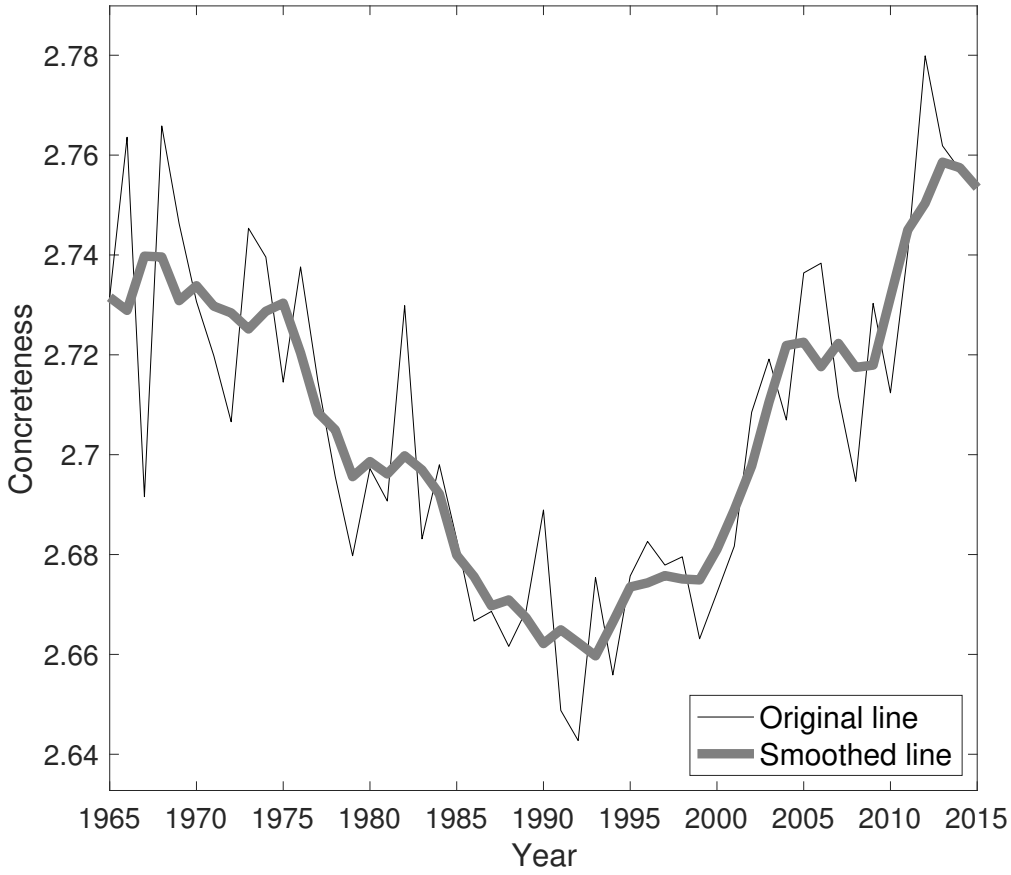

Figure 3.2: Concreteness time series for song lyrics

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

3.3.2 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, this chapter calculated counts and average concreteness scores of individual genre categories. Table 3.3.2 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 3.3.2 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.

For further analysis of the relationship between genres and concreteness over time, this chapter shows the time series of concreteness scores of some dominant genres (see in Figure 3.3). 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

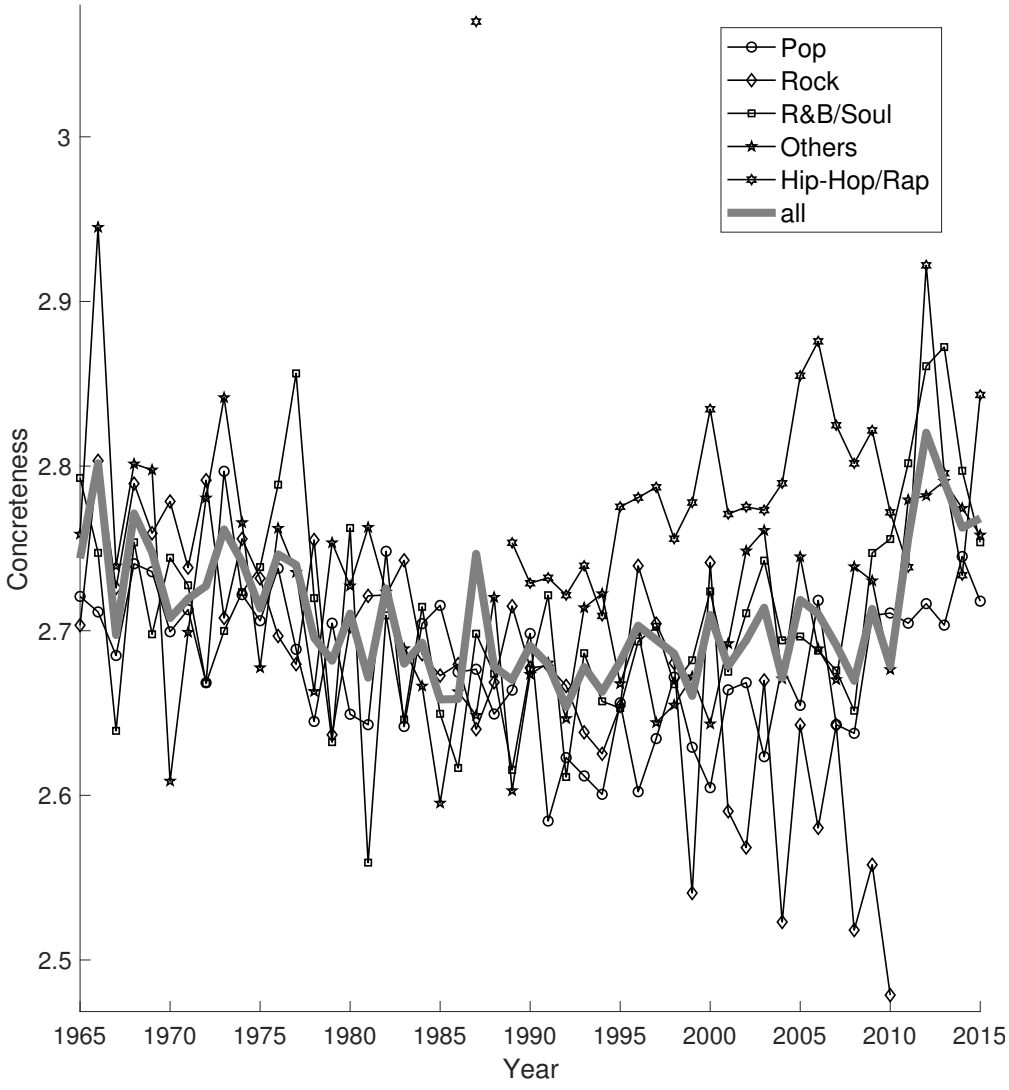

Figure 3.3: Concreteness time series for song lyrics broken down by major genres

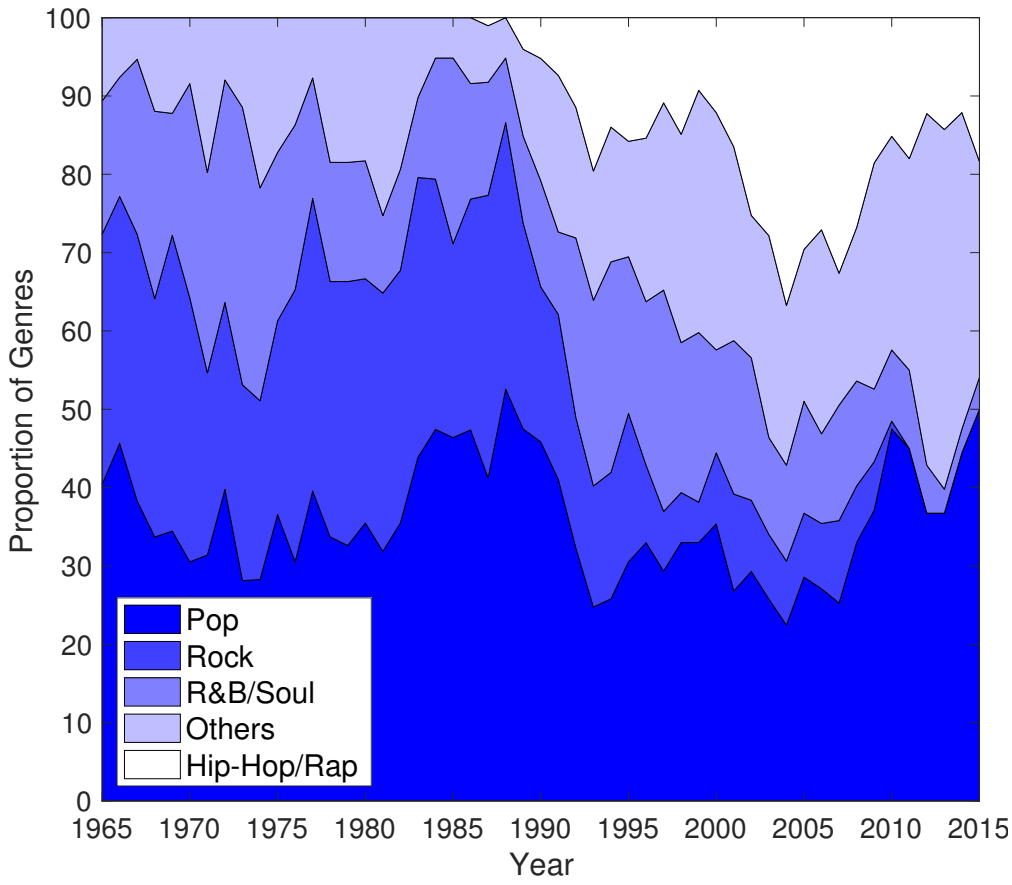

Figure 3.4: A stacked area plot of proportion of major genres

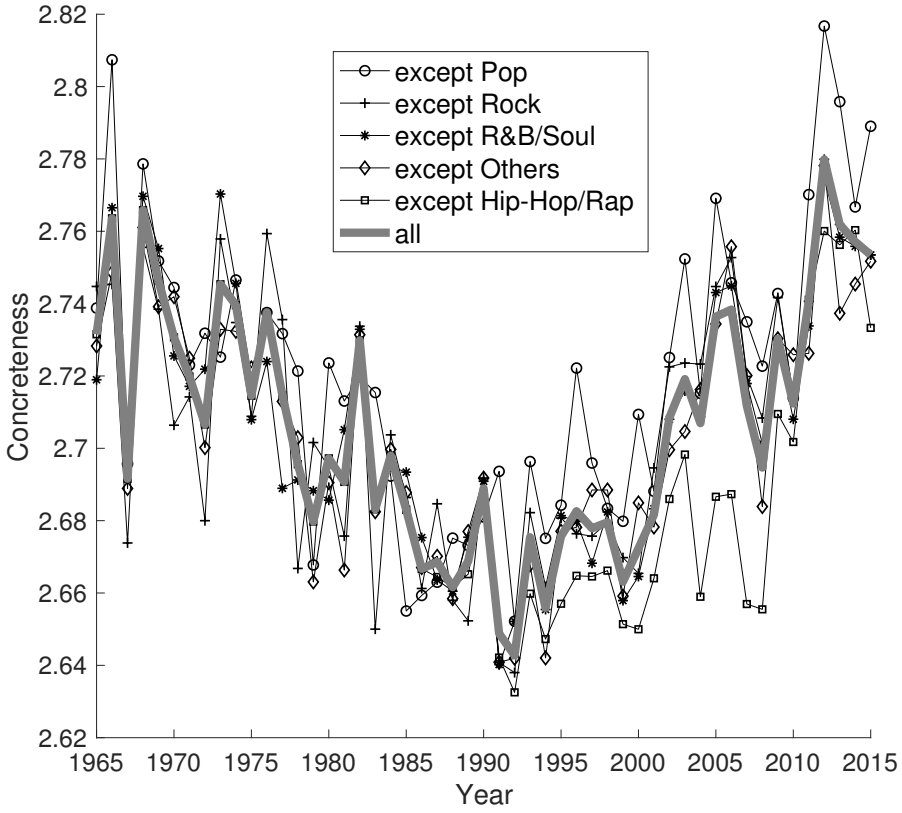

Figure 3.5: Concreteness time series for song lyrics except each major genre

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

also calculated, as shown in Figure 3.4. The noticeable change is the advent of *Hip-Hop/Rap* in the late 1980s, which *Hip-Hop/Rap* eventually overtook *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 3.5), 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.

3.3.3 Closed/Open Word Classes

Changes of shares of closed/open word classes over time may also influence the concreteness trend. In Table 3.3.3, Hills et al. [41] reported the average concreteness values and standard deviations of 11 word classes in the concreteness ratings from the data collection publicized by Brysbaert et al. [40]. 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 3.6 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

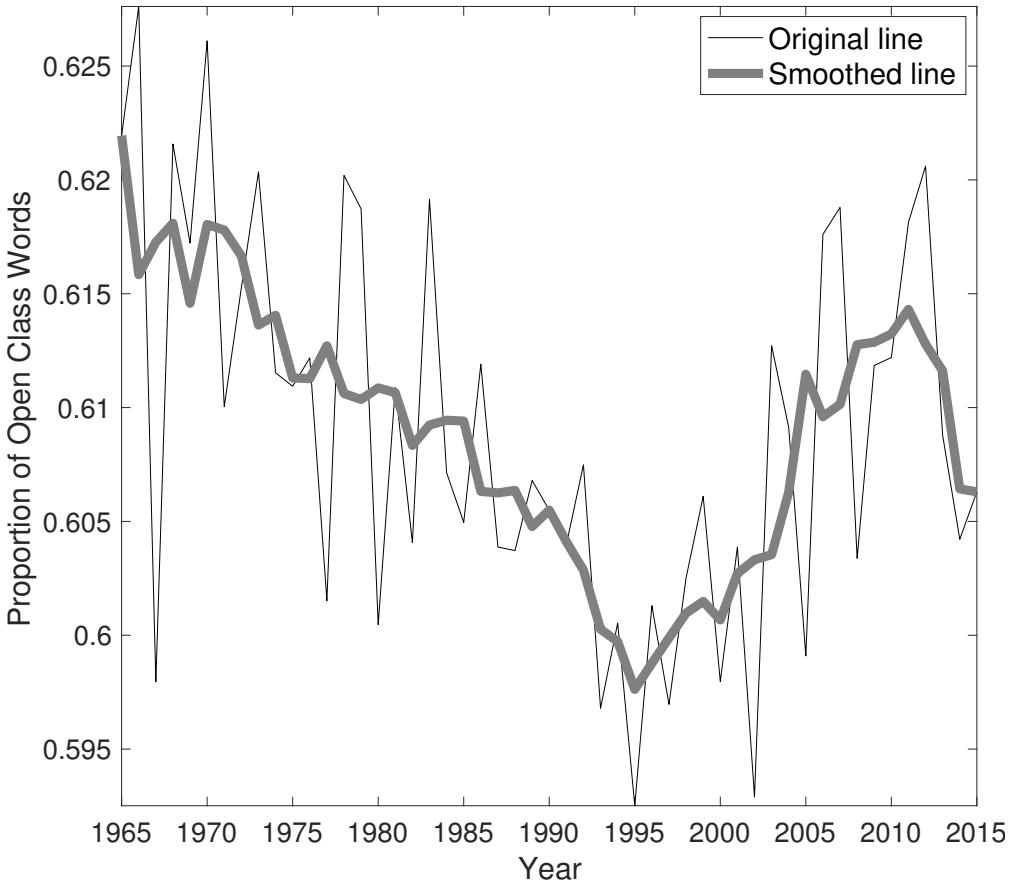

Figure 3.6: A portion of open word classes

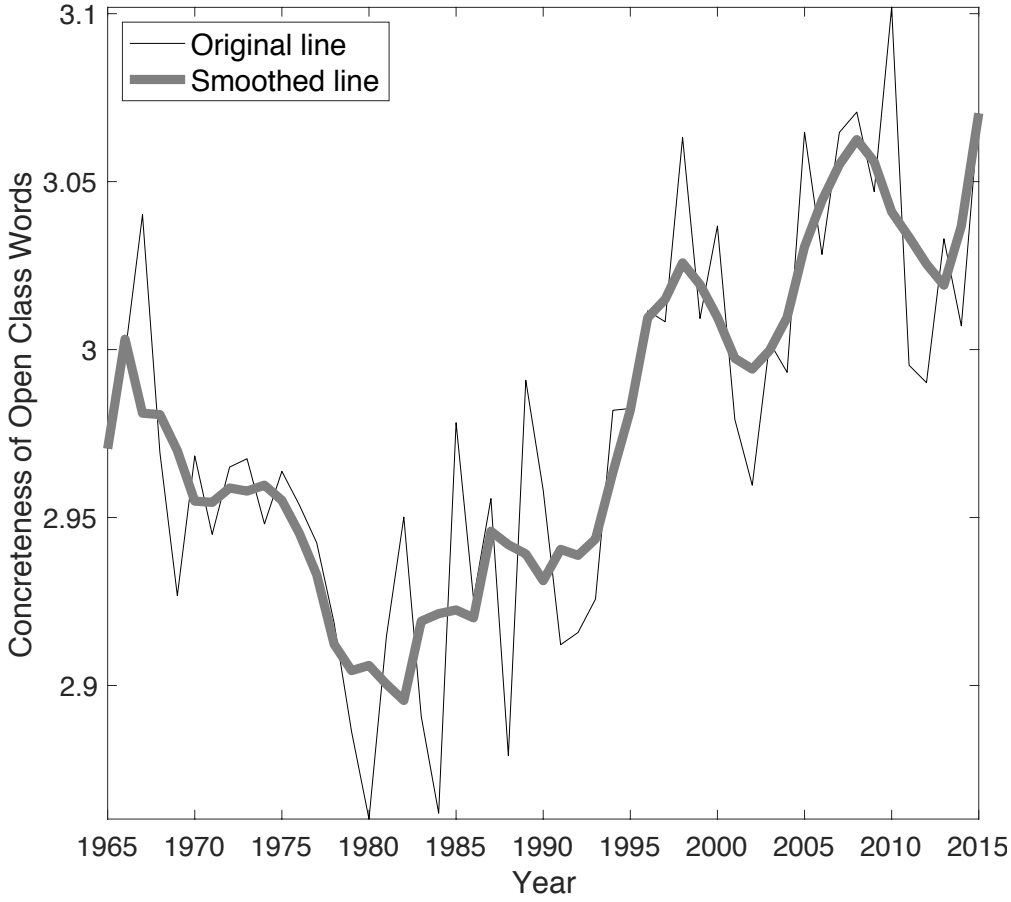

Figure 3.7: Concreteness time series for open word classes in song lyrics

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 3.3: Concreteness for a selection of word classes in the Brysbaert et al. (Hills et al. [41])

two. The trend of overall concreteness values of song 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 3.7 ($r = 0.6105$, $p < 0.001$). 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.

3.3.4 Word Length

This chapter also examines how basic linguistic characteristics, such as numbers of words in song lyrics, may relate to the overall concreteness trend. As shown in Figure 3.8 and Figure 3.9, 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: $r = -0.5971$, $p < 0.001$).

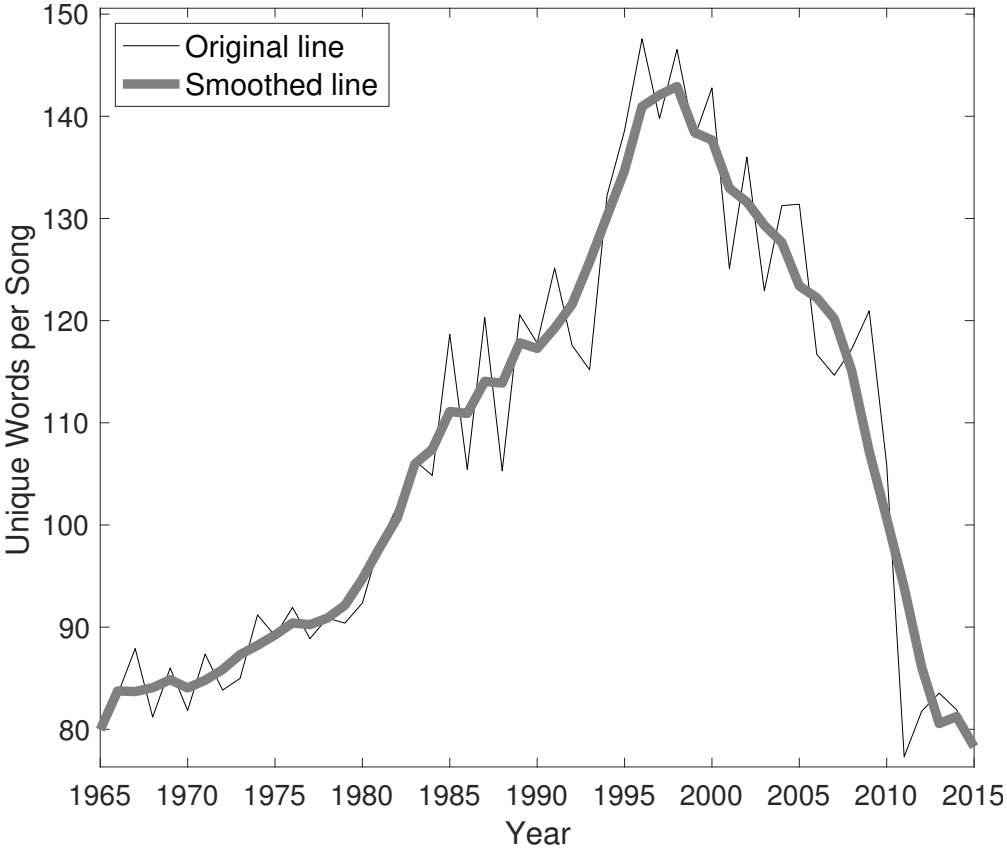

Figure 3.8: Average number of unique words in song lyrics

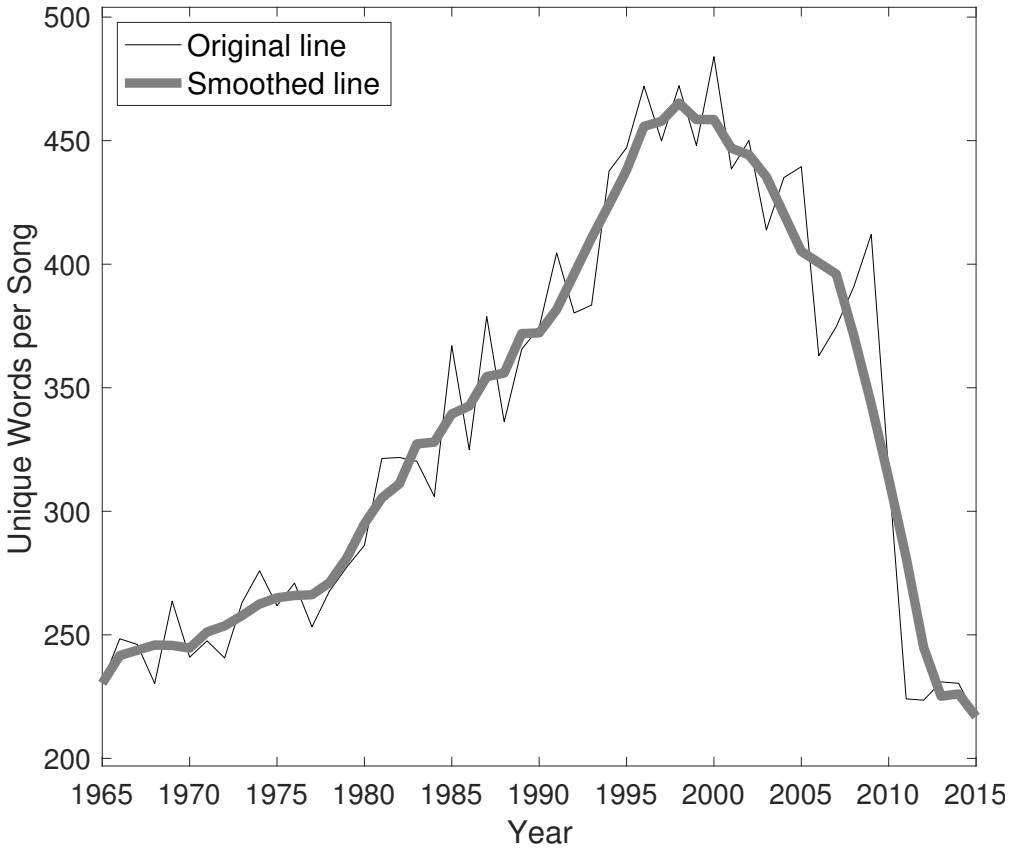

Figure 3.9: Average number of words in song lyrics

3.4 Discussion and Conclusion

This chapter identified concreteness trends in pop song lyrics between 1965 and 2015. The trends went down until the early 1990s and went up after that, which answers research question 1. The high correlation with "Hip-Hop/Rap" may explain the rise after the early 1990s (answering research question 2-1). 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 (answering research question 2-2).

Among the various quantitative dimensions of text complexity, this chapter paid particular attention to concreteness. Because concrete 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 3.10, 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. [98] 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.

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 3.11, the first three lines of Queen's "Bohemian Rhapsody" seem to be about murder. However, according to the songwriter's partner, the song, in fact, addresses his sexual identity

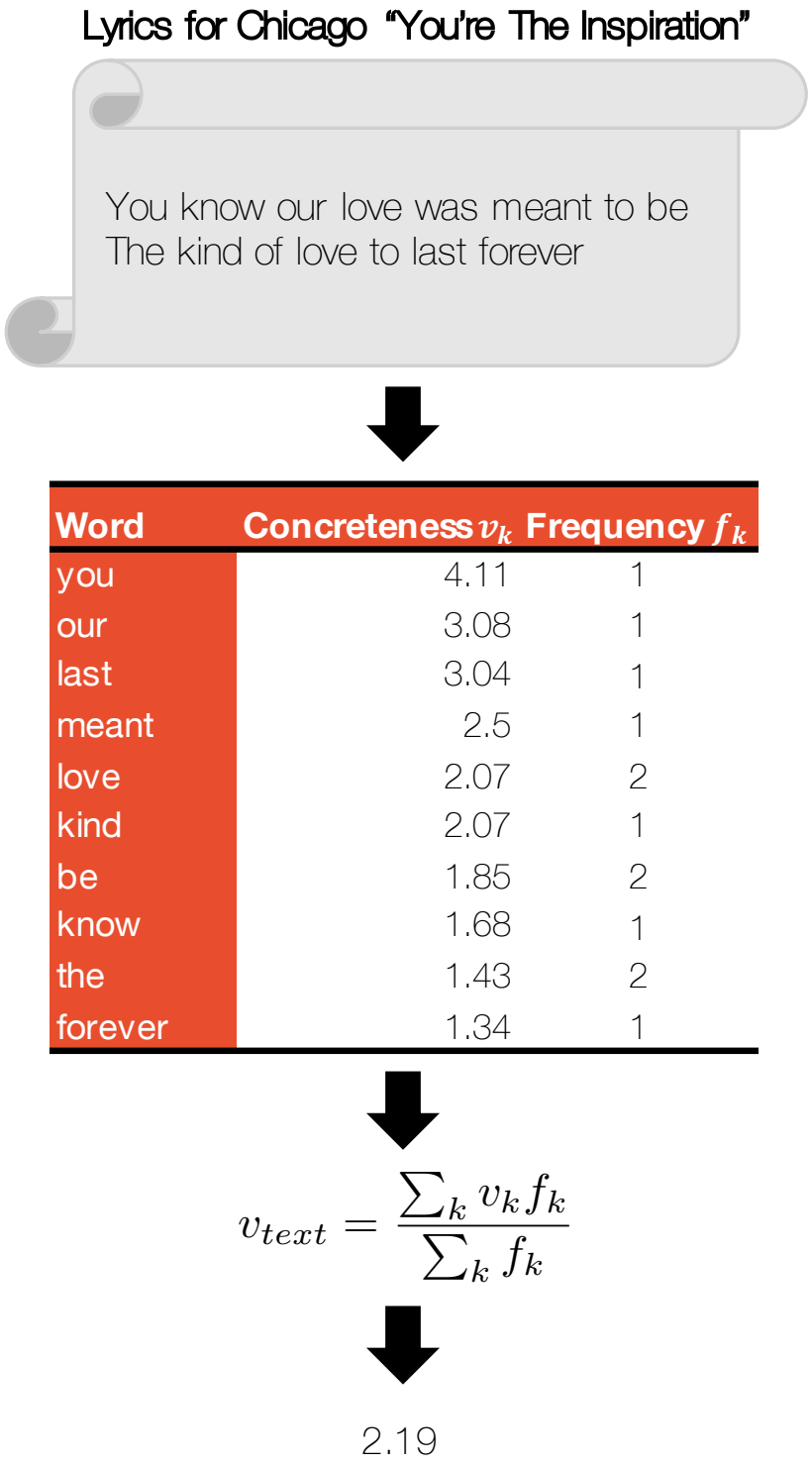

Figure 3.10: A pictorial example of how the overall concreteness is calculated by using the example of an excerpt of the lyrics of "You're the Inspiration" written by Chicago

Bohemian Rhapsody - Queen

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

Is this about murder?

Maybe not:
"Mercury's lover Jim Hutton suggested
the singer was confessing his bisexuality
– murdering his old image (sexuality)."

Figure 3.11: An excerpt from song lyrics of Queen's "Bohemian Rhapsody"

instead of murdering someone. 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.

3.5 Summary

This chapter 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.

Chapter 4

Study 2: Evaluating Usefulness of User-generated Interpretations

4.1 Introduction

This chapter investigates the assumption that user-generated interpretations of song lyrics found on the web are useful for inferring topics. It is an important assumption in this dissertation, because another lyric complexity method proposed in Chapter 5 heavily relies on the quality of the topics automatically extracted from the interpretations. Chapter 5 introduces this additional lyric complexity method because traditional quantitative text complexity metrics, such as concreteness (used in Chapter 3), cannot capture qualitative aspects of text complexity. Although various qualitative analyses can potentially complement the traditional quantitative metrics as listed in Section 2.2.4, it is not easy to conduct a large-scale qualitative analysis due to resource constraints. To this end, the proposed method in Chapter 5 utilizes user-generated interpretations from a website called *songmeanings.com* to grasp some qualitative dimensions of text complexity in a quantitative fashion. Although the users on the website do not explicitly comment on the level of lyric complexity, we hypothesize that their opinions about the meaning of song lyrics indirectly inform about lyric complexity. More specifically, we assume that the amount of agreement among users might be correlated with lyric complexity, operating under the belief that a variety of different interpretations are related to higher complexity. It is crucial to check the validity of the assumption that quantitative analysis of the interpretations can extract the topics of the song lyrics.

This chapter compares the interpretations and song lyrics as input to automatic lyric topic classification tasks based on the hypothesis that interpretations are written in more direct sentences, which are easier for a topic classifier to work on than the lyrics. As

mentioned in Section 2.3, this problem was first addressed by me and my colleagues [7, 9], where user interpretations outperformed song lyrics as a feature for this classification task. The comparisons between the most representative terms of each input showed that user interpretations contain richer topic-related information in a more straightforward form compared to ambiguous song lyrics. Evolving from our earlier work, this chapter uses a more advanced way to extract some higher-level features from both song lyrics and user interpretations. Instead of the primitive Term-Frequency (TF) representation that the topic classification systems in our previous work was based on, this study proposes to use a more advanced word embedding representation, *fastText* [108]. While the TF representation can effectively describe the simple statistics associated with the bag-of-words representation of a document, the independent and sparse nature of the elements of the TF vector requires the feature to be too high dimensional. Word embedding techniques can address this issue by learning a model that converts a word into a dense vector representation, which should subsequently be able to recover the other neighboring words [109]. In this way, the vector representation can encode the co-occurrence information of the words spread in the large corpus. This chapter thus employs *fastText* as the main representation of words to utilize the co-occurrence information among words rather than the TF representation used in our previous work [7, 9]. This representation also aligns with the probabilistic topic modeling experiments in Chapter 5, where the term co-occurrence is a fundamental criterion in defining topics.

This chapter aims to answer research question 2: "Can an automatic algorithm successfully identify underlying topics of song lyrics from user-generated interpretations?" The experiments are designed to answer this question by answering research question 2-1: "Are users' interpretations of song lyrics more useful than song lyrics for the topic classification task?" and research question 2-2: "How different are the most representative words of interpretations and lyrics?"

4.2 Experiment Design

4.2.1 Data

Lyric Topic Dataset

This study follows the same experimental setup with the same dataset used in our previous work [9], which also compared song lyrics and their interpretations as input to lyric topic classification systems. The dataset consists of 800 popular songs. This balanced dataset

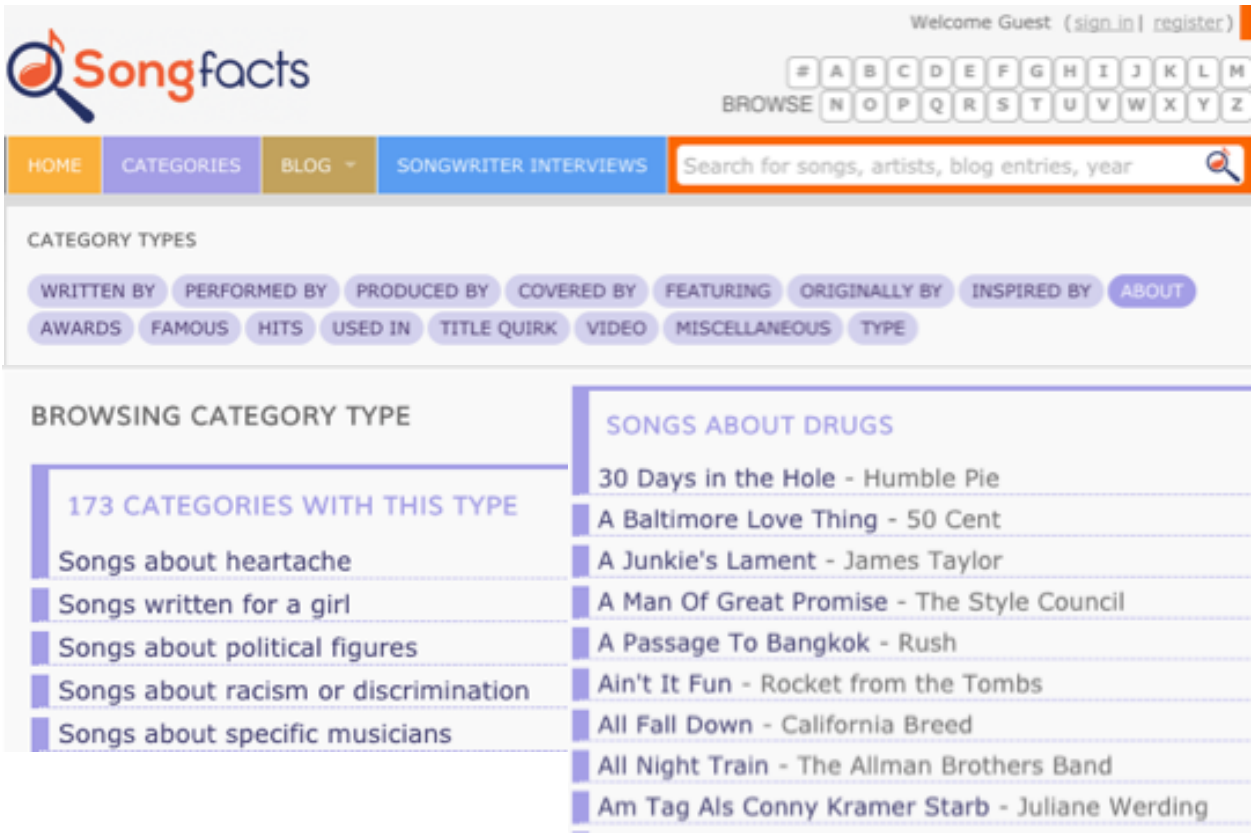

Figure 4.1: A screenshot image of the names of the first five categories of "about" category type and the song titles of the first nine "songs about drugs" on *songfacts.com*

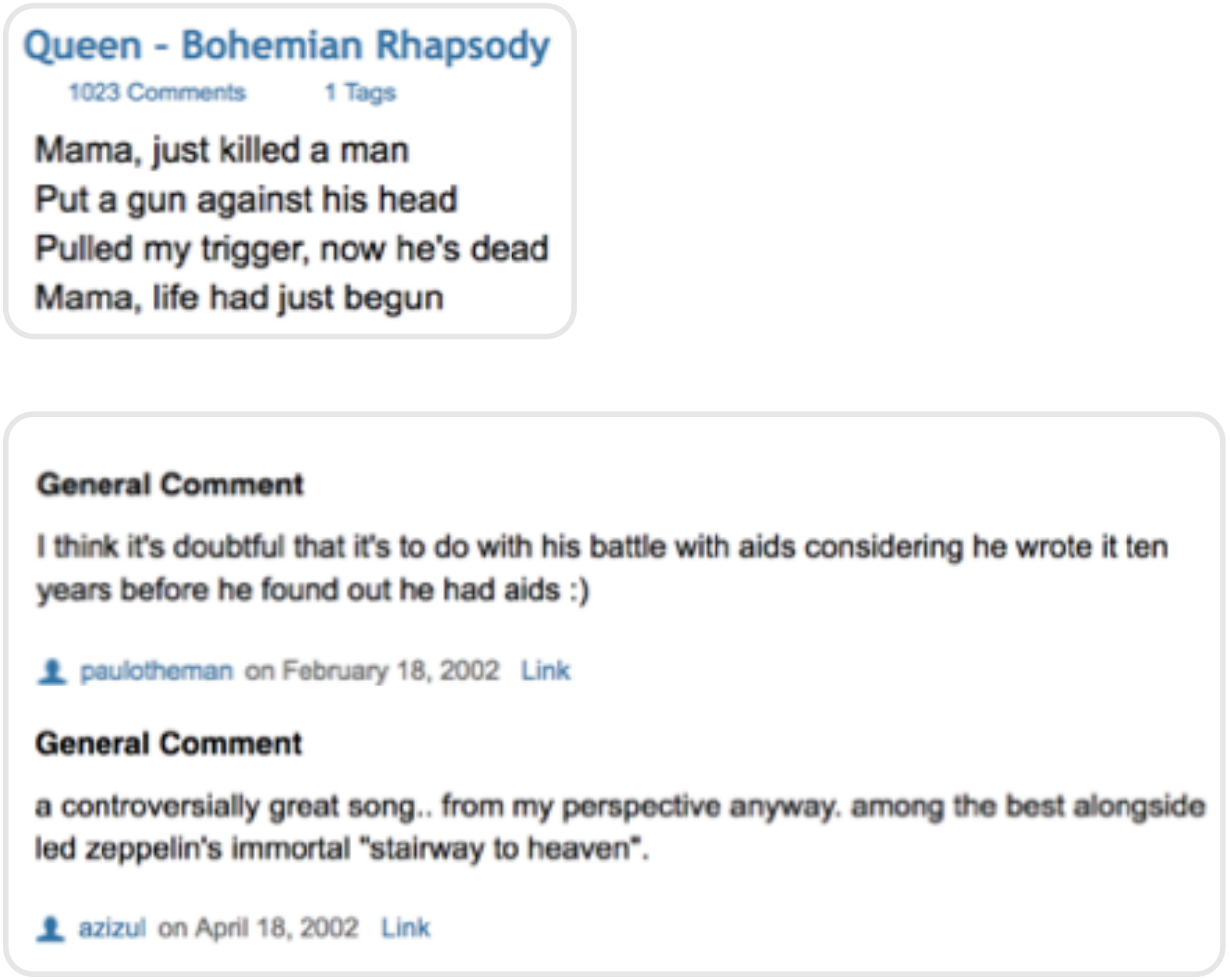

Figure 4.2: A screenshot image of an excerpt from lyrics of Queen's "Bohemian Rhapsody" and two users' interpretations of the lyrics in the form of comments on *songmeanings.com*

has eight topic categories, each of which has 100 songs. The topic labels are from *songfacts.com*, which categorizes their music collection into 173 subjects by experts, as shown in Figure 4.1. Among 173 subject categories, the eight most popular ones were selected: *Places*, *Sex*, *Ex-lover*, *Drugs*, *War*, *Parent*, *Religion*, and *Death*. Song lyrics and their user-generated interpretations were collected from *songmeanings.com*, where millions of users post their comments about the meanings of song lyrics, as shown in Figure 4.2.

Word Embedding

For word embedding, this chapter used an existing solution called *fastText* [108], which provides vector representations of a large vocabulary of words in English-language Wikipedia. The 300 dimensional *fastText* vectors are trained based on the skipgram model [109]. Because this skipgram model utilizes subword information, it shows better performance than others in multiple tasks including human similarity judgment and word analogy tasks than *word2vec* models. As an already trained neural network model, *fastText* predicts the 300-dimensional vector representation of a given word through its ordinary feedforward process.

4.2.2 Preprocessing

Song lyrics and user comments were lowercased and broken down into words. Punctuation marks were eliminated because they are irrelevant in this context. Next, common stop words[1] in English and infrequent words were removed because their discriminative power is low. Because words in *fastText* were not lemmatized, lemmatization was skipped in the preprocessing step.

4.2.3 Classification

For comparing lyrics and their interpretations as an input source to classification systems, the two inputs were fed to a supervised classification system respectively and together as shown in Figure 4.3. The combination of both input sources were also fed to the classifier to determine whether they have complementary relationship. Among many classifiers, we used naïve Bayes for this preliminary since it performs almost as well as Support Vector Machines (SVM) in our previous study [9], while it is faster than SVM. Considering the small size of the dataset, we excluded deep learning-based classification algorithms,

[1] https://code.google.com/archive/p/stop-words/

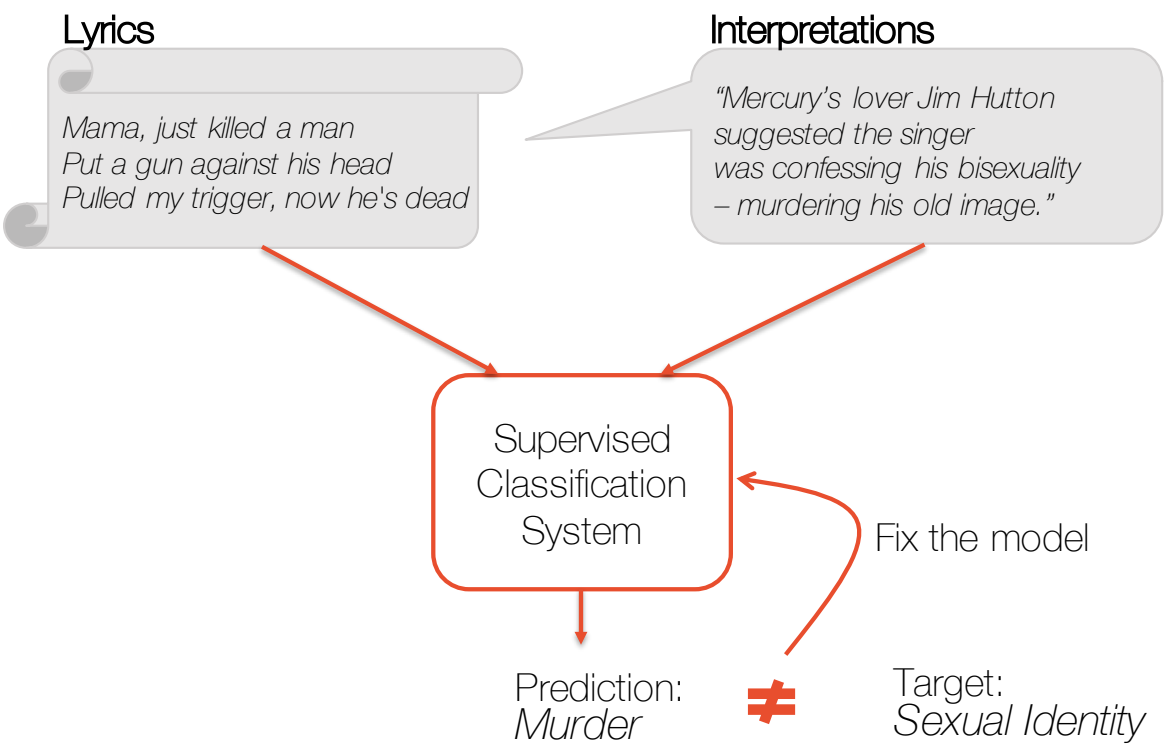

Figure 4.3: A pictorial example of the training process of classification systems

although we believe that they could perform better than other classifiers once a bigger dataset is available for training. The average accuracy was reported by using 10-fold cross validation.

4.2.4 Feature Representations

Since *fastText* produces a vector per word, a document, (i.e., song lyrics or a user comment) is represented with a sequence of vectors in which the length varies depending on the number of words. Instead of treating this sequential input data as it is, this study uses their weighted average as the representative word vector, where the weights are proportional to the frequency of the words. For comparison, this study also reports a case where the frequency of the words was not considered. This proposed feature aggregation is based on the observation that averaging word vectors work reasonably well when classifying short texts such as tweets, news headlines, and text messages [110].

Input Texts	TF	Weighted Average of Word Vectors	Average of Unique Word Vectors
Lyrics	39.37%	36.37%	37.50%
Interpretations	60.25%	50.37%	41.00%
Combinations	60.75%	57.88%	51.13%

Table 4.1: Classification accuracy across all input sources and feature representations

4.2.5 Evaluation Method

Classification accuracy is used as a performance measure in this chapter because the lyric topic dataset is balanced in that each topic category has the same number of songs. Accuracy is not a proper performance measure of classifiers with an imbalanced dataset because of the accuracy paradox, which refers to the case where high accuracy is achieved by favoring the most dominant classes [111, 112]. By using the balanced dataset, the accuracy paradox was avoided.

4.3 Results

4.3.1 Classification Accuracy

Table 4.1 compares the average accuracy of classification systems with different feature representations across multiple text inputs: 1) lyrics; 2) interpretations; and 3) the combination of the two. The classifier based on term frequency yielded the highest accuracy across all three input texts. On the one hand, when the word vectors were averaged while considering the frequency of each word, the classifier performed well with only an accuracy difference of 3% between the word embedding representation and term frequency representation. On the other hand, when the frequency of words was ignored, the classification accuracy dropped to 52%. This indicates that the frequency of words still plays a role in topic classification. As for the text inputs, the classification results from the combination of lyrics and interpretations outperformed the cases with interpretations or lyrics, across all the feature representations, which is consistent with the findings from our previous study [9].

Table 4.2 provides detailed accuracy information of the classification systems by using different input texts across all the subject categories. Particularly, the classification systems with all cases of input in Table 4.2 used the frequency-weighted average of word embedding representation. The Friedman's ANOVA test was applied to determine whether

Subjects	Lyrics	Interpretations	Combinations
Places	46.00%	36.00%	59.00%
Sex	53.00%	48.00%	57.00%
Ex-lover	43.00%	41.00%	63.00%
Drugs	19.00%	62.00%	54.00%
War	67.00%	70.00%	69.00%
Parent	16.00%	48.00%	45.00%
Religion	26.00%	64.00%	70.00%
Death	21.00%	34.00%	46.00%
Average	36.38%	50.38%	57.88%

Table 4.2: Classification accuracy across all subject categories

there are statistically significant differences between the two classification systems using the two different input texts. This test was chosen since the accuracy data do not follow the normal distribution due to the small size of samples [113]. The statistical test shows that the classification results using interpretations statistically outperformed the cases with lyrics ($p < 0.05$), although lyrics yielded better results in three categories. *War* is the most accurately classified category across all types of input. On the other hand, *Parent* and *Death* are the two most challenging subjects in all cases of input. In general, with some exceptions, interpretations are more useful input for the topic classification systems than song lyrics are; particularly, interpretations are much more useful than lyrics for *Religion* and *Drugs* categories. Contrarily, when it comes to classifying *Sex* and *Places*, lyrics led to higher performance than interpretations. Overall, adding interpretations to lyrics led to performance improvements in all topic cases.

4.3.2 Analysis of Confusion Matrices

To further examine which topic categories are often misclassified when different feature representations were used, confusion matrices of the classifiers that take the combination of lyrics and their interpretations were analyzed because they yielded the highest accuracy. Figure 4.4 is the confusion matrix of the classification system that takes the combination of lyrics and interpretations as input data and uses the term frequency-based feature representation. In contrast, Figure 4.5 shows the confusion matrix of the classification system that takes the combination of the two input sources by frequency-weighted averaging word vectors. First, the confusion matrices show how the categories are confused with each other. *Sex* and *Ex-lover* are the most confused pair of categories in Figure 4.5

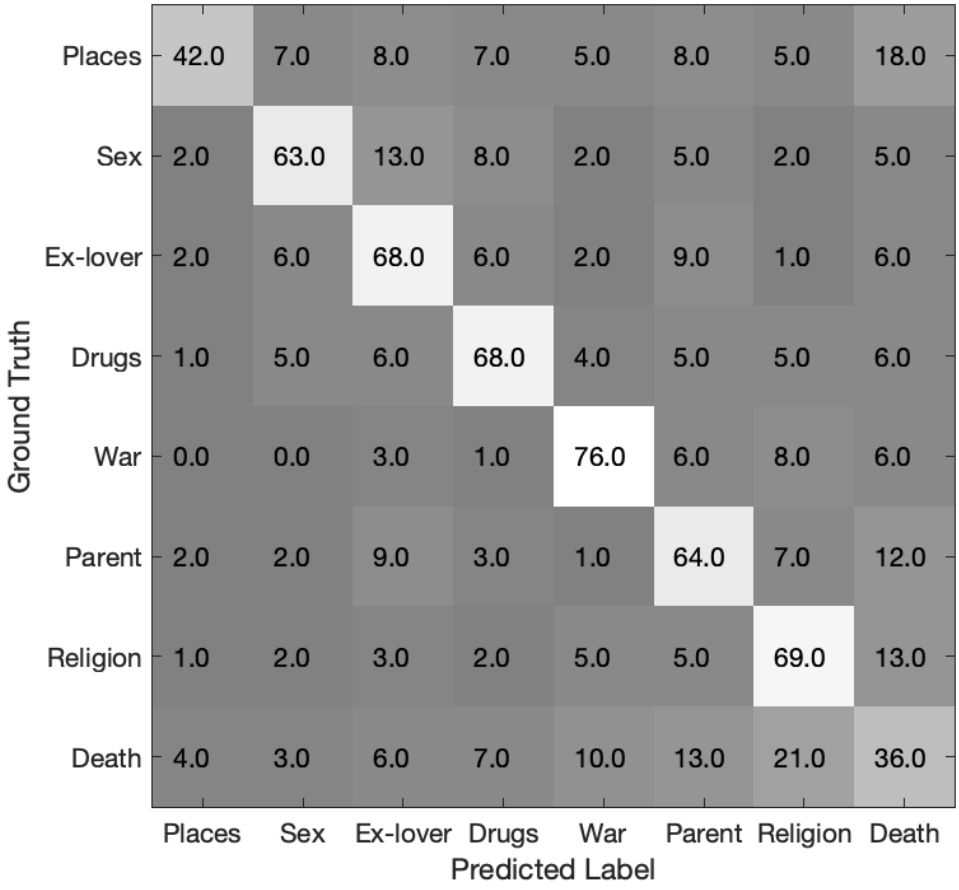

Figure 4.4: Confusion matrix of the classifier with TF

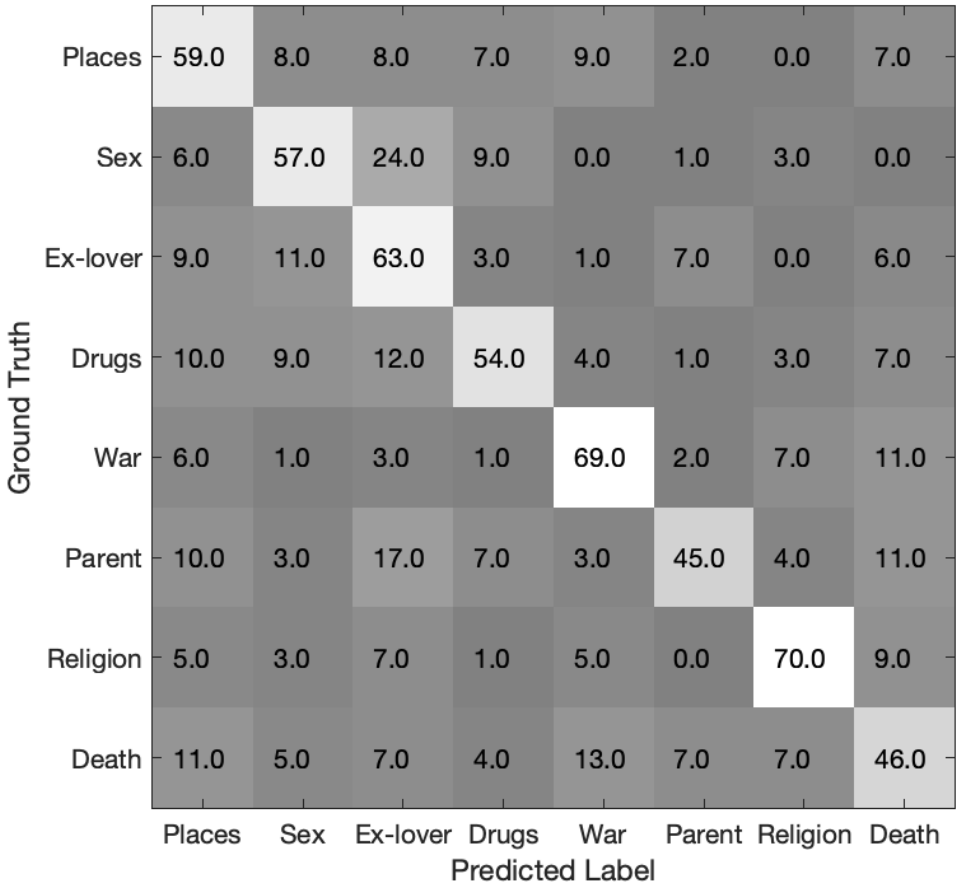

Figure 4.5: Confusion matrix of the classifier with averaged word vectors where the frequency of words was allowed

and one of the most confused in Figure 4.4: 24% of *Sex* songs were misclassified as *Ex-lover* in Figure 4.4, and 13% of *Sex* songs were misclassified as *Ex-lover* in Figure 4.5. It is not surprising because the two subjects are closely related to each other. The most confused pair in Figure 4.4 is *Death* and *Religion*: 21% of songs about *Death* are misclassified as *Religion*. The fact that only 7% of songs about *Death* are misclassified as *Religion* in Figure 4.5, however, indicates that the term frequency-based feature representation might cause the confusion. It was observed that the classifiers predict a few categories more often than others. The Term Frequency-based classification system prefers *Ex-lover* (116%), *Parents* (115%), and *Religion* (118%), while the word embedding-based system prefers *Ex-lover* (141%) and *Places* (116%). *Ex-lover* is favored by both of the systems. The classification system based on term frequencies heavily penalized *places* (54%) (Figure 4.4), while one based on word vectors favors the category (Figure 4.5).

4.3.3 Representative Words of Subject Categories

To identify the most representative words in each category, the top 20 words from each of the eight categories in each of the two input sources were studied: interpretations (Table 4.3) and lyrics (Table 4.4). As word vectors have continuous values, naïve Bayes assumes that data follow the Gaussian distribution. Therefore, mean and variance are available for each category after training the classifiers. The representative words for each category refer to those that are the closest to the mean of Gaussian distribution for each category. The words in the tables are ranked in order of the Euclidean distance from the mean of each Gaussian model for each category.

We can see that the top 20 descriptive words for each category seem to be reasonably related with each other and the categories. Particularly good are the first ranked words in Table 4.3, collected from interpretations that describe topics directly (e.g., *sexy* for *Sex*, *love* for *Ex-lover*, *drug* for *Drugs*, *war* for *War*, *father and mother* for *Parents*, *god* for *Religion*, and *death* for *Death*). The rest of the words in the top 20 descriptive words are relevant to each category. For instance, most of the words for the *Places* category are descriptive terms for places and names of areas (e.g., *riverside*, *city*, and *california*). Similarly, the *Drugs* category contains different types of drugs (e.g., *cocaine*, *heroin*, and *marijuana*). However, sometimes the most highly ranked terms in Table 4.4 based on song lyrics are not the most relevant terms for each category (e.g., *bitch* for *Sex*, *guess* for *Ex-lover*, *stasis* for *Death*) We can also see the song lyrics convey topics with indirect words in the case of the *Death* category. While the highest ranked term in interpretations is *death*, it is *stasis* in lyrics.

Places	Sex	Ex-lover	Drugs	War	Parent	Religion	Death
riverside	sexy	love	drug	war	father	god	death
city	bitch	loves	cocaine	civilians	mother	scripture	dying
california	wanna	jealous	drugs	troops	grandmother	divine	dies
york	funky	feelings	heroin	army	parents	believers	died
park	disco	flirting	marijuana	soldiers	grandfather	worship	buried
coast	bitches	unrequited	narcotics	armies	daughter	teachings	succumbed
boston	trippin	heartbroken	cannabis	casualties	stepfather	christianity	funeral
san	rihanna	infatuation	nicotine	military	son	religion	survived
angeles	somethin	feels	medication	forces	remarried	deity	demise
orleans	fuckin	boyfriend	addiction	wwii	wife	religions	illness
clearwater	catchy	girlfriend	addicts	battles	uncle	christ	mourn
los	fuck	liking	alcohol	wars	daughters	belief	surviving
vancouver	britney	infatuated	antidepressants	warfare	brother	christians	deathbed
located	rap	thinks	meth	armed	cousin	yahweh	corpse
chicago	iggy	worries	overdose	invasion	grandson	prophets	deceased
southern	dudes	boyfriends	pills	combat	childhood	beliefs	dead
detroit	gettin	feeling	addictions	civilian	fiance	religious	suicide
francisco	gotta	kisses	lsd	atrocities	fiancee	omnipotent	fate
bay	outkast	affection	hashish	wwi	dad	gods	alive
north	song	pretends	addictive	allies	heartbroken	biblical	sudden

Table 4.3: Top 20 words from interpretations' mean of the Gaussian model for each category

Places	Sex	Ex-lover	Drugs	War	Parent	Religion	Death
city	bitch	guess	cocaine	soldiers	father	god	stasis
riverside	wanna	dunno	heroin	casualties	parents	divine	remain
town	gotta	messed	drugs	war	mother	gods	reflections
beach	sexy	kinda	marijuana	troops	grandchildren	christ	dying
bay	daddy	admit	drug	destruction	cared	salvation	recede
fairgrounds	somethin	apologize	nicotine	forces	son	spiritual	disappear
coast	gonna	bother	medication	army	childhood	believers	shadows
skyline	poppin	feel	painkillers	atrocities	uncle	resurrection	life
promenade	shawty	worry	pills	killed	daughter	prophets	darkness
quay	fuckin	wondering	meth	armed	youngest	eternal	survive
park	trippin	everytime	aspirin	wounded	stepfather	faithfulness	entombment
boardwalk	bitches	guessing	addicted	prisoners	remembers	divinity	ascend
montego	hey	yeah	addiction	executions	unhappy	heavenly	beneath
street	feelin	oops	overdose	invasion	divorced	prayers	reflection
orleans	gettin	forgot	crack	perished	brother	sanctified	dwell
north	sassy	pretend	bag	destroyed	wife	sinned	fading
bungalows	momma	wouldn	choking	battle	loved	holy	descending
towns	baby	suppose	anaesthetic	allied	cousin	jesus	fragmented
suburbs	funky	realize	alcohol	allies	friend	prophecy	uncertainty
york	nigga	alright	itching	siege	disappointed	humankind	sunlight

Table 4.4: Top 20 words from lyrics' mean of the Gaussian model for each category

Moreover, many words are not tightly related to subject categories for *Sex* and *Ex-lover*. Overall, the top ranked words in interpretations are generally more obviously related to the categories, whereas the words in lyrics are less related to the categories.

4.4 Discussion and Conclusion

We can observe that user-generated interpretations outperform song lyrics as input to lyric topic classifiers by big margins, which is a positive answer to research question 2-1. This finding was consistently observed in both cases of feature representations: TF and word embedding. The analysis of the top 20 words of interpretations and song lyrics reveals that the representative words from the interpretations tend to be more straightforward than those from the lyrics. This finding is an answer to research question 2-2. Therefore, we can conclude that interpretations collected on the web can provide quality topic information as input to automatic topic classification systems. In general, as the empirical results in this chapter answer research question 2 positively, it is reasonable to believe that another automatic topic analysis algorithm (i.e., Latent Dirichlet Allocation) can also benefit from the same interpretation dataset for its probabilistic topic modeling.

There are a few other issues to discuss. First, introducing word embedding was motivated by positive experiences of other researchers with regard to its theoretical advantages; however, it does not improve the performance of the classification systems as we hypothesized. One reason would be because the frequency-weighted averaging strategy, which consolidated all the word vectors into a single vector, was not sophisticated enough, although such a frequency-weighted averaging method has been reported to well represent short texts. More advanced ways to describe documents, such as taking sequences into account and applying *doc2vec*, may improve the classification performance.

It is also noteworthy that the dataset and experimental setup are not realistic enough because (a) the numbers of topics are limited to eight; (b) the lyrics for each song can have only one topic; and (c) the dataset was balanced among the class labels. In the real world, however, (a) there can be potentially infinitely many topics in song lyrics; (b) the lyrics for one song can contain multiple topics; and (c) some topics are more popular than others. These clearly limit the applicability of this topic classification system if it were meant for the real world application scenarios, e.g. estimating topics of a new song. However, the experiments in this chapter had to control all the other experimental setups to be the same, because the main purpose of these experiments was to evaluate the usefulness of the interpretation over song lyrics themselves. Indeed, the empirical results showed that the

interpretations tend to outperform the lyrics.

4.5 Summary

This chapter compared song lyrics and their user-generated interpretations collected on the web as input for lyric topic classification systems to determine whether user-generated interpretations can be a good source for automatic topic detection algorithms. The assessment of the quality of the topic information automatically extracted from user-generated interpretations is important because the method proposed in the following chapter assumes that different topics from the interpretation data set are readily available for further processing. This work expands our previous work by using word embedding as a feature representation to directly utilize the semantic relationship between words-found from the co-occurrence information in a large-scale collection-and to ensure that the interpretation data contains the topic information in a simpler form.

The experiments showed that user-generated interpretations outperformed lyrics in the classification task by 14% when the frequency-weighted average strategy was used. This finding suggests that the interpretations may be a better input source for automatic topic detection algorithms than song lyrics. The comparison of the top 20 words for each topic between the two input sources confirms that the language used in the interpretations is more straightforward and closely related to each topic.

Chapter 5

Study 3: Topical Diversity of Interpretations as a Lyric Complexity Metric

5.1 Introduction

Although direct text complexity measurements, such as word length, sentence length, and lexical novelty, are always viable options, they can often miss some internal and external factors that contribute to lyric text complexity. These internal factors, such as many layers of meaning, call for manual qualitative analysis because current natural language processing technologies cannot decipher complex meanings. Luckily, users' interpretations on song lyrics provide such qualitative data. As for the external factors, they cannot be easily captured by simply analyzing lyrics because social context and the author's intention are usually not included in lyrics. On the contrary to song lyrics, users' interpretations of them posted on the web tend to have such information. The experiments in Chapter 4 revealed that user-generated interpretations collected on *songmeanings.com* are good sources for automatic topic detection algorithms because they convey topic-related information with straightforward words.

To this end, this chapter proposes a Lyric Topic Diversity Score (LTDS), a method that can quantify the level of difficulty by analyzing the users' understanding, which is readily available in web communities. *Based on the assumption that lyrics with a complex meaning allow diverse interpretations made of various topics, this chapter proposes to measure the topical diversity of the user interpretations of the same song as an indirect measurement of the complexity of song lyrics.* This study applies probabilistic topic models (i.e., Latent Dirichlet Allocation, or LDA)[81] to learn the latent topics from the user comments found in *songmeanings.com*. Then, the diversity of the topics for each song is measured to find its relationship to the other direct text complexity measure, a Lexical Novelty Score (LNS).

The topical diversity of song lyrics is of interest to this study, in addition to the text complexity of song lyrics. How to quantify diversity has been widely studied in various fields [114]. For instance, Magurran [115] pointed out that diversity has been a main theme in ecology because diversity often indicates the well-being of ecological systems. The topical diversity of documents has been used in the field of natural language processing to measure the topical diversity of academic communities, such as conferences and journals by Hall et al. and Bache et al. [82, 114]. It is also been used to identify the relationship between the topical diversity of documents and other characteristics. For example, Azarbonyad et al. [116] explored how topical diversity of text are related with interestingness. Hall et al. [82] introduced topic entropy as a measurement of topical diversity in conferences. However, they did not take the conceptual distance between the topics into account. Recently, Bache et al. [114] devised a more sophisticated topical diversity measure of text documents by considering both topic distributions and distances between topics. The study expanded Rao's coefficients, a diversity measure in biology [117], into topical diversity of topics learned from topic modeling. More details on this measure is found in Section 5.2.4. I will adopt the recent topical diversity measure to quantify the diversity in the interpretations of a song lyric. The following two indirect evaluation methods are used to validate the usefulness of the proposed method. The first evaluation method determines whether the LTDS can capture differences in topical diversity between popular songs and less popular songs, and the second evaluation method analyzes relationships between the LTDS and LNS.

This chapter aims to answer the following research questions.

- Research Question 3: Would the diversity of topics in interpretations of song lyrics be useful for measuring the complexity of song lyrics?
  - Research Question 3-1: Can the proposed measure in Chapter 5 capture differences in topical diversity between popular songs and less popular songs?
  - Research Question 3-2: Can we understand the measure better by analyzing the relationship between diversity of topics in song interpretations and lexical novelty of song lyrics?

Section 5.2 provides detailed information about the experiments to answer this question, including datasets for the experiments and evaluation, preprocessing techniques, and the topical diversity metric adopted from Bache et al. [114]. Section 5.3 reports experimental results, and Section 5.4 discusses the limitations of this study.

5.2 Experiment Design

5.2.1 Datasets

Song Interpretations Dataset (SID)

The interpretation data was obtained by using the API that *songmeanings.com* provides, which takes a title and an artist as parameters and returns a JSON file containing interpretations of a song lyric. This dataset is different from the Lyric Topic Dataset used in Chapter 4 in that this one is much larger but does not have topic labels. Among the millions of songs available on *songmeanings.com*, overlapping songs with the LNS Dataset (LNSD) below were used for the experiment to examine the relationship between the two metrics. Lexical Novelty Score (LNS) is a first lyric text complexity metric (more detailed information can be found in Section 2.1.2). To align SID and LNSD, we conservatively match the songs by using the title and artist information. Another important criterion in selecting those songs is the number of interpretations. As the number of interpretations depends not only on the level of disagreement but also on the popularity of the song, we select only the top five interpretations that obtained the most votes from the other users. In this way, the topical diversity can be more robust to the number of interpretations, too. Eventually, 4,642 artists, 36,041 songs, and 180,205 interpretations appear in SID.

LNS Dataset (LNSD)

To evaluate the proposed topical diversity metric, this study compares it with the LNS values calculated from the bag-of-words representation of lyrics. The LNS values are from the bag-of-words LyricFind corpus, which is associated with the LNS values and metadata information of 275,905 songs.[1] After rigorous matching between SID and LNSD, LNS scores of 36,041 songs were used for the experiment. LNS scores range from 0 to 100.

Billboard 100 Artist Dataset (BAD100)

An additional matching process between the 4,642 artists and the Billboard top 100 artists [2] yields 42 artists and 43,121 songs. BAD100 serves as an auxiliary dataset to examine

[1] http://www.smcnus.org/lyrics/

[2] https://www.billboard.com/articles/columns/chart-beat/5557800/hot-100-55th-anniversary-by-the-numbers-top-100-artists-most-no

whether there is a relationship between popularity and the level of agreement among user-generated interpretations.

5.2.2 Preprocessing

As the original data found on SID contains a lot of noise, such as user IDs and emoticons, only words with alphabetic symbols are considered. Lemmatization was performed using Morphadoner[3] to group words that share the same base form. The reason lemmatization is chosen over stemming is that it increases the readability of the words in the discovered topics, unlike stemming, which usually cuts out the ends of words[118]. General stopwords[4] that usually have low discriminative power are filtered out. In addition, collection-specific stop words that appear in more than 20% of documents are also removed. Finally, proper nouns recognized by the named-entity recognizer in Morphadoner are excluded [119]. This is because this study is interested in discovering general topics that do not depend on frequently appearing artists' names.

5.2.3 Parameters for Topic Modeling

From the preprocessed interpretation data, K different LDA topics were extracted in this study using MAchine Learning for LanguagE Toolkit (MALLET) [120]. Both unigrams and bigrams were used to improve the performance of topic modeling [121]. To investigate the effect of the different number of topics and the different choices of the Dirichlet smoothing parameter α, the following values were varied: $K = \{20, 50, 100\}$ and $\alpha = \{1.5/K, 5/K, 10/K, 15/K, 20/K\}$. Among those choices, $K = 50$ and $\alpha \geq 5/K$ showed reasonable results.

5.2.4 Topical Diversity Metrics

Entropy is widely used in many fields to measure the diversity of a population [114]. It was first defined and used by Hall et al. [82] to measure the diversity of topics of a certain conference proceedings learned from a probabilistic topic modeling algorithm. The topic entropy of a certain conference on a certain year is defined as

$$H(z|c, y) = -\sum_{i=1}^{T} p(z_i|c, y) \log(p(z_i|c, y)) \tag{5.1}$$

[3] http://morphadorner.northwestern.edu/morphadorner/
[4] https://code.google.com/archive/p/stop-words/

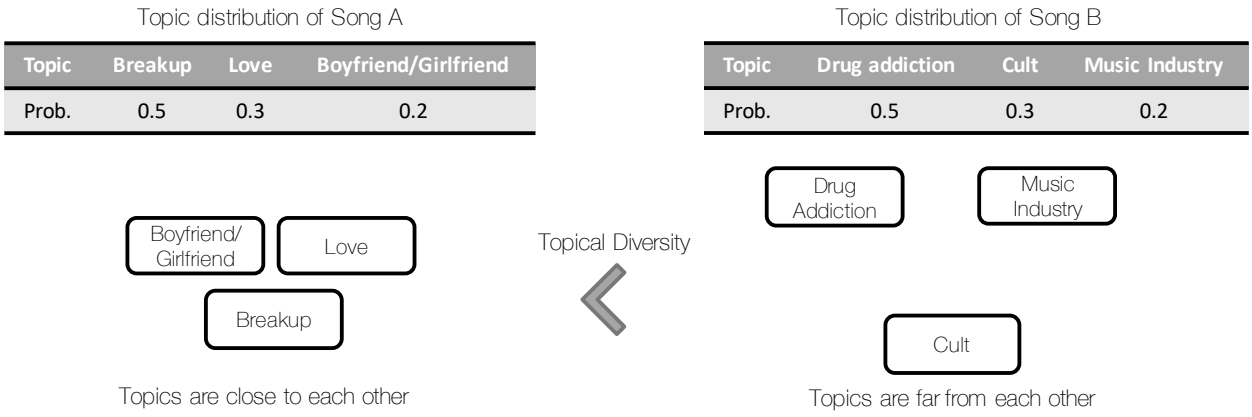

Figure 5.1: An imaginary example of two song lyrics with the same topic distribution with different average distances between topics. Entropy does not take into account the semantic distance between topics. Bache et al.'s topical diversity metric can measure them by using the equation 5.5

where $p(z_i|c, y)$ is the probability distribution of a topic z at a conference c on the year y and T is the number of topics.

As a diversity measure entropy makes sense, because the topic distribution of a document will be more disordered (i.e., closer to the uniform distribution), if there are many active topics in the interpretations of the same song. However, a straightforward application of entropy to measure topical diversity cannot always be accurate, especially if some topics are semantically similar to each other than to the rest. Figure 5.1 illustrates the case where the entropy measure cannot capture the gap of topical diversity between two fictional song examples. Song A has three active topics (*Boyfriend/Girlfriend*, *Love*, and *Breakup*) and they are closely related to each other because they are all about relationships. Conversely, song B has three active topics (*Drug Addition*, *Music Industry*, and *Cults*) that are not closely related to each other. The two songs have the same topic distribution probability, but entropy does not take into account the semantic distances between topics; therefore, the entropy scores of song A and song B are the same. Bache et al. [114] designed another topical diversity measure to overcome this shortcoming. Compared to the topic entropy, Bache et al.'s metric takes the distance between topics into consideration as well. There, the diversity of d-th document is defined as

$$div^{(d)} = \sum_{i=1}^{K} \sum_{j=1}^{K} P(i|d)P(j|d)\delta(i, j), \tag{5.2}$$

where $P(i|d)$ is the proportion of words in document d that are assigned to topic i, which can be calculated by using the LDA results. For example, let $\mathbf{C}$ be a $D \times K$ matrix whose (d, k)-th element contains the number of words in the d-th document that belong to the i-th topic. Then, the proportion of the topics in the d-th document can be calculated as follows:

$$P(k|d) = \frac{\mathbf{C}_{d,k}}{\sum_k \mathbf{C}_{d,k}}. \tag{5.3}$$

$\delta(i, j)$ is the distance between topic i and topic j, which can be defined based on the same document-topic matrix $\mathbf{C}$. First, we calculate the cosine similarity between the two co-occurrence vectors $\mathbf{C}_{:,i}$ and $\mathbf{C}_{:,j}$:

$$s(i, j) = \frac{\sum_d \mathbf{C}_{d,i}\mathbf{C}_{d,j}}{\sqrt{\sum_d \mathbf{C}_{d,i}^2}\sqrt{\sum_d \mathbf{C}_{d,j}^2}}. \tag{5.4}$$

As this is a similarity metric ranged between 0 and 1 (1 means they are same), not a distance, we can simply convert it into a distance measure as follows:

$$\delta(i, j) = 1 - s(i, j). \tag{5.5}$$

5.2.5 Evaluation Methods

Evaluation 1: The Differences in Topic Diversity between Popular Songs and Less Popular Songs

Previously, Ellis et al. [13] evaluated their lyric complexity metric, LNS, by examining whether the lexical novelty scores of the most popular songs appearing in the Billboard charts tend to be lower than those of other songs based on the assumption that highly complex lyrics tend not to be chart-worthy. This study adopts the same evaluation method used by Ellis et al. [13] because it is the only lyric complexity measure that has been developed so far. Empirical Cumulative Distribution Functions (ECDF) is used to determine whether the LTDSs of the most popular songs tend to be low. Next, statistical significance between LTDSs of the two groups, the most popular songs and the remaining songs, is assessed using a nonparametric two-sample Mann–Whitney (MW) test.

ECDF is an empirical method to approximate the Cumulative Distribution Function (CDF). It can be tricky to calculate the CDF function (1) if the probabilistic density function cannot be easily integrated, or (2) if we do not know the probabilistic density

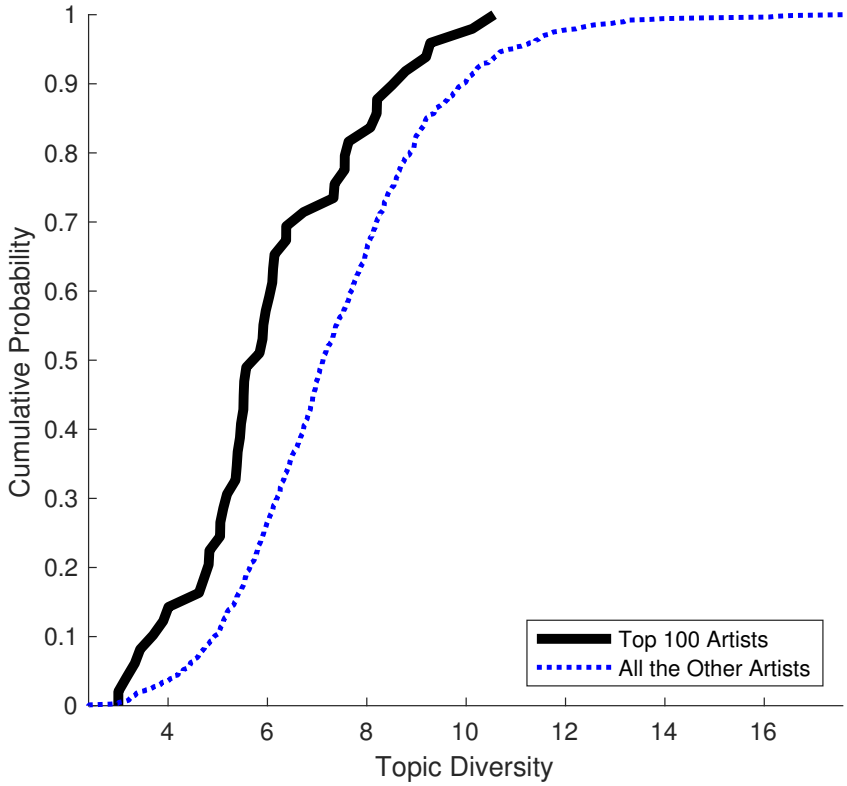

Figure 5.2: Empirical cumulative distribution functions of the averaged topical diversity of the two sets: BAD100 and all the others

function. ECDF approximates CDF based on the sample distribution. For a given interval, the count of the samples that fall in that interval divided by the total number of samples can serve as the incremental gap. By accumulating these gaps over the observed range of the random variable, the ECDF is formed with multiple steps in its shape. Based on the similar argument for the MW test, it is reasonable to expect that the ECDF from topical diversity found in the very popular songs will be different from the ECDF calculated from the less popular ones; the ECDF curve will hit 1 earlier if there are more songs with small topical diversity values in the sample distribution.

The MW test is a flexible statistical test that can compare two data sets without assuming that either of them follows the Gaussian distribution. Its null hypothesis is that a random sample from one distribution can be bigger or smaller than the other random sample from the other distribution with an equal probability. Therefore, if the test result rejects the null hypothesis, we can say that topical diversity from the popular songs tends to be different than the less popular ones. Both the MW test and the comparison of the ECDF graphs were also used in the context of evaluating the LNS metrics in Ellis et al. [13].

Evaluation 2: Relationship to LNS

The validity of the LTDS is further evaluated by conducting a regression analysis. For the same set of songs, by having the LNS as the predictor while the LTDS is the response variable, a linear regression model is learned to draw the level of correlation between the two measurements. F-test and the R^2 values are traditional techniques used for this kind of evaluation.

5.3 Results

5.3.1 Topic Modeling Results

To get a general idea about the topic modeling results, Table 5.1 lists the top 25 topics out of 50 total topics ordered by their associated Dirichlet parameter α_k. First, because the user comments tend to include some topics that do not really discuss the meaning of the lyrics, we see some noisy, but popular topics, such as topic 1, 2, 4, 7, and 16. Topic 1 consists of terms that frequently appear in interpretations of song lyrics, such as *verse*, *word*, and *refer*; topic 2 and topic 4 are about music-related terms, such as *play*, *band*, and *record* and terms to express their feelings, such as *awesome*, *amaze*, and *cool*; topic 7

ID	α_k	Top Words
1	0.47501	mind verse word point refer idea place change sense lose true reference understand read interpretation turn long
2	0.34321	play year band lot boy record big rock kind work fun read probable start drink word cool obvious night
3	0.30865	relationship break heart feeling hurt lose fall long change hard realize bad happen wrong care wait hope hold start
4	0.16905	band awesome amaze cd comment rock agree guitar kick ass cool yeah kick_ass voice lol stuff play total solo
5	0.08093	beautiful amaze sad voice perfect true sweet simple remind comment summer wonderful absolute happy agree year
6	0.07985	dream night eye sleep light fall wake rain hand dark wait sky heart morning cold sun water sea turn
7	0.05577	fuck shit hate ass suck bitch fuckin stupid hell band rap sick kick piss funny agree blah bad kill
8	0.04041	die death dead kill heaven alive ghost wife suicide pass funeral story die_die cancer friend memory accident loss grave
9	0.0374	friend friendship friend_friend boyfriend close_friend remind_friend fake remind hang lose_friend crazy grow
10	0.03367	god religion christian faith religious sin church belief heaven bible christianity save lord pray holy question evil
11	0.03354	child father mother parent son dad family daughter kid mom baby abuse grow young sister abortion brother
12	0.03348	war fight country government american soldier political bush freedom power history battle bomb army nation
13	0.03341	dance baby wanna yeah gonna ha catchy gotta ya lol beat yeah_yeah fun night ha_ha club da party play
14	0.03316	drug addiction heroin addict high drink alcohol cocaine drug_addiction smoke reference problem needle refer
15	0.03198	sex woman sexual gay male female rape boy lover sexy sexuality pleasure sex_sex desire lust obvious virginity
16	0.03111	version cover original beatle original_version acoustic acoustic_version version_version cover_original
17	0.0248	school young boy kid grow high high_school year youth age parent college bully childhood teenage adult
18	0.02083	voice beautiful amaze romantic dream mouse feeling meat incredible emotion sweet intense powerful
19	0.02047	cheat woman wife girlfriend affair husband marry steal lover marriage relationship secret promise obvious doesn
20	0.02023	suicide commit depression commit_suicide gun kill depress sister suicidal singer shoot attempt jump glass imply
21	0.01981	video watch watch_video http http_watch sonic film youtube kid sonic_youth odd phone clip video_video cool
22	0.01969	band fan stand member label musician beatle indie record pearl band_member artist crowd fame tribute deaf garden
23	0.01876	movie la writer story la_la film character lose whoa tonight goodbye soundtrack beautiful kill vampire burn titanic
24	0.01851	burn light heart flame soul candle burn_burn bit sting heart_soul set gambler light_light alive sacred sleep die
25	0.01811	punk band punk_rock clash pistol punk_band yea fat fox rock nofx fan sex_pistol rhythm money gun fight emo rich

Table 5.1: Top 25 topics out of 50 total topics ordered by the Dirichlet parameter α_k. Their top words are displayed along with the Dirichlet parameter α_k. Bigrams are connected by an underscore

consists of terms that express negative emotions, such as *fuck*, *shit*, and *hate*; and Topic 16 contains terms that refer to types of version of songs, such as *version*, *cover*, and *original*. However, most of the other topics represent a certain subject that the lyrics are about. For example, topic 3 is about *Relationship* and *Break-up*; topic 8 contains words about *Death*; topic 10 is mostly for *Religion*; topic 12 represents *War*; and so on.

5.3.2 The Differences in Topic Diversity between Popular Songs and Less Popular Songs

From the procedure described in Section 5.2.4, the topical diversity of all interpretations is calculated, which gives one diversity measure per song. Then, the diversity values are averaged over the songs from the same artist.

First, a Mann-Whitney U-test was conducted as suggested in [13] to test the hypothesis that the two topical diversity distributions of BAD100 and the others have the same median. It turned out that the test rejects this hypothesis with $p = 1.9973 \times 10^{-5}$, meaning that the topic diversities of the two groups are from distributions with different median values.

Figure 5.2 provides evidence that the empirical cumulative distribution functions of the two sets of artists are quite different from each other. As shown in [13], more lexically complex lyrics are less "chart-worthy." Here we observe that the more popular artists tend to write lyrics that provoke a low level of topical diversity, which is reflected as a graph (thick line) with low LTDSs.

5.3.3 Relationship between LNS and Topical Diversity

This experiment also tests if the LNSs over the artists have a correlation with the LTDSs. For the 1,348 artists with more than five songs we calculate their averaged LTDSs. A linear regression shows a linear relationship between the two variables, as shown in Figure 5.3 ($p = 6.87 \times 10^{-38}$). The R^2 value of this model is 0.116, meaning that the model explains 11.6% of the variability in the response variable, averaged topical diversity. The value itself is not very high, but the significant p-value shows that we can still reject the null hypothesis that the estimated slope is zero.

Figure 5.4 shows that the sectional distribution of the topical diversity per period of LNS values ranged between $0-10, 10-20, \cdots, 90-100$. First, we can see that the lyrics with a lower lexical complexity are associated with lower as well, while the group with a high LNS is associated with a distribution with a large mean. This clear correlation supports the

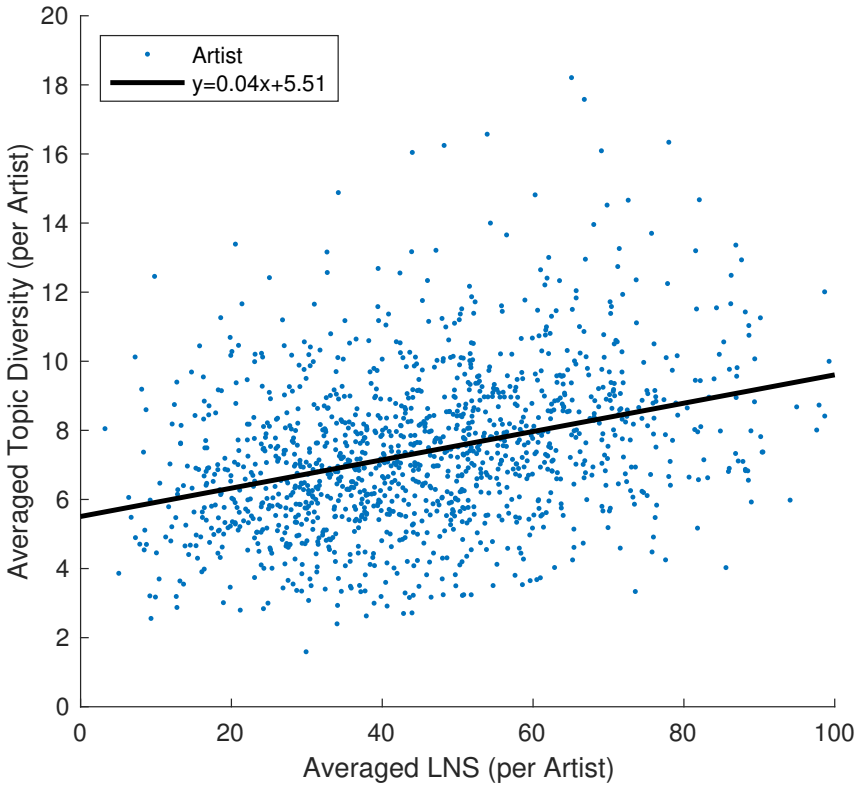

Figure 5.3: A linear regression result between LNS and the proposed topical diversity. An average value for an artist is reported as a dot

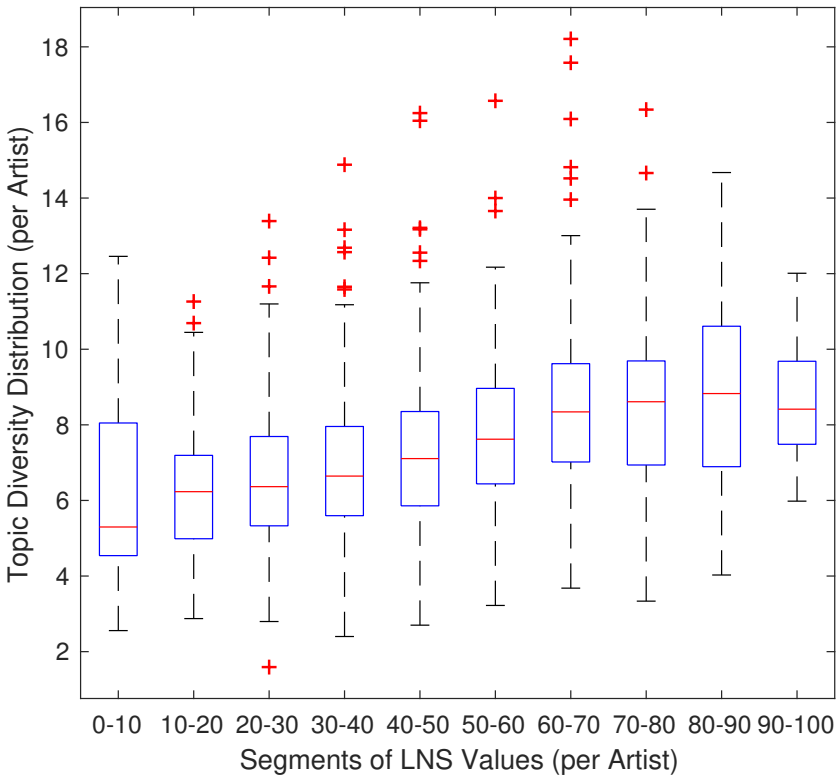

Figure 5.4: Artists with more than 5 songs are grouped into ten sections based on their averaged LNSs. In each section we draw a box plot using the averaged topical diversity of the artists belonging to the section

assumption of this study that the complexity of song lyrics is related to the level of disagreement among the user interpretations, and the level of disagreement is successfully captured by the LTDS.

5.4 Discussion and Conclusion

The comparison of LTDSs between songs from the most popular artists and the remaining artists positively answers research question 3-1: the proposed measure in Chapter 5 can capture the differences in topical diversity between popular songs and less popular songs. In particular, diversity of topics in popular songs tends to be lower than in less popular songs. The correlation analysis between LTDS and LNS also positively answers research question 3-2: the two measures are highly correlated, which might capture how lexically novel words lead to different kinds of interpretations. In conclusion, the answer to research question 3 is that the diversity of topics in interpretations of song lyrics be useful for measuring the complexity of song lyrics while also raising several interesting discussion points.

First, it is challenging to design and conduct user studies to collect reliable lyric complexity scores. As assumed in this chapter, each user can rate lyric complexity differently depending on that individual's particular knowledge about and interest in songs and their artists. Suppose user A reads lyrics of song S for the first time and does not know about the artists, while user B has liked song S for years and knows a lot about the artists, such as the story of how song S was created. This difference can lead to different interpretations and judgments of lyric complexity: user A can capture the most superficial meaning and consider it simple, and user B can capture a deeper level of meaning and consider it complex. Therefore, instead of simply averaging lyric complexity scores from multiple users, a more complicated method is necessary to consolidate the collected scores by taking into account how different users interpret the same song lyrics. Therefore, effective user studies need to collect not only lyric complexity scores but also users' interpretations as well as any information implies users' interest and knowledge. In essence, the user comments collected from *songmeanings.com* are not very different from the target data that effective user studies try to collect except for the existence of final lyric complexity scores.

Due to the challenges of user studies mentioned above, indirect evaluation methods were chosen in this chapter as an alternative. In the literature, it is not rare to rely on indirect evaluation methods to validate quantitative analysis results, especially those learned in an unsupervised manner from large-scale datasets, where the indirect evaluation is

characterized by a comparison with the other known metrics. Dodds et al. [122] proposed the happiness measure that calculates the valence level of documents by averaging valence scores of words. It is difficult, however, to verify that the resultant document-specific happiness scores are due to the lack of the ground-truth, which can be collected only via a thorough and complex user study. Instead, they evaluated the usefulness of the measure by examining valence trends of song lyrics in an indirect fashion, that is, by identifying a positive relation to the historical events. They also found the relationships between valence values of song lyrics and music-related metadata, genres, and artists. Another study by Ellis et al. [13] also assessed the effectiveness of the proposed lyric complexity measure using indirect evaluation methods, an approach this chapter adopts. Therefore, to overcome the bias towards a particular external method, it is important to consider different relationships to various metrics and analyze the results. In sum, because both happiness and complexity are complicated concepts, it might be more reasonable to identify tendency and relationship with other factors first before conducting user studies.

Despite the rationale behind selecting the indirect evaluation methods, those methods do have clear limitations. First, there is no known relationship between lyric complexity and song popularity except that identified by LNS. In fact, according to the work of Berlyne and of North and Hargreaves [123, 124], there is an inverted-U relationship between music complexity and preference, which means that people tend to prefer music with a moderate level of complexity [123]. However, both LNS and LTDS showed that people tend to prefer songs with a lower level of lyric complexity. The difference might be explained by the fact that those experiments done by Berlyne and by North and Hargreaves used instrumental music , which does not contain song lyrics, while the evaluations of LNS and LTDS used music with lyrics. Second, the positive relationship to the LNS does not necessarily mean that the proposed measure is always valid. The LTDS is meant to capture an additional aspect of lyric complexity rather than to reaffirm the validity of LNS. However, it is also reasonable to assume that lexically novel words may invoke deeper thought processes and a greater diversity in interpretations.

5.5 Summary

This chapter proposed a lyric complexity metric that tries to capture qualitative dimensions of text complexity. The metric is based on user-generated interpretations posted on the web that provide qualitative analyses of song lyrics at no cost. Its usefulness as an input to automatic topic detection algorithms was verified in Chapter 4. The

assumption behind this metric is that more complex song lyrics can lead to many different interpretations with various topics. The LTDS tries to measure topical diversity of interpretations of a song lyrics by adopting Bache et al.'s [114] topical diversity metric.

The LTDSs of songs written by the 42 most popular artists were compared to the LTDSs of the remaining artists in the SID. The comparison showed that (1) the LTDSs of the two groups are statistically significantly different; and (2) the songs of popular artists tend to have low level of topical diversity. Another experiment identified a correlation between the LTDS, the metric proposed in this chapter and another lyric complexity metric, the LNS. These findings indicate that the topical diversity-based lyric complexity metric can measure some aspects of lyric complexity.

The approach based on user-generated data introduced in this chapter has potential for various applications. These can be applied to measure complexity of a variety of objects if user-generated reviews are available, including other genres of literature, such as fictions and poetry, as well as non-text objects, such as movies and products sold by *Amazon.com*.

Chapter 6

Conclusions and Future Work

6.1 Introduction

This dissertation conducted three studies to investigate computational methods to measure the complexity of song lyrics in the interest of providing better music services. Like complexity in other text genres, lyric complexity refers to the level of understanding the difficulty of lyric texts in both quantitative and qualitative dimensions. The quantitative dimensions of lyric complexity can be measured effectively by using computational methods. These quantitative dimensions include word frequency, word familiarity, and concreteness (more information about quantitative dimensions of lyric complexity can be found in Chapter 2.2.3). The qualitative dimensions of lyric complexity require attentive reading to reach a deeper level of understanding. Some song lyrics have multiple levels of meaning, which human readers understand through their comprehension and extra knowledge beyond the lyric text (more information about qualitative dimensions of lyric complexity can be found in Chapter 2.2.4). Compared to qualitative dimensions of lyric complexity, quantitative dimensions of lyric complexity can be measured efficiently even on a large scale but need to be complemented by qualitative dimensions. Acquiring qualitative dimensions of lyric complexity is expensive because attentive readers are required. To overcome these limitations of the two dimensions of lyric complexity, this dissertation not only explored concreteness, one of the existing quantitative dimensions of lyric complexity, but also proposed a way to measure lyric complexity by utilizing user-generated data that reflect qualitative dimensions of lyric complexity.

The chapter begins with conclusions of each study that reflect upon research questions. Subsequently, overall limitations of this dissertation are presented with regard to three aspects: data, techniques, and evaluation methods. Future research opportunities are

explored for each of the individual studies.

6.2 Conclusions

The overarching research question of this dissertation was, "How can the complexity of song lyrics be measured computationally?" To answer this question, the author conducted three studies. Study 1 in Chapter 3 took the traditional quantitative text complexity approach that analyzes linguistic properties of song lyrics. Study 2 in Chapter 4 introduced user-generated interpretations of song lyrics and evaluated how much topic-related information can be obtained from the data using automatic text analysis methods. Study 3 in Chapter 5 proposed a lyric complexity metric that utilizes topic diversity of user-generated interpretations of song lyrics. The conclusions from each study are provided below.

6.2.1 Conclusions of Study 1: Concreteness of Words as a Lyric Complexity Metric

Study 1 focused on concreteness, one of the traditional text complexity metrics. In particular, Study 1 investigated changes in concreteness of western popular music over the last 55 years to answer research question 1: "How has text complexity of popular song lyrics changed over time in terms of concreteness?" Study 1 also explored how music-related metadata and linguistic information are correlated with the trend, to further explain the general trend in concreteness and to answer research question 1-1: "What is the relationship between the concreteness trends and genres?" and research question 1-2: "What is the relationship between the concreteness trends and word statistics in song lyrics?" Further analyses of selective examples were conducted to examine the limitations of concreteness as a text complexity metric.

There was a clear downward trend in concreteness of popular music until the early 1990s, followed by an upward trend. To further analyze the V-shaped trend, correlation analyses with genres, word classes, and the length of song lyrics were conducted. The fact that "Hip-Hop/Rap" first appeared in 1990s might partially explain the upward trend of the overall concreteness. Increased popularity of "Hip-Hop/Rap" might have led to increased concreteness in other genres as they began to include rapped sections or passages, however, this hypothesis needs to be tested thoroughly in the future studies. As for word classes, the similar V-shaped trend of concreteness was also observed when only open class words were

considered. Particularly, it turned out that among the open class words, Nouns tremendously contributed to the overall concreteness trends of open class words.

The lyric complexity metric based on concreteness scores has advantages and disadvantages. As for the advantages, like other quantitative text complexity measures, this automatic method is easy to use and does not require humans to read the lyric texts to determine the level of difficulty. In addition, this data-driven metric covers a large amount of English words and can be improved by incorporating the concreteness scores of even more words. However, analyses of selective examples, such as non-narrative love/relationship songs and highly figurative songs, showed aspects of text complexity that concreteness cannot capture. The former type has many words with low concreteness values to convey love or sadness, the most common themes in popular songs that are easily understood. However, song lyrics with metaphoric expressions look easy to understand with highly concrete words, they require more knowledge for their comprehension. These analyses suggest that for a comprehensive understanding of lyric complexity, instead of solely using concreteness scores, it is important to consider other metadata, such as topics of song lyrics, as well as qualitative dimensions.

6.2.2 Conclusions of Study 2: Evaluating Usefulness of User-generated Interpretations

Since the quantitative text complexity metrics have their own limitations, it is recommended to complement them with qualitative analyses. There is currently no existing dataset of human evaluation of lyric complexity in the public domain, however, but users' interpretations of song lyrics are freely available. To justify Study 3's approach to utilize those user-generated interpretations as a proxy for the missing human evaluation of lyric complexity, Study 2 sought to answer research question 2: "Can an automatic algorithm successfully identify underlying topics of song lyrics from user-generated interpretations?" In particular, Study 2 investigated usefulness of user-generated interpretations and song lyrics as input to lyric topic classification systems to answer research question 2-1: "Are users' interpretations of song lyrics more useful than song lyrics for the topic classification task?" and research question 2-2: "How different are the most representative words of interpretations and lyrics?"

Interpretations were found to be more useful than lyrics in different settings of classification systems. The accuracy gaps were 20.88% (term frequency-based feature representation) and 14% (word embedding-based feature representation). Interpretations

were also found to be more useful than lyrics in five topic categories by a wide margin. The analysis of the most representative words from the two sources revealed that words in song lyrics tend to be poetic and figurative while words in their interpretations tend to be straightforward. These findings suggest that interpretations may be more promising than lyrics for the task of computational understanding song lyrics.

6.2.3 Conclusions of Study 3: Topical Diversity of Interpretations as a Lyric Complexity Metric

To overcome limitations of existing quantitative text complexity metrics, such as the inability to capture multiple layers of meaning, Study 3 explored research question 3: "Would the diversity of topics in interpretations of song lyrics be useful for measuring the complexity of song lyrics?" This unprecedented approach takes into account how users on the web agree or disagree with others' interpretations of song lyrics. The assumption behind the measure is that users tend to disagree with each other if song lyrics are difficult to understand, thereby leading to a wide range of interpretations. The level of agreement was modeled by the diversity metric defined using the LDA topics, which consider both the semantic distance between topics as well as their contribution in a given document. To verify the usefulness of the lyric complexity metric proposed in Study 3, the following two research questions were answered: research question 3-1: "Can the proposed measure in Chapter 5 capture differences in topical diversity between popular songs and less popular songs?" and research question 3-2: "Can we understand the measure better by analyzing the relationship between diversity of topics in song interpretations and lexical novelty of song lyrics?"

The two experiments to evaluate the proposed measure showed promising results. The first experiment revealed that the proposed measure can capture the difference between very popular songs and the remaining songs. The second experiment showed that the statistically significant correlation between the proposed topical diversity and LNS, another lyric complexity metric, is significant. Although the evaluation methods are not as effective as user studies, these two findings indicate that the measure based on the interpretations is worth further investigation.

The strongest point of LTDS is utilizing qualitative dimensions of lyric complexity in a quantitative manner to take advantages of both approaches. LTDS has inherent weaknesses, however, stemming from the nature of the user-generated interpretations that the proposed measure heavily relies on. For instance, LTDS is applicable to songs with a

sufficient number of quality interpretations, but very popular songs tend to have many quality comments while less popular songs are likely to have no comments at all. LTDS also has another weakness coming from topic modeling. To be specific, LDA can identify highly abstract levels of topics while the levels of real-world topics of song lyrics can range from extremely abstract to extremely specific. Taken together, these findings indicate that to improve the proposed lyric complexity method, future research needs be done to 1) encourage people to write more comments on less popular song lyrics, and 2) replace LDA with other topic identification algorithms that can capture more specific topics as well.

6.3 Limitations

As one of the initial studies on lyric text complexity, this work has limitations that offer opportunities for further research. They are reported here in terms of data, techniques, and evaluation methods.

Data

This dissertation focused on western popular song lyrics. Study 1 in Chapter 3, about 5,500 song lyrics on the Billboard 100 Hot chart between 1960 and 2015 were examined. Study 2 in Chapter 4 used 800 western popular songs that are listed on *songfacts.com*, and finally, Study 3 in Chapter 5 handled 275,905 songs that appear both the LyricFind corpus and *songmeanings.com*. Because *billboard.com*, *songfacts.com*, *lyricfind.com*, and *songmeanings.com* are music services mainly targeting people in western countries who use the English language, the results from this dissertation cannot be generalized to all music, including non-western music. To determine whether the methods in this dissertation are applicable to non-Western music, such as K-Pop, further cross-cultural investigations are necessary.

It is also important to note that the data collection for each individual chapter has its own limitations. First, the 5,500 Billboard songs cannot truly represent western popular music as a whole. According to the Long Tail theory [125], a huge number of songs in the long tail of a demand curve are as important as the small number of songs in the head of the demand curve. The findings of Study 1 only reflect the head; therefore, they are not applicable to the long tail. Second, the dataset for the classification in Study 2 has its own limitations: the balanced dataset is not realistic in that some topics are much more prevalent than others, and song lyrics can have multiple topics. However, I used the

dataset because it was important to control other variables to better compare lyrics and interpretations. Finally, as for the corpus of the user interpretations of song lyrics, a popularity bias exists because songs with less than five comments were excluded to attain a sufficient amount of interpretations for each song lyrics. For this reason, the findings cannot represent a general property for all popular music. Instead, they only represent songs that attract a lot of attention from users of *songmeanings.com*.

Techniques

The bag-of-words model, where a document is represented as a set of words, was mostly used to analyze the text. This dissertation starts from this simple but strong approach; however, it disregards the other linguistic characteristics, such as word order and grammar. More advanced features detailing necessary linguistic information could lead to better computational text complexity metrics.

Naïve Bayes was chosen in Study 2 because the size of the collection was too small for deep learning algorithms, and naïve Bayes was almost as good as SVM in a similar task [9]. However, since more powerful classification algorithms are available, they might provide different results. For instance, deep learning algorithms might improve the performance when a large number of song lyrics are available.

Evaluation

Further user studies could strengthen the evaluation of the topical diversity measure in Study 3. In this case, the proposed measure was evaluated by indirect evaluation methods: 1) to determine whether it could recognize the differences in topic diversity between very popular songs and less popular songs and 2) to discern any correlation between the proposed measure and LNS, another lyric complexity metric. Although the two indirect evaluation methods showed promising results, additional user studies could directly validate the effectiveness of the proposed measure.

6.4 Future Research

The findings and the proposed methods of this dissertation can be used to solve other research problems. Moreover, the limitations of this dissertation suggest the need for a number of further studies to better measure lyric text complexity and to improve researchers' understanding of the user data. The section starts with discussions on

complexities of measuring lyric complexity, followed by possible topics of future research for each study.

6.4.1 Future Research Related to Complexities of Measuring Lyric Complexity

This dissertation explored two lyric complexity metrics, but there are many more questions to be addressed regarding this topic going forward, because measuring lyric complexity is a very complex task.

Subjectivity of Lyric Complexity

As CCSS's three-part model of text complexity suggests, opinions on complexity of song lyrics can be subjective, so it is important to assess the reader's knowledge, experience, and motivation as well as inherent text complexity when measuring text complexity. Readers who have relevant knowledge and experiences to understand lyrics of a certain song can perceive it less complex than those who do not. On one hand, readers who are interested in the artist might consider the song lyrics more complex, thereby adding meanings. On the other hand, they might perceive songs as less complex if their background knowledge enables them to understand them easily. Subjectivity of understanding in music, such as similarity and mood, is a well-known topic in MIR research [126]. Therefore, personalization has become an important direction in MIR research[127, 128]. An important research question for future work is how to personalize lyric complexity metrics.

Meaningless Lyrics

It is challenging to determine the level of complexity of nonsensical or meaningless song lyrics. Unlike other literature, such as novels and poems, song lyrics are always coupled with music. Therefore, a song can be considered a complete artwork even if part or all of its lyrics do not make sense. This is more common in certain genres than others, for instance, it is known that lyrics are often meaningless in rock and roll music [129]. Some examples of meaningless lines are: "Mmmbop, ba duba dop ... Ba du, yeah" from Hanson's "MMMBop" and "Rah rah ah-ah-ah! Ro mah ro-mah-mah Gaga oh-la-la!" from Lady Gaga's "Bad Romance." There are many questions to explore with regard to such lyrics in the future: how does one automatically detect nonsensical lines? Do these meaningless lines increase or decrease lyric complexity? Which genres have more meaningful lyrics? What are the types of nonsensical lines?

6.4.2 Future Research Related to Study 1: Concreteness of Words as a Lyric Complexity Metric

In the study, concreteness was applied to thousands of the most popular songs and the correlations were analyzed along with the diachronic trend and various metadata. The possible directions of further studies include investigating diverse text complexity metrics on large-scale song data and examining how other important metadata relate to lyric complexity. Specific future studies are as follows.

Exploring Other Text Complexity Metrics

Since this is an initial study of lyric text complexity, concreteness was only explored among various text complexity metrics. The next step is to investigate other text complexity metrics. The metric with the highest potential for lyric complexity analysis is lexical tightness, as proposed by Flor et al. [130]. Lexical tightness measures semantic relationships between pairs of content words in a document by using their co-occurrence information in a large reference dataset. To the best of my knowledge, this is the only automatic text complexity metric that showed promising results in analysis of poetry; it was evaluated using a small poetry dataset, of just 66 poems. Given the similarities between poems and song lyrics, lexical tightness might be a useful metric for conducting trend analyses of lyric complexity. Another possible future work is to apply the readability formulas provided by *https://readable.io/* to analyze the readability trends of music on a larger scale.

Large-Scale Text Complexity Analysis

This research examined the text complexity of 5,500 lyrics of the most popular songs in terms of concreteness. As mentioned in the limitation section, the findings from the experiments on the songs at the head of a demand curve might not be applicable to songs at the middle and long tail. To expand the depth and breadth of this study, the author plans to explore the entire LyricFind dataset of 1 million songs, which includes songs at the middle and long tail as well. The bigger dataset can also enable the analysis to include larger genre categories beyond the four major genres.

Gender Trend Analysis of Popular Music

Another topic to explore is how the gender of artists influences the concreteness trends. The trend of concreteness of popular song lyrics has changed dramatically from downward to upward in the early 1990s. That time coincides with the period of American popular music when female artists started standing out on the charts more often than before [96]. The author plans to determine whether the gender information of artists and the concreteness trend are correlated with each other.

6.4.3 Future Research Related to Study 2: Evaluating Usefulness of User-generated Interpretations

The study found the usefulness of user-generated interpretations of song lyrics by comparing them to song lyrics as input to automatic topic classification systems. Future experiments on large-scale data with a great number of topics could assess their value in a realistic scenario. Furthermore, content analysis of the data would be another way to improve understanding of the data.

Large-Scale Topic Classification

The author also plans to expand the experiments to include a large-scale unbalanced dataset to be more realistic and to exploit the state-of-the-art deep learning algorithms. Furthermore, it is worth investigating various advanced NLP technologies to improve the current measures.

Content Analysis on Song Lyrics and Their Interpretations

Users' interpretations about song lyrics on the web have proven to be useful in the automatic song lyrics topic classification tasks. Although the analysis of the most representative words in the data revealed their characteristics to some degree, content analysis could answer more questions about their nature and value. To improve understanding of the data and come up with better ways to exploit the data, the author plans to conduct a content analysis.

6.4.4 Future Research Related to Study 3: Topical Diversity of Interpretations as a Lyric Complexity Metric

The study showed promising applications of supplementary user-generated data to improve understanding of the target data, which is known to be difficult for humans as well as machines. One direction of future work could be to apply this approach in other research areas where auxiliary user data can benefit other users. Another direction is to incorporate a user study to evaluate the proposed measures. More details of future studies are as follows.

Using User-generated Data for Studying other Literature

This dissertation research utilized user-generated data on the web to comprehend complex text data. This approach can be applied to other genres of literature where auxiliary user-generated data are available. For example, *goodreads.com* provides quality reviews and ratings for a variety of genres of books ranging from Children to zombies. Users of the website can also ask questions about books to other users. Another example website is *poems-and-quotes.com* where poems are categorized by dozens of topics and users can share their reviews in the form of comments. In addition to such websites, more discussions between users take place on a wide range of websites, including blog postings, articles, and digital libraries.

Using User-generated Data for the Analysis of Non-text Objects

User-generated data can also be beneficial for the analysis of different forms of objects beyond text data. For instance, the approach taken in this research can be applied to analyze movies and TV shows, as reviews are available from various websites, including *imdb.com*, *rottentomatoes.com*, *metacritic.com*. and *fandango.com*. Given that text analysis is easier than multimedia analysis and that more information can be enriched by users, it can help users better understand video content. The supplementary data-based understanding objects are applicable to non-cultural data, for example, products on online shopping sites such as *amazon.com*. Products with controversial reviews can be interpreted differently than ones with high levels of agreement.

Providing User-generated Data to Immerse Users in Cultural Objects

One possible directions of future work is to study how to provide users with other user-generated data on the web in real-time while they experience some cultural objects beyond song lyrics. For example, people who visit museums could enjoy artworks better if they could take advantage of other users' knowledge and opinions. To make this possible, many questions could be answered, including the following: how does one present user-generated data effectively?; is an expert's explanation still as important as crowd-sourced one?; how does a user pick quality data and penalize unreliable data?

User Studies

To strongly validate the proposed complexity measures, the author plans to conduct user studies in the future. For the user studies, user data about lyric text complexity will be collected through Amazon Mechanical Turk, which is a widely used crowdsourcing platform in MIR [131–133] and other fields [134, 135]. The respondents will rate perceived difficulty in understanding the meaning of songs. They will also be asked to answer other questions about the difficulty of words and sentences in lyrics, as shown in 6.1. For further analysis, songs will be sampled with an even distribution over time, genre, and artist.

6.5 Summary

This dissertation investigated computational methods to measure the complexity of song lyrics to expand the scope of research on automatic music complexity annotation and to provide better MIR services. The key findings are as follows:

- This is the first study that explored concreteness of song lyrics in terms of text complexity. The V-shaped trend of concreteness between 1965 and 2015 are correlated with genres, the proportion of word classes, and the number of words in song lyrics. More specifically, the growing popularity of hip-hop and rap may have contributed to the upward trend after 1991; Nouns tremendously contributes to the overall concreteness trends; and the longer song lyrics, the lower the concreteness scores.

- User-generated interpretations outperform song lyrics as input to lyric topic classifiers by big margins. The comparison of the top 20 words for each topic between the two input sources confirms that the language used in the interpretations is more

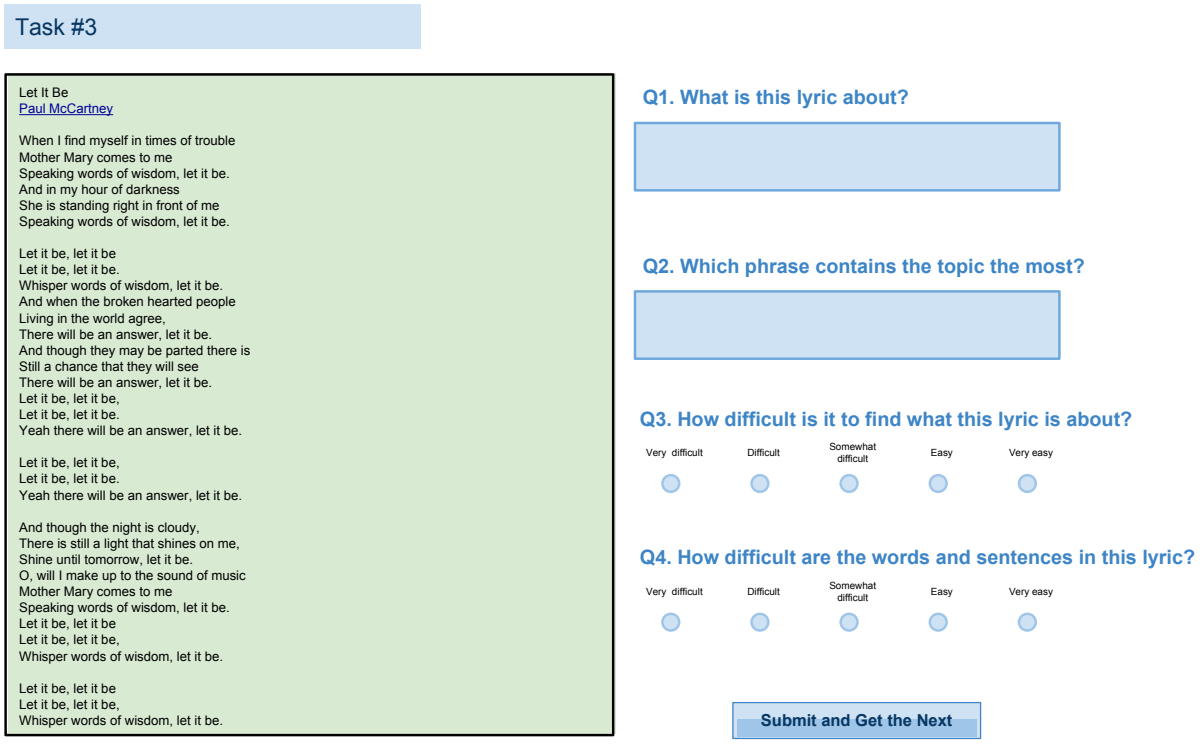

Figure 6.1: Survey questions for the user study in a mockup image

straightforward and closely related to each topic. These findings suggest that the interpretations may be a better input source for automatic topic detection algorithms than song lyrics.

- This work also proposes a lyric complexity metric that tries to capture qualitative dimensions of text complexity. It measures topical diversity of interpretations of a song lyrics based on the assumption that more complex song lyrics can lead to many different interpretations with various topics. This measure captured the different tendency of LTDSs of song lyrics of the most popular artists and the remaining artists. A correlation to another lyric complexity metric, LNS is also identified. These findings indicate that this unique lyric complexity measure based on topical diversity scores of user-generated interpretations is promising.

References

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

[2] J Stephen Downie, Xiao Hu, Jin Ha Lee, Kahyun Choi, Sally Jo Cunningham, and Yun Hao, "Ten years of MIREX: reflections, challenges and opportunities," in *Proceedings of the 15th International Conference on Music Information Retrieval*, 2014.

[3] Zhouyu Fu, Guojun Lu, Kai Ming Ting, and Dengsheng Zhang, "A survey of audio-based music classification and annotation," *IEEE Transactions on Multimedia*, vol. 13, no. 2, pp. 303–319, 2011.

[4] Jose PG Mahedero, Álvaro MartÍnez, Pedro Cano, Markus Koppenberger, and Fabien Gouyon, "Natural language processing of lyrics," in *Proceedings of the 13th Annual ACM International Conference on Multimedia*, 2005, pp. 475–478.

[5] Bin Wei, U Rochester, Chengliang Zhang, and Mitsunori Ogihara, "Keyword generation for lyrics," *Marriage*, vol. 195, no. 47, pp. 64, 2007.

[6] Florian Kleedorfer, Peter Knees, and Tim Pohle, "Oh oh oh whoah! towards automatic topic detection in song lyrics.," *Proceedings of the 9th International Conference on Music Information Retrieval*, pp. 287–292, 2008.

[7] Kahyun Choi, Jin Ha Lee, and J Stephen Downie, "What is this song about anyway?: Automatic classification of subject using user interpretations and lyrics," in *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*. IEEE Press, 2014, pp. 453–454.

[8] Kahyun Choi, Jin Ha Lee, Craig Willis, and J Stephen Downie, "Topic modeling users' interpretations of songs to inform subject access in music digital libraries," in *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*. ACM, 2015, pp. 183–186.

[9] Kahyun Choi, Jin Ha Lee, Xiao Hu, and J Stephen Downie, "Music subject classification based on lyrics and user interpretations," in *Proceedings of the American Society for Information Science and Technology*, 2016, pp. 1–10.

[10] Kahyun Choi, Jin Ha Lee, and J Stephen Downie, "Exploratory investigation of word embedding in song lyric topic classification: promising preliminary results," *Proceedings of the 18th ACM/IEEE-CS Joint Conference on Digital Libraries*, 2018.

[11] Shoto Sasaki, Kazuyoshi Yoshii, Tomoyasu Nakano, Masataka Goto, and Shigeo Morishima, "Lyrics radar: A lyrics retrieval system based on latent topics of lyrics.," in *Proceedings of the 15th International Conference on Music Information Retrieval*, 2014, pp. 585–590.

[12] Lucas Sterckx, Thomas Demeester, Johannes Deleu, Laurent Mertens, and Chris Develder, "Assessing quality of unsupervised topics in song lyrics," *European Conference on Information Retrieval*, pp. 547–552, 2014.

[13] Robert J Ellis, Zhe Xing, Jiakun Fang, and Ye Wang, "Quantifying lexical novelty in song lyrics," in *Proceedings of the 15th International Conference on Music Information Retrieval*, 2015, pp. 694–700.

[14] Robert Neumayer and Andreas Rauber, "Integration of text and audio features for genre classification in music information retrieval," in *Proceedings of the European Conference on Information Retrieval*. Springer, 2007, pp. 724–727.

[15] Rudolf Mayer, Robert Neumayer, and Andreas Rauber, "Rhyme and style features for musical genre classification by song lyrics.," *Proceedings of the 9th International Conference on Music Information Retrieval*, pp. 337–342, 2008.

[16] Alexandros Tsaptsinos, "Lyrics-based music genre classification using a hierarchical attention network," *arXiv preprint arXiv:1707.04678*, 2017.

[17] Cyril Laurier, Jens Grivolla, and Perfecto Herrera, "Multimodal music mood classification using audio and lyrics," in *Proceedings of the International Conference on Machine Learning and Applications*. IEEE, 2008, pp. 688–693.

[18] Xiao Hu, J Stephen Downie, and Andreas F Ehmann, "Lyric text mining in music mood classification," *Proceedings of the 10th International Conference on Music Information Retrieval*, 2009.

[19] Xiao Hu and J Stephen Downie, "Improving mood classification in music digital libraries by combining lyrics and audio," in *Proceedings of the 10th Annual Joint Conference on Digital Libraries*. ACM, 2010, pp. 159–168.

[20] Xiao Hu and J Stephen Downie, "When lyrics outperform audio for music mood classification: A feature analysis.," in *Proceedings of the 11th International Conference on Music Information Retrieval*, 2010, pp. 619–624.

[21] Jey Han Lau and Timothy Baldwin, "An empirical evaluation of doc2vec with practical insights into document embedding generation," *arXiv preprint arXiv:1607.05368*, 2016.

[22] Yi Yu, Suhua Tang, Francisco Raposo, and Lei Chen, "Deep cross-modal correlation learning for audio and lyrics in music retrieval," *arXiv preprint arXiv:1711.08976*, 2017.

[23] Xiao Hu, J Stephen Downie, Kris West, and Andreas F Ehmann, "Mining music reviews: Promising preliminary results," in *Proceedings of the 6th International Conference on Music Information Retrieval*, 2005.

[24] Markus Schedl, "Web-based and community-based music information extraction," *LiT. OgiharaM. TzanetakisG.(Eds.), Music Data Mining*, pp. 219–249, 2011.

[25] Sergio Oramas, Luis Espinosa-Anke, Aonghus Lawlor, et al., "Exploring customer reviews for music genre classification and evolutionary studies," in *Proceedings of the 17th International Society for Music Information Retrieval Conference*, 2016.

[26] Sergio Oramas, Oriol Nieto, Francesco Barbieri, and Xavier Serra, "Multi-label music genre classification from audio, text, and images using deep features," *arXiv preprint arXiv:1707.04916*, 2017.

[27] Sebastian Streich, *Music complexity: A multi-faceted description of audio content*, Ph.D. thesis, Universitat Pompeu Fabra, 2006.

[28] Keith Rayner and Susan A Duffy, "Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity," *Memory & Cognition*, vol. 14, no. 3, pp. 191–201, 1986.

[29] William H DuBay, "The principles of readability," *Impact Information*, 2004.

[30] Elfrieda H Hiebert, *Readability and the Common Core's staircase of text complexity*, Santa Cruz, CA: TextProject Inc, 2012.

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

[32] George R Klare, "The measurement of readability: useful information for communicators," *ACM Journal of Computer Documentation*, vol. 24, no. 3, pp. 107–121, 2000.

[33] Janice Redish, "Readability formulas have even more limitations than klare discusses," *ACM Journal of Computer Documentation*, vol. 24, no. 3, pp. 132–137, 2000.

[34] Nancy Frey and Douglas Fisher, *Rigorous reading: 5 access points for comprehending complex texts*, Corwin Press, 2013.

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

[36] Songmeanings.com, "Songmeanings main page," http://songmeanings.com/, Retrieved January 30, 2017.

[37] Genious.com, "Genious.com main page," http://genius.com/, Retrieved January 30, 2017.

[38] Lyricinterpretations.com, "Lyricinterpretations.com main page," http://lyricinterpretations.com/, Retrieved January 30, 2017.

[39] Similarweb.com, "Similarweb.com main page," https://www.similarweb.com/, Retrieved September 5, 2016.

[40] Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman, "Concreteness ratings for 40 thousand generally known English word lemmas," *Behavior Research Methods*, vol. 46, no. 3, pp. 904–911, 2014.

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

[42] Yoav Goldberg and Omer Levy, "word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method," *arXiv preprint arXiv:1402.3722*, 2014.

[43] Robert Mitchell Parry, "Musical complexity and top 40 chart performance," Tech. Rep., Georgia Institute of Technology, 2004.

[44] Sebastian Streich and Perfecto Herrera, "Detrended fluctuation analysis of music signals: Danceability estimation and further semantic characterization," in *Proceedings of the 118th Audio Engineering Society Convention*, 2005.

[45] Aline K Honingh and Rens Bod, "Pitch class set categories as analysis tools for degrees of tonality," in *Proceedings of the 11th International Conference on Music Information Retrieval*, 2010, pp. 459–464.

[46] Matthias Mauch and Mark Levy, "Structural change on multiple time scales as a correlate of musical complexity," in *Proceedings of the 12th International Conference on Music Information Retrieval*, 2011, pp. 489–494.

[47] Manuela M Marin and Helmut Leder, "Examining complexity across domains: relating subjective and objective measures of affective environmental scenes, paintings and music," *PLoS One*, vol. 8, no. 8, pp. e72412, 2013.

[48] Peter Foster, Matthias Mauch, and Simon Dixon, "Sequential complexity as a descriptor for musical similarity," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 1965–1977, 2014.

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

[50] Christof Weiß and Meinard Müller, "Quantifying and visualizing tonal complexity," in *Proceedings of the Conference on Interdisciplinary Musicology*, 2014.

[51] Christof Weiss and Meinard Müller, "Tonal complexity features for style classification of classical music," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 2015, pp. 688–692.

[52] Guillaume Robal and Tim Blackwell, "Live algorithms with complexity matching," Tech. Rep., University of London, 2014.

[53] Junghyuk Lee and Jong-Seok Lee, "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, 2015, pp. 3–6.

[54] Bruno Di Giorgi, Simon Dixon, Massimiliano Zanoni, and Augusto Sarti, "A data-driven model of tonal chord sequence complexity," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 11, pp. 2237–2250, 2017.

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

[56] J Stephen Downie, "Music information retrieval," *Annual Review of Information Science and Technology*, vol. 37, no. 1, pp. 295–340, 2003.

[57] Karl Kristoffer Jensen and David Hebert, "Predictability of harmonic complexity across 75 years of popular music hits," in *Proceedings of the 11th International Symposium on Computer Music Multidisciplinary Research*. The Laboratory of Mechanics and Acoustics, 2015, pp. 198–212.

[58] Takuya Fujishima, "Real-time chord recognition of musical sound: A system using common lisp music," in *Proceedings of the International Computer Music Conference*, 1999, pp. 464–467.

[59] Emilia Gómez and Perfecto Herrera, "Estimating the tonality of polyphonic audio files: Cognitive versus machine learning modelling strategies," *Proceedings of the 5th International Conference on Music Information Retrieval*, 2004.

[60] Marc Brysbaert and Boris New, "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*, vol. 41, no. 4, pp. 977–990, 2009.

[61] Xiaobin Chen and Detmar Meurers, "Characterizing text difficulty with word frequencies," in *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, 2016, pp. 84–94.

[62] Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai, "Coh-metrix: Analysis of text on cohesion and language," *Behavior Research Methods, Instruments, & Computers*, vol. 36, no. 2, pp. 193–202, 2004.

[63] Marci Glaus, "Text complexity and young adult literature: Establishing its place," *Journal of Adolescent & Adult Literacy*, vol. 57, no. 5, pp. 407–416, 2014.

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

[65] Rudolph Flesch, "A new readability yardstick.," *Journal of Applied Psychology*, vol. 32, no. 3, pp. 221, 1948.

[66] Edgar Dale and Jeanne S Chall, "A formula for predicting readability: Instructions," *Educational Research Bulletin*, pp. 37–54, 1948.

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

[68] Nina H Macdonald, Lawrence T Frase, Patricia S Gingrich, and Stacey A Keenan, "The writer's workbench: Computer aids for text analysis," *Educational Psychologist*, vol. 17, no. 3, pp. 172–179, 1982.

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

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

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

[72] Kathleen M Sheehan, Irene Kostin, Diane Napolitano, and Michael Flor, "The textevaluator tool: Helping teachers and test developers select texts for use in instruction and assessment," *The Elementary School Journal*, vol. 115, no. 2, pp. 184–209, 2014.

[73] George R Klare, "The role of word frequency in readability," *Elementary English*, vol. 45, no. 1, pp. 12–22, 1968.

[74] Edward L Thorndike, *The Teacher's Word Book*, Teacher's College, Columbia University, 1921.

[75] Edward Fry, "A readability formula that saves time," *Journal of Reading*, vol. 11, no. 7, pp. 513–578, 1968.

[76] G Harry Mc Laughlin, "Smog grading-a new readability formula," *Journal of Reading*, vol. 12, no. 8, pp. 639–646, 1969.

[77] Edgar Dale, "Evaluating thorndike's word list," *Educational Research Bulletin*, pp. 451–457, 1931.

[78] Gondy Leroy and David Kauchak, "The effect of word familiarity on actual and perceived text difficulty," *Journal of the American Medical Informatics Association*, vol. 21, no. e1, pp. e169–e172, 2013.

[79] Arthur C Graesser, Danielle S McNamara, and Jonna M Kulikowich, "Coh-metrix: Providing multilevel analyses of text characteristics," *Educational Researcher*, vol. 40, no. 5, pp. 223–234, 2011.

[80] Kirill Kireyev and Thomas K Landauer, "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, 2011, pp. 299–308.

[81] David M Blei, Andrew Y Ng, and Michael I Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[82] David Hall, Daniel Jurafsky, and Christopher D Manning, "Studying the history of ideas using topic models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2008, pp. 363–371.

[83] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li, "Comparing twitter and traditional media using topic models," in *Proceedings of the European Conference on Information Retrieval*. Springer, 2011, pp. 338–349.

[84] Anton Barua, Stephen W Thomas, and Ahmed E Hassan, "What are developers talking about? An analysis of topics and trends in stack overflow," *Empirical Software Engineering*, vol. 19, no. 3, pp. 619–654, 2014.

[85] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin, "Automatic evaluation of topic coherence," in *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 100–108.

[86] Claudia Monica Ferradas Moi, "Rock poetry: The literature our students listen to.," *Journal of the Imagination in Language Learning*, pp. 56–59, 1994.

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

[88] Jeanette Altarriba, Lisa M Bauer, and Claudia Benvenuto, "Concreteness, context availability, and imageability ratings and word associations for abstract, concrete, and emotion words," *Behavior Research Methods*, vol. 31, no. 4, pp. 578–602, 1999.

[89] Mark Sadoski, Ernest T Goetz, and Joyce B Fritz, "Impact of concreteness on comprehensibility, interest, and memory for text: Implications for dual coding theory and text design.," *Journal of Educational Psychology*, vol. 85, no. 2, pp. 291, 1993.

[90] Mark Sadoski, "Resolving the effects of concreteness on interest, comprehension, and learning important ideas from text," *Educational Psychology Review*, vol. 13, no. 3, pp. 263–281, 2001.

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

[92] 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., "Quantitative analysis of culture using millions of digitized books," *Science*, vol. 331, no. 6014, pp. 176–182, 2011.

[93] Mark Davies, "The 385+ million word corpus of contemporary American English (1990–2008+): Design, architecture, and linguistic insights," *International Journal of Corpus Linguistics*, vol. 14, no. 2, pp. 159–190, 2009.

[94] Peter Sheridan Dodds and Christopher M Danforth, "Measuring the happiness of large-scale written expression: Songs, blogs, and presidents," *Journal of Happiness Studies*, vol. 11, no. 4, pp. 441–456, 2010.

[95] Margaret M Bradley and Peter J Lang, "Affective norms for English words (ANEW): Instruction manual and affective ratings," Tech. Rep., Citeseer, 1999.

[96] Marc Lafrance, Lara Worcester, and Lori Burns, "Gender and the billboard top 40 charts between 1997 and 2007," *Popular Music and Society*, vol. 34, no. 5, pp. 557–570, 2011.

[97] Marc Lafrance, Casey Scheibling, Lori Burns, and Jean Durr, "Race, gender, and the billboard top 40 charts between 1997 and 2007," *Popular Music and Society*, pp. 1–17, 2017.

[98] Peter G Christenson, Silvia de Haan-Rietdijk, Donald F Roberts, and Tom FM ter Bogt, "What has America been singing about? Trends in themes in the us top-40 songs: 1960–2010," *Psychology of Music*, 2018.

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

[100] Alexandros Tsaptsinos, "Lyrics-based music genre classification using a hierarchical attention network," in *Proceedings of the 18th International Conference on Music Information Retrieval*, 2017.

[101] 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, "The English lexicon project," *Behavior Research Methods*, vol. 39, no. 3, pp. 445–459, 2007.

[102] Emmanuel Keuleers, Paula Lacey, Kathleen Rastle, and Marc Brysbaert, "The british lexicon project: Lexical decision data for 28,730 monosyllabic and disyllabic English words," *Behavior Research Methods*, vol. 44, no. 1, pp. 287–304, 2012.

[103] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Association for Computational Linguistics System Demonstrations*, 2014, pp. 55–60.

[104] David Roxbee Cox and Alan Stuart, "Some quick sign tests for trend in location and dispersion," *Biometrika*, vol. 42, no. 1/2, pp. 80–95, 1955.

[105] Marc Lavielle, "Using penalized contrasts for the change-point problem," *Signal Processing*, vol. 85, no. 8, pp. 1501–1510, 2005.

[106] Rebecca Killick, Paul Fearnhead, and Idris A Eckley, "Optimal detection of changepoints with a linear computational cost," *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1590–1598, 2012.

[107] W Sichert-Hellert, M Kersting, U Alexy, and F Manz, "Ten-year trends in vitamin and mineral intake from fortified food in german children and adolescents," *European Journal of Clinical Nutrition*, vol. 54, no. 1, pp. 81, 2000.

[108] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov, "Enriching word vectors with subword information," *arXiv preprint arXiv:1607.04606*, 2016.

[109] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[110] Cedric De Boom, Steven Van Canneyt, Thomas Demeester, and Bart Dhoedt, "Representation learning for very short texts using weighted word embedding aggregation," *Pattern Recognition Letters*, vol. 80, pp. 150–156, 2016.

[111] Xingquan Zhu, *Knowledge Discovery and Data Mining: Challenges and Realities: Challenges and Realities*, IGI Global, 2007.

[112] Francisco J Valverde-Albacete and Carmen Peláez-Moreno, "100% classification accuracy considered harmful: The normalized information transfer factor explains the accuracy paradox," *PloS One*, vol. 9, no. 1, pp. e84217, 2014.

[113] J Stephen Downie, "The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research," *Acoustical Science and Technology*, vol. 29, no. 4, pp. 247–255, 2008.

[114] Kevin Bache, David Newman, and Padhraic Smyth, "Text-based measures of document diversity," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2013, pp. 23–31, ACM.

[115] Anne E Magurran, *Why diversity?*, Springer, 1988.

[116] Hosein Azarbonyad, Ferron Saan, Mostafa Dehghani, Maarten Marx, and Jaap Kamps, "Are topically diverse documents also interesting?," in *Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages*. 2015, pp. 215–221, Springer.

[117] C Radhakrishna Rao, "Diversity and dissimilarity coefficients: a unified approach," *Theoretical Population Biology*, vol. 21, no. 1, pp. 24–43, 1982.

[118] Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al., *Introduction to Information Retrieval*, Cambridge university press Cambridge, 2008.

[119] Philip R Burns, "Morphadorner v2: A java library for the morphological adornment of English language texts," *Northwestern University*, 2013.

[120] Andrew Kachites McCallum, "Mc," http://mallet.cs.umass.edu, 2002.

[121] Chade-Meng Tan, Yuan-Fang Wang, and Chan-Do Lee, "The use of bigrams to enhance text categorization," *Information Processing & Management*, vol. 38, no. 4, pp. 529–546, 2002.

[122] Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth, "Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter," *PloS one*, vol. 6, no. 12, pp. e26752, 2011.

[123] Daniel E Berlyne, *Aesthetics and psychobiology*, Appleton-Century-Crofts, 1973.

[124] Adrian C North and David J Hargreaves, "Subjective complexity, familiarity, and liking for popular music.," *Psychomusicology: A Journal of Research in Music Cognition*, vol. 14, no. 1-2, pp. 77, 1995.

[125] Chris Anderson, "The long tail," *Wired magazine*, pp. 170–177, October 2004.

[126] Markus Schedl, Arthur Flexer, and Julián Urbano, "The neglected user in music information retrieval research," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 523–539, 2013.

[127] Cheng-Che Lu and Vincent S Tseng, "A novel method for personalized music recommendation," *Expert Systems with Applications*, vol. 36, no. 6, pp. 10035–10044, 2009.

[128] Ju-Chiang Wang, Yi-Hsuan Yang, Hsin-Min Wang, and Shyh-Kang Jeng, "Personalized music emotion recognition via model adaptation," in *Proceedings of Signal & Information Processing Association Annual Summit and Conference*. IEEE, 2012, pp. 1–7.

[129] Carl I Belz, "Popular music and the folk tradition," *The Journal of American Folklore*, vol. 80, no. 316, pp. 130–142, 1967.

[130] Michael Flor, Beata Beigman Klebanov, and Kathleen M Sheehan, "Lexical tightness and text complexity," in *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*, 2013, pp. 29–38.

[131] Jin Ha Lee, "Crowdsourcing music similarity judgments using mechanical turk," in *Proceedings of the 11th International Conference on Music Information Retrieval*, 2010.

[132] Michael I Mandel, Douglas Eck, and Yoshua Bengio, "Learning tags that vary within a song," in *Proceedings of the 9th International Conference on Music Information Retrieval*, 2010.

[133] Julián Urbano, Jorge Morato, Mónica Marrero, and Diego Martín, "Crowdsourcing preference judgments for evaluation of music similarity tasks," *ACM SIGIR Workshop on Crowdsourcing for Search Evaluation*, 2010.

[134] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling, "Amazon's mechanical turk a new source of inexpensive, yet high-quality, data?," *Perspectives on Psychological Science*, vol. 6, no. 1, pp. 3–5, 2011.

[135] Aniket Kittur, Ed H Chi, and Bongwon Suh, "Crowdsourcing user studies with mechanical turk," in *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2008, pp. 453–456.