Lyrics Analysis of Korean Hit Songs

Wongsang Kwak (20173022), Nakyung Lee (20173388), Jeongwhan Oh (20183362), Poyan Lai (20184166), Joongun Park (20185128)

1. Introduction

Lyrics might be more powerful than we thought. The American singer-songwriter Bob Dylan not only touches many peoples' hearts with his beautifully written, poetic lyrics, but even won himself a Nobel Prize in Literature in 2016. As a key element of music, lyrics plays a big role in conveying the message of a song which emotionally influences people a lot. In return, people often consider lyrics as an expression of their own emotions. While music is an important part of our cultural life, we believe that lyrics worth an in-depth investigation into it.

When investigating Korean songs in recent years, we observed that many of them have repetitive lyrics. Lyricists in Korea tend to used simple, somehow meaningless, but catchy phrases repeatedly to make the songs more addictive. As the length of a song is limited, when the same word is used over again within the song, it is reasonable to assume that the diversity of words will decrease. It makes us wonder how diverse is the lyrics of Korean hit songs.

Lexical diversity of a text is often analysed in linguistics studies. Johansson (2008) defines it as the measure of how many different words that are used in a text. Jarvis (2013) pointed out that it is also known as variability, which is usually operationalized into measures capturing the proportion of unique words in a text. On top of diversity, the lexical sophistication and richness, which refer to the number of rare words and the number of word types respectively, are another two commonly used measures.

Jarvis proposed that language researchers should consider seven properties of diversity, which include size, richness, effective number of types, evenness, disparity, importance and dispersion. Table1 captures how he defined each of this property along with the possible measures that can be used to calculate that property.

Property	Measure
Size	Number of tokens
Richness	Number of types
Effective number of types	The exponential function applied to Shannon's index
Evenness	The degree to which tokens are distributed equally across types
Disparity	The proportion of words in a text that are semantically related
Importance	The relative frequency with which the words in a text occur in the language as a whole
Dispersion	The average interval between tokens of the same type

Table1. Properties of Lexical Diversity And Their Measures

We developed our model for statistical analysis based on the above properties proposed by Jarvis[1][10][11]. In this project, we examined whether the lyrics of Korean hit songs have an improving diversity in the recent two decades. Based on our observation that most songs top the chart are addictive songs, our hypothesis is that lyrics of Korean hit songs do not have improving diversity. In the following, we will discuss our dataset and methods in detail, followed by our findings.

2. Methodology

Data Preparation

To gather the dataset, we conducted data crawling with Python on the website of the music service provider Melon. Melon records Monthly Active User(MAU) 5.2 million in 2017 in mobile(Android), which takes 52% of MAU in top 8 music streaming services in Korea[8]. Either it takes 39% in survey for music streaming service used in Korea[2]. We chose Melon to ensure our data will be reliable and representative enough because Melon has the highest market share.

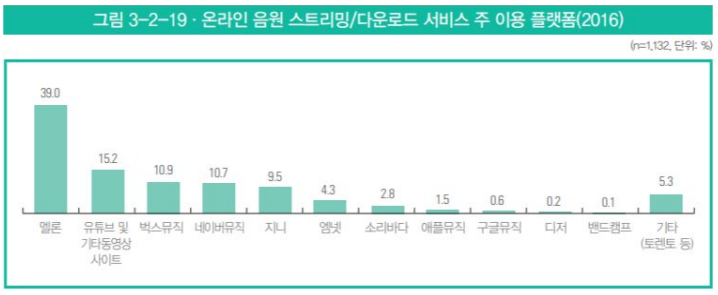


Figure 1. Survey on usage of music streaming services

We extracted the title, song ID and lyrics of the Top 50 songs from the monthly chart from 2000 to 2017. After parsing the documents with BeautifulSoup package[3], we formed the dataset of 4654 songs with the year they are on the chart, the title, the ID and the lyrics. Apart from the total dataset, we also created subset data according to the year for the convenience of data analysis.

Before tidying up the text, we first conducted morphological analysis. Morphological analysis is the study of the internal structure of words. In Korean, the same word can appear as different part of speech. At the same time, some words are just a part of the syntax without any semantic meaning. The general approach of morphological analysis is to identify the meaning of words by their part of speech according to the context, and then conduct the text segmentation by their meaning. To be specific, the same word would be categorized as different word types as listed in Table 2[5]. It is therefore a necessary step to avoid ambiguity and select only meaningful corpora from the text.

To classify tags for corpora and extract them, we used KoNLP library which is an open source project for natural language processing. We used KoNLP library and mecab-ko analyzer[4] both in R and Python code. The usages and details are in the code that we attached.

· Common nouns · Proper nouns · Bound nouns · Numerals · Pronouns · Verbs · Adjectives	· Auxiliary predicate · Positive copula · Negative copula · Determinatives · Adverbs · Exclamations

Table2. Examples of Word types (Part of Speech)

After extracting all the meaningful corpora from the text, we developed the model to analyse the text based on the properties of lexical diversity defined by Jarvis[1]. Our model includes four factors, which are variability, volume, evenness and rarity, used to explain the difference of lyrics between each year.

Data Analysis

After evaluating different measures, functions and libraries available in R for diversity analysis, we choose four factors based on prior works[1][9][10] which define lexical diversity for their purpose. We are capable to operate, including variability, volume, evenness and rarity to examine the difference of lyrics between each year. We computed the data by this four factors and then conducted the Multivariate Analysis of Variance (MANOVA) to test whether the four factors are sufficient to explain the difference.The operational definition of the four factors is described as below.

Variability. As mentioned, lexical diversity is also known as variability, meaning the number or the proportion of unique words[6]. In other words, the higher number or proportion of unique words, the higher its variability. By this definition, we computed the number of unique words of our corpora.

Volume. This is similar to variability, but volume refers to the size of the corpora, i.e. the total number of words in the lyrics. The higher number of words in the lyrics might also imply a higher lexical diversity of the text because a longer text must contain more words than a shorter text. By this definition[10], we computed the total number of words of our corpora.

Evenness. As shown in Table 1, evenness is defined as the degree to which tokens are distributed equally across types. One of the measures is to calculate the standard deviation of words per types[11]. By examining the standard deviation of words per types, it is able to know the variation of words per type. A larger variation indicates a higher diversity. In our implementation, we extracted the number of word types and computed the evenness value according to the formula below:

$$Evenness = \frac{1}{Standard\ Deviation\ (SD)}$$

Figure 2. The formula of Evenness

Rarity. In our model, rarity is operationally defined as the relative frequency with which the words in a text occur in the language as a whole. Jarvis[1] ranks a word's rarity with British National Corpus(BNC)[12] which has 100 million-word text samples in English. Since we are focusing on analysing Korean lyrics, we computed the Term Frequency-Inverse Document Frequency (TF-IDF)[13][14] for measuring the rarity instead of BNC rank. TF-IDF is a

measure which reflects how important a word is to a document in a corpus, which can be calculated by the following formula:

$$tfidf_x = tf_x \times \log(\frac{N}{df_x})$$

where

tf_x = Frequency of a **term x** occurs in the lyrics each year

df_x = Number of **songs** containing each **term x** in other years

idf_x = Inverse of df

N = Total number of **songs**

Figure 3 The formula of TF-IDF

In the implementation, we computed the TF-IDF value for all words, which reflects how important that word in that particular year comparing to other years. In other words, the closer the TF-IDF value to zero, the less important is that word in that year. The word with the highest value within particular year implies that it is a particularly unique word being used in that year.

3. Results & Discussion

Variability

Figure 4 is the boxplot which shows the number of unique words categorized by year. The green box indicates that about half of the songs of that year contain no more than 50 unique words in the lyrics. A number of outliers which are represented by dots are also visible. Figure 5 is a closer look into the songs with no more than 50 unique words. The black line is the line for linear regression, which the summary of it is shown in Figure 6, and the blue is the line plotting the mean value.

Based on the significant p-value and the adjusted R-squared value of 0.023. We can conclude that there is a positive correlation with acceptable strength between the variability of songs over the years. In other words, there is an improving lexical diversity over years in terms of variability.

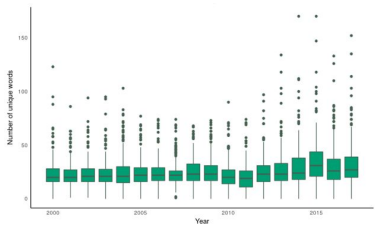


Figure 4. Boxplot (Year-Variability)

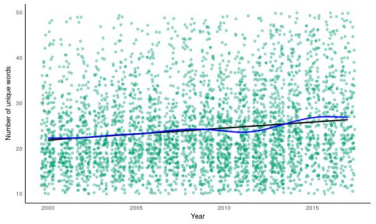


Figure 5. Scatter Plot (Year-Variability)

```
> summary(lm(MTLD.lex_diversity ~ MTLD.year))

Call:
lm(formula = MTLD.lex_diversity ~ MTLD.year)

Residuals:
    Min      1Q  Median      3Q     Max
-29.414  -8.978  -2.927   5.503 142.048

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -953.42782   93.48495  -10.20   <2e-16 ***
MTLD.year      0.48728    0.04654   10.47   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.76 on 4616 degrees of freedom
Multiple R-squared:  0.0232,    Adjusted R-squared:  0.02299
F-statistic: 109.6 on 1 and 4616 DF,  p-value: < 2.2e-16
```

Figure 6. Summary of Regression

Volume

Each dot in Figure 7 represents a song, which is positioned according to the number of words within the lyrics. The black line is the linear regression with the summary as shown in Figure 8, and the blue one is the variation of the mean value. Similarly, the significant p-value and the adjusted R-squared value of 0.026 showed that there is a positive correlation with acceptable strength between the volume of songs over the years. In other words, there is an improving lexical diversity over years in terms of volume.

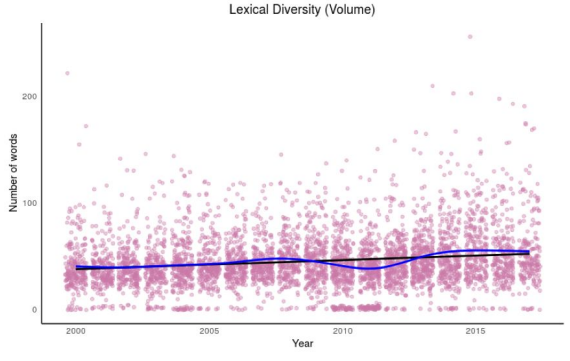


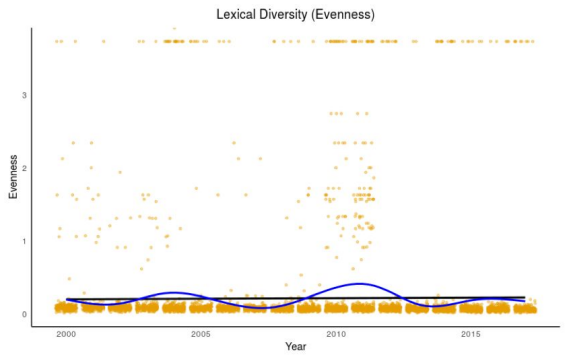
Figure 7. Scatter Plot (Year-Volume)

```
> summary(lm(Volume.lex_diversity ~ MTLD.year))

Call:
lm(formula = Volume.lex_diversity ~ MTLD.year)

Residuals:
    Min      1Q  Median      3Q     Max
-52.348 -14.570  -3.968   9.793 205.342

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1652.2438   152.0683  -10.87   <2e-16 ***
MTLD.year       0.8451     0.0757   11.16   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25.64 on 4616 degrees of freedom
Multiple R-squared:  0.02629,   Adjusted R-squared:  0.02608
F-statistic: 124.6 on 1 and 4616 DF,  p-value: < 2.2e-16
```

Figure 8. Summary of Regression

Evenness

Figure 9 depicts the standard deviation of each word type according to the year. The black line is the linear regression with the summary as shown in Figure 10, and the blue one is the variation of the mean value. Most songs are with a low value of evenness; it is also noteworthy that the mean values are greatly affected by the outliers. Yet, the insignificant p-value and R-squared value which is almost zero indicated that there is no correlation between the evenness of the songs over years. We are unable to conclude whether there is any difference over years in terms of evenness.

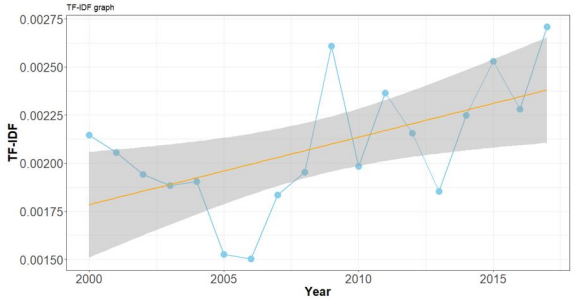



```
> summary(lm(Evenness.lex_diversity ~ MTLD.year))

Call:
lm(formula = Evenness.lex_diversity ~ MTLD.year)

Residuals:
    Min      1Q  Median      3Q     Max
-0.2133 -0.1602 -0.1383 -0.1129  3.5351

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.664155   3.709157  -0.718    0.473
MTLD.year    0.001435   0.001846   0.777    0.437

Residual standard error: 0.6253 on 4615 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.0001309, Adjusted R-squared:  -8.573e-05
F-statistic: 0.6043 on 1 and 4615 DF,  p-value: 0.437
```

Figure 9. Scatter Plot (Year-Evenness) *Figure 10. Summary of Regression*

Rarity

Figure 11 plotted the average of the TF-IDF values of words appearing in each year. The graph shows that there is an increasing trend in overall except for the drop in 2005 and 2013. The significant p-value and the adjusted R-squared value of 0.027 showed that there is a positive correlation with certain strength between the TF-IDF values of words over the years. In other words, there is an improving lexical diversity over years in terms of rarity.




```
> summary(lm(tfidf_value~year))

Call:
lm(formula = tfidf_value ~ year)

Residuals:
       Min         1Q     Median         3Q        Max
-4.932e-04 -1.419e-04 -2.332e-05  2.131e-04  5.096e-04

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.837e-02  2.612e-02  -2.617   0.0187 *
year         3.508e-05  1.301e-05   2.697   0.0159 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0002863 on 16 degrees of freedom
Multiple R-squared:  0.3125,    Adjusted R-squared:  0.2695
F-statistic: 7.273 on 1 and 16 DF,  p-value: 0.01587
```

Figure 11. Linear Plot (Year-TFIDF) *Figure 12. Summary of Regression*

MANOVA

The result of MANOVA shown in Figure 13 showed that variability, volume and rarity with significant p-value are sufficient factors to explain the difference between lyrics over years. We can therefore conclude that our model, excepting evenness, is sufficient to explain the difference between lyrics over years in terms of diversity. Since there is a positive correlation of variability, volume and rarity over years, we conclude that there is an improving diversity of Korean hit songs over the recent two decades.

```
> summary.aov(fit)
 Response 1 :
              Df  Sum Sq Mean Sq F value Pr(>F)
cdt$MTLD.year    1   27125   27125     109 <2e-16 ***
Residuals     4615 1146297     248
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response 2 :
              Df  Sum Sq Mean Sq F value Pr(>F)
cdt$MTLD.year    1   81581   81581     124 <2e-16 ***
Residuals     4615 3032829     657
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response 3 :
              Df Sum Sq Mean Sq F value Pr(>F)
cdt$MTLD.year    1      0   0.236     0.6   0.44
Residuals     4615   1805   0.391

 Response 4 :
              Df   Sum Sq  Mean Sq F value Pr(>F)
cdt$MTLD.year    1 0.000133 1.33e-04    1843 <2e-16 ***
Residuals     4615 0.000334 1.00e-07
```

Figure 13. Summary of MANOVA

4. Limitations & Future Work

There are a few limitations of our findings. In this section, we will discuss the limitations of our study, the future work we might do and the contribution of this study.

Firstly, the dataset we used would be incomplete because there are a number of missing lyrics data on Melon. Since it is time-consuming to examine and amend the lyrics data one by one, there is a possibility that we might miss out some important lyrics data.

Secondly, we considered mainly the frequency of tokens over the year instead of the context of each song. However, there might be some words which are highly correlated and tend to co-occur even in different songs. As a result, we are unable to capture the hidden networks of different tokens in our project.

Thirdly, our methods to compute each measure of diversity is imperfect. There are many existing methods devised by previous researchers to compute each of the property of lexical diversity. We are forced to give up some of the more reliable measures and methods due to the time limits of the project, as well as the technical difficulties we encountered. Therefore, there is a room of improvement for the methods we used.

Except the above limitations mentioned, we discovered that it will also be interesting to conduct sentiment analysis of Korean lyrics. There seems to be a trend that songs released in the same season are more similar in terms of the atmosphere and sentiment. For example, songs of Spring and Summer tend to be brighter and more joyful while songs of Autumn and Fall tend to be more emotional. These kind of patterns might be discovered through a systematic investigation.

Lastly, Jarvis[1] pointed out that the range, variety, or diversity of words found in learners' language use is believed to reflect the complexity of their vocabulary knowledge as well as the level of their language proficiency. As we mentioned, people are influenced a lot by lyrics. If the lexical diversity of lyrics can be further increased, it might also help people to improve their language ability. Our analysis contributed to the understanding of the lexical diversity of Korean lyrics, we hope to see a continually improving diversity of lyrics.

References

[1] JARVIS, Scott. Capturing the diversity in lexical diversity. *Language Learning*, 2013, 63.s1: 87-106. Available:https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1467-9922.2012.00739.x
[2] Korea Creative Content Agency (KOCCA), '2016 MUSIC INDUSTRY WHITE PAPER', 2017. Available:http://www.kocca.kr/industry/16_industry_m_3_2.pdf
[3] Leonard Richardson, 'Beautiful Soup', 2017. Available: https://www.crummy.com/software/BeautifulSoup/.
[4] Junhewk Kim, 'RmecabKo', 2018. Available: https://github.com/junhewk/RmecabKo/blob/master/readme.rmd.
[5] Lucy Park, 'Korean POS tags comparison chart', 2014. Available: https://docs.google.com/spreadsheets/d/1OGAjUvalBuX-oZvZ_-9tEfYD2gOe7hTGsgUpiiBSXI8/edit#gid=0.
[6] Debbie Liske, 'Tidy Sentiment Analysis in R', 2018. Available: https://www.datacamp.com/community/tutorials/sentiment-analysis-R.
[7] Junhewk Kim, '한국 가요 50년사, 가사 분석', 2017. Available: https://junhewk.github.io/text/2017/11/10/meloncharts-lyrics/.
[8] Taekyun Kim, '모바일 음원서비스 1위는 멜론... 월 실사용자 522만명', 2017. Available: http://www.yonhapnews.co.kr/bulletin/2017/02/28/0200000000AKR20170228043900017.HTML
[9] McCarthy, P. M. (2005). An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD) [Microfiche]. Doctoral dissertation, University of Memphis.
[10] Wikipedia, Lexical diversity, Available: https://en.wikipedia.org/wiki/Lexical_diversity.
[11] JARVIS, Scott. Vocabulary Knowledge: Human Ratings and Automated Measures(pp. 29), 2013.
[12] Wikipedia, British National Corpus. Available: https://en.wikipedia.org/wiki/British_National_Corpus.
[13] Wikipedia, tf-idf, Available: https://en.wikipedia.org/wiki/Tf%E2%80%93idf
[14] cran.r-prject, koRpus, Available : https://cran.r-project.org/web/packages/koRpus/vignettes/koRpus_vignette.pdf