K-pop music video popularity as influenced by English in song lyrics

William M. Scott
(Dated: March 6, 2022)

Korean music has become a powerful economic and cultural phenomenon worldwide. Interestingly, in spite of the name, most Korean pop (K-pop) songs contain English to varying degrees. It is hypothesized that the fraction of English words within K-pop songs might impact how successful they are (measured by music video views on Youtube in this paper). In particular, it is supposed that more English should make a song more internationally accessible and should therefore increase its likelihood of reaching a wider audience. Contrary to this more specific hypothesis, however, it is shown at a 95% confidence level that K-pop songs with only Korean lyrics outperform their pure-English counterparts on average. It is subsequently demonstrated that if ϕ_{E} is the fraction of English words in a K-pop song, then the most successful songs have $\phi_{\mathrm{E}} = 0.40$ on average. A 95% confidence interval is constructed for optimal ϕ_{E}, resulting in $(0.37, 0.43)$.

I. INTRODUCTION

There has been a remarkable rise in the popularity of Korean pop music (K-pop) internationally over a relatively short period of time. This incredible success is evidenced by the fact that individual K-pop production agencies have begun to generate the equivalent of hundreds of millions of USD in revenue each year [1–3].

Such success, no doubt, may be attributed to a complex interplay of factors. In examining what sets K-pop apart from music industries in other parts of the world, a few elements come to mind: the high production values, the large amount of artist-focused ancillary content, and the relatively strong focus on merchandise over digital sales—in essence turning the music into the advertisement, rather than the product. One thing that goes less discussed, however, is the preponderance of English lyrics within the songs.

Those outside Korea who are first introduced to the genre might be surprised to find that not only do a majority of successful K-pop songs contain at least some English, but that often English lyrics make up an appreciable fraction, sometimes even most, of the overall word count. For comparison, Japan, a neighbouring country with its own well-developed music industry, produces pop songs with markedly less English; and though the use of English in songs is quite common across Asia [4], K-pop stands out nevertheless, particularly when one considers just how many K-Pop songs are written entirely in English or get fully-translated English releases. It is thus hypothesized that part of K-pop's international success might be down to its extensive use of the world's default lingua franca.

In order to ascertain what sort of effect the inclusion of English has on the success of K-pop songs, some success metric is needed. The best metric would be the revenue generated by a particular song, but those data are not readily available. In this paper, then, music video views on Youtube are used as a surrogate measure of success and are compared against the fraction of English words within K-pop songs. The results of this comparison are then used to predict an optimal level of English content to help maximize a song's chance of success.

II. DATA COLLECTION AND REFINEMENT

As a starting point, a large database was found on `kaggle.com` [5] containing a list of 3762 K-pop songs. Within the database, each song had listed the artist name, the song name, the release date, the so-called *release type*, and the URL associated with the song's music video on Youtube. Other descriptors were available within the database but are not relevant to this study.

Neither the video views nor any sort of indication of the fraction of English within the songs was originally listed in the database. Therefore, both the video views and the word-by-word breakdown of the songs were scraped from the Internet. The listed Youtube video URLs were used to gather the number of views each song had on Dec. 1, 2021. As for the lyrics, the artists and song names were copied into automatic queries to Google, and the words comprising each song were then extracted from the query results.

The song release dates vary widely, with the earliest being released on March 23, 1992 and the latest on May 5, 2020. In preliminary data analysis, it was decided that early releases, many of which necessarily had to be uploaded to Youtube after their original time of release, should not be included in the final analysis. There were two reasons for this decision. First, the views of such videos are all considerably lower on average than more recent releases. Second, in selecting which older songs to include in the original database, the people who compiled it were no doubt subject to forms of bias, such as favouring more popular songs. While it goes without saying that more recent songs in the database should also suffer from a selection bias based on popularity, they should all suffer from it in more or less the *same way*. To deal with these issues and to lessen the effect of a song's release date acting as a confounding variable, it was arbitrarily but conservatively decided to only include songs released on or after Jan. 1, 2012 in the final analysis.

To further homogenize the data, it was decided to only make use of songs of a particular release type. The release types in the database were as follows: Major, Minor, Japanese, English, Special (as in holiday specials), CF (a release meant specifically to advertise a third-party

product), and OST (meaning *original soundtrack* and typically used for film or tv). The most common release type in the database was the major release, which included the most influential K-pop songs, so it was chosen. One might wonder why the English releases were left out, given the topic of this paper. English releases are typically songs that may or may not have had English in them originally but that got fully-translated releases later on. Such secondary releases do not fall within the scope of this research.

In considering only major releases with release dates falling on or after Jan. 1, 2012, the number of songs was cut from 3762 to 2339. Of these 2339 songs, 69 had listed URLs that no longer linked to Youtube videos but ran successful lyrics queries, 600 had operational links but unsuccessful lyrics queries, and an additional 72 ran into both problems. This further reduced the number of songs in the final analysis to 1598. It should be noted that the retrieval of the lyrics was successful or unsuccessful in the same way that a typical query including the artist, song name, and the word *lyrics* typed directly into the Google search engine would succeed or fail to bring up the song lyrics within the search engine itself. This indicates that the final data are likely biased more toward popular songs, which is deemed acceptable for the purposes of this study because it is unlikely that equally unpopular songs with differing English makeup should fail such lyrics queries at different rates.

As a final note on data collection, some caveats should be provided in relation to the word counts. Similar to English, but unlike many other East Asian languages, Korean features white space between its words. It is therefore relatively simple to automate the word counting process. Linguists might argue that in some cases one ought to count morphemes rather than words [6], or that an overall count of words might not be a good indicator of how much of a song might be in one language or another. After all, what does it mean to ask how much of a text is in a particular language? Are we counting time spent in the language by syllable? In that case, word count would not necessarily be a good estimate of the "amount" of the language. What if one language tends to convey more information per word than another, regardless of the time it takes to utter each word? In considering such questions, it becomes clear that judging the strict "amount" of a language within a multilingual text is difficult and open to interpretation. Even so, it is argued that for the purposes of this study, such concerns are irrelevant. The goal here is to compare song performance based on relative English content, so in comparing songs, if one has 20% English words and another has 10% English words, then while it might not be incredibly meaningful to say that the former is 20% English, it *is* meaningful to say that the former has more English than the latter in a precisely quantifiable way.

Besides the above issue regarding how the frequency of words within a multilingual text is interpreted, there is also a small matter related to the way the words were tallied. Words were counted as being either Korean or English without consideration for other languages. Any word containing only unaccented Latin letters was counted as English, and any other type of word was counted as Korean. It is therefore likely, for example, that a few Spanish words were counted as English, and a few others as Korean. Among major K-pop releases, the incidence of words in languages besides Korean and English is quite small, however, so this should not be cause for much concern. Exceptions to this binary classification were the common words "ooh," "oh," "ha," and "haha," all of which, because they are not necessarily English per se, were simply dropped from the word counts.

III. ANALYSIS

For a given song, the number of Youtube video views is given by v, and the fraction of English, ϕ_{E}, is defined as

$$\phi_{\mathrm{E}} = \frac{W_{\mathrm{E}}}{W_{\mathrm{K}} + W_{\mathrm{E}}} \,, \tag{1}$$

where W_{K} and W_{E} are the number of Korean and English words, respectively.

These definitions in hand, the data are analyzed first considering only songs where $\phi_{\mathrm{E}} = 0$ and $\phi_{\mathrm{E}} = 1$. Next, songs are examined with the full range of ϕ_{E}. Finally, a brief argument is made to discount song release date as a possible confounding variable.

A. Pure Korean and Pure English

Of the $N = 1598$ songs available for analysis, $N_{\mathrm{K}} = 104$ have only Korean words, and $N_{\mathrm{E}} = 146$ have only English. A box and whisker plot of the views may be found in Fig. 1. Note the logarithmic scale. Surprisingly, the pure Korean songs appear to outperform the pure English songs with a mean number of views that is 2.5 times larger. This may seem to be an open-and-shut case for the pure Korean songs' dominance, yet there is quite a large range in the data points, so to be sure the difference in the means is statistically significant, a Welch's t-test is performed.

First, while the data themselves are certianly not normally distributed (see Fig.2), according to the central limit theorem, the mean of the data points from newly acquired and independent data sets should be. Given this fact, estimates for the variance of the mean of the pure Korean and pure English video views are $\hat{V}_{\mathrm{K}} = {s_{\mathrm{K}}}^2/N_{\mathrm{K}}$ and $\hat{V}_{\mathrm{E}} = {s_{\mathrm{E}}}^2/N_{\mathrm{E}}$, with corresponding sample standard deviations $s_{\mathrm{K}} = 24\,539\,204$ and $s_{\mathrm{E}} = 18\,738\,316$. If the means of the video views for the pure Korean and pure English songs are $\mu_{v_{\mathrm{K}}}$ and $\mu_{v_{\mathrm{E}}}$, respectively, then unbiased estimators for $\mu_{v_{\mathrm{K}}}$ and $\mu_{v_{\mathrm{E}}}$ are simply the arithmetic means: $\hat{\mu}_{v_{\mathrm{K}}} = \bar{v}_{\mathrm{K}} = 10\,084\,546$, and $\hat{\mu}_{v_{\mathrm{E}}} = \bar{v}_{\mathrm{E}} = 3\,982\,321$.

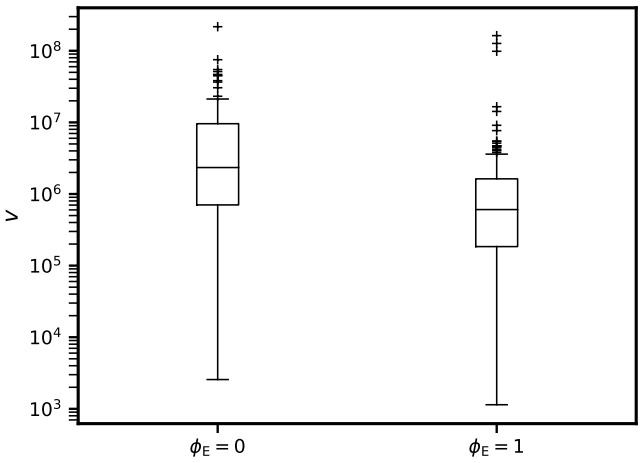

FIG. 1. Box and whisker plot of the views for 104 pure Korean songs (left) and 146 pure English songs (right) on a logarithmic scale. The pure Korean songs have a mean of 10 084 546 and a median of 2 341 422, while the pure English songs have a mean of 3 982 321 and a median of 606 118. The whiskers extend to a maximum of 1.5 times the interquartile range.

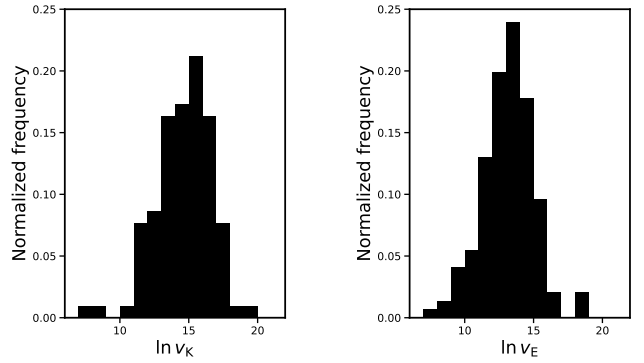

FIG. 2. Normalized histograms of the natural log of the pure Korean songs (left) and the pure English songs (right). It is clear that the data are more easily interpreted and visualized after performing a logarithmic transformation.

The Welch's t-test implies that

$$t = \frac{\bar{v}_\mathrm{K} - \bar{v}_\mathrm{E}}{\sqrt{\hat{V}_\mathrm{K} + \hat{V}_\mathrm{E}}} \tag{2}$$

should be Student-t distributed with

$$d \approx \frac{\left(\hat{V}_\mathrm{K} + \hat{V}_\mathrm{E}\right)^2}{\frac{\hat{V}_\mathrm{K}^2}{d_\mathrm{K}} + \frac{\hat{V}_\mathrm{E}^2}{d_\mathrm{E}}} \tag{3}$$

degrees of freedom, where $d_\mathrm{K} = 104 - 1$ and $d_\mathrm{E} = 146 - 1$ are the degrees of freedom associated with calculating $\hat{V}_\mathrm{K}$ and $\hat{V}_\mathrm{E}$, respectively. A two-tailed test yields a P value of $0.034 < 0.05$, meaning that pure Korean songs do, in fact, outperform pure English songs at a confidence level of 95%.

Having established that the two means are different from one another within an acceptable level of statistical uncertainty, confidence intervals for the means may be worked out. If the mean of the views of N videos is $\bar{v}$, and if s is the sample standard deviation, then a 95% confidence interval is given by

$$\bar{v} \pm \frac{s}{\sqrt{N}} F^{-1}(0.975) \ , \tag{4}$$

where F^{-1} is the inverse cumulative distribution function for the Student's t-distribution with $N - 1$ degrees of freedom. Using the corresponding previously calculated values, it may be shown that the 95% confidence intervals for μ_{v_K} and μ_{v_E} are $(5\times10^6, 15\times10^6)$ and $(1\times10^6, 7\times10^6)$, respectively.

It could be argued that N_K and N_E are not sufficiently large to justify the use of the above method. Therefore, a second method was employed to independently verify the results. A bias-corrected and accelerated bootstrap with 10 000 resamplings was performed using the SciPy method `scipy.stats.bootstrap`, which showed that $\hat{V}_\mathrm{K} = 5.73 \times 10^{12}$ and $\hat{V}_\mathrm{E} = 2.40 \times 10^{12}$. In making use of Eq. (2) and Eq. (3) again, it can be demonstrated that the two means are distinct with a P value of 0.013, and because $0.013 < 0.05$, it can again be concluded with 95% confidence that the pure Korean songs outperform the pure English songs. In addition, the bootstrap returned a confidence interval of $(7 \times 10^6, 18 \times 10^6)$ for μ_{v_K} and $(2 \times 10^6, 9 \times 10^6)$ for μ_{v_E}. This latter pair of confidence intervals is reasonably close to the former set. If it is deemed necessary to put stock in one over the other, the bootstrap method is preferable; though, to be very conservative, the two methods could be combined to yield a 95% confidence interval of $(5 \times 10^6, 18 \times 10^6)$ for μ_{v_K} and a 95% confidence interval of $(1 \times 10^6, 9 \times 10^6)$ for μ_{v_E}.

Both of the above methods are in agreement that K-pop music videos on Youtube with 100% Korean lyrics outperform their 100% English counterparts on average. This result is in direct opposition to the hypothesis that more English in a K-pop song should make it more internationally accessible and should therefore improve its chance of success.

B. Continuous $\phi_\mathbf{E}$

Returning to all $N = 1598$ songs, a fit for a $\ln v$ vs ϕ_E plot was determined using least squares regression with the SciPy method `scipy.optimize.curve_fit`. This fit may be seen in Fig. 3. Regarding the transformation of the dependent variable, it may be true that so long as there are many independent data points, it is reasonable to perform a least squares fit, even if the residuals are not normally distributed [7]. That said, it is preferable to transform the data in such a way that the residuals are normally distributed, or at least roughly so. The

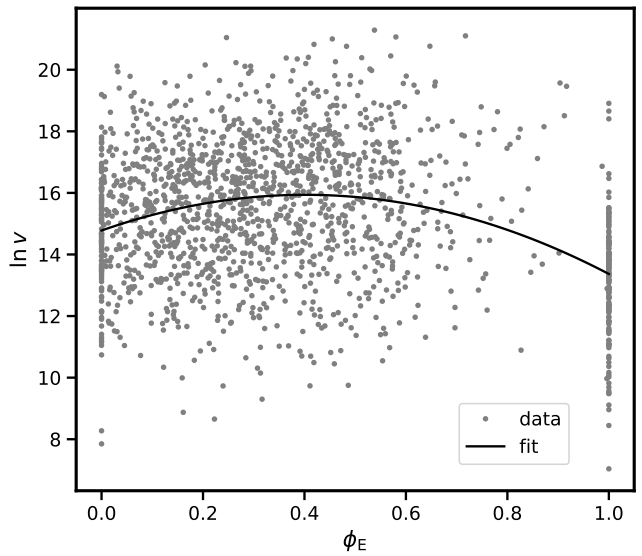

FIG. 3. The natural log of the video views as a function of the fraction of English in the songs. The quadratic fit is given by $\ln v = -7.18{\phi_E}^2 + 5.77\phi_E + 14.8$.

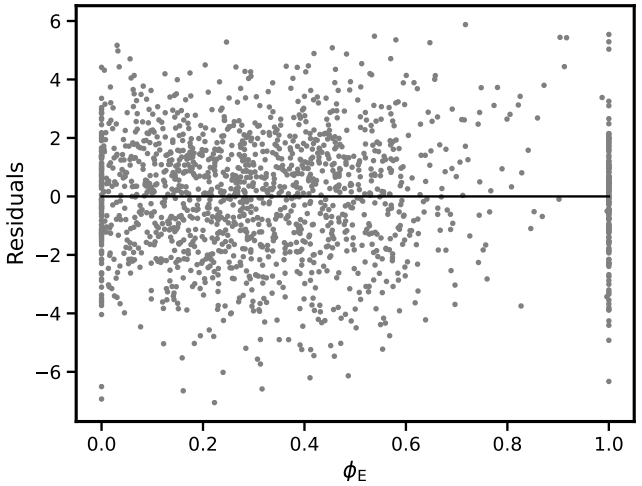

FIG. 4. The residuals for the fit in Fig. 3. Note the fairly even spread about the $y = 0$ line. A histogram of the residuals suggests (roughly) Gaussian spread.

plots in Fig. 2 demonstrate that a logarithmic transformation of the views is a simple way to get approximately Gaussian behaviour, particularly in comparison with the untransformed views. From a different perspective, the log transform is important because it prevents some of the extremely high v values from being overly-domineering in the squares of the least squares fit. As for why a quadratic fit was chosen, a second order polynomial is the lowest-order polynomial that results in a well-behaved residual plot, as seen in Fig. 4.

The fit in Fig. 3 shows that the average views climb for increasing ϕ_E until some optimal point and then fall off, suggesting that while pure Korean song are more popular than pure English songs, a mixture of the two is better still. It is unfortunate that the data are so paltry in the right half of the plot, but the pure English songs do at least provide a reasonably certain endpoint for the fit. The equation describing the fit is

$$\ln v = a{\phi_E}^2 + b\phi_E + c\,, \tag{5}$$

where $a = -7.18$, $b = 5.77$, and $c = 14.8$; and the covariance matrix is given by

$$\begin{bmatrix} V_{aa} & V_{ab} & V_{ac} \\ V_{ba} & V_{bb} & V_{bc} \\ V_{ca} & V_{cb} & V_{cc} \end{bmatrix} = \begin{bmatrix} 0.338 & -0.330 & 0.0487 \\ -0.330 & 0.358 & -0.0605 \\ 0.0487 & -0.0605 & 0.0145 \end{bmatrix}. \tag{6}$$

Next, an estimate and confidence interval for the optimal fraction of English in a K-pop song, ${\phi_E}^*$, is identified. A point estimate is calculated by taking the derivative of Eq. (5) with respect to ϕ_E and setting it equal to zero:

$$\begin{aligned} \frac{d\ln v}{d\phi_E} &= 2a\phi_E + b \\ 0 &= 2a{\phi_E}^* + b \\ {\phi_E}^* &= -\frac{b}{2a}\,. \end{aligned} \tag{7}$$

Substituting in the values for a and b yields ${\phi_E}^* = 0.40$.

As for constructing the related confidence interval, it can be noted that because $\sigma_a = \sqrt{V_{aa}} = 0.581 \ll a$, Eq. (7) is not significantly nonlinear in the relevant parameter space. Therefore, the standard error of ${\phi_E}^*$, $\sigma_{{\phi_E}^*}$, may be found by employing regular error propagation resulting from a Taylor series expansion of the ${\phi_E}^*$ function about the point ${\phi_E}^*(a = -7.18, b = 5.77)$ [7]:

$$\begin{aligned} {\sigma_{{\phi_E}^*}}^2 \approx& \left(\frac{\partial {\phi_E}^*}{\partial a}\right)^2 V_{aa} + \left(\frac{\partial {\phi_E}^*}{\partial b}\right)^2 V_{bb} \\ &+ 2\left(\frac{\partial {\phi_E}^*}{\partial a}\frac{\partial {\phi_E}^*}{\partial b}\right) V_{ab} \\ \sigma_{{\phi_E}^*} \approx& \sqrt{\frac{1}{4a^2}\left(\frac{b^2}{a^2}V_{aa} + V_{bb} - \frac{2b}{a}V_{ab}\right)}\,. \end{aligned} \tag{8}$$

Substituting the optimal parameters and the relevant covariance values into Eq. (8) and then multiplying by 1.96 leads to a 95% confidence interval on ${\phi_E}^*$ of $(0.37, 0.43)$.

C. Confounding Variables

There are a number of possible confounding variables that could impact the interpretation of this analysis. For example, it is conceivable that there might be an assumption within the most successful K-pop production agencies that some minimal amount of English is needed for a song to become successful internationally. Even if this notion ultimately is not true, it would result in songs with the highest production values and best advertising having

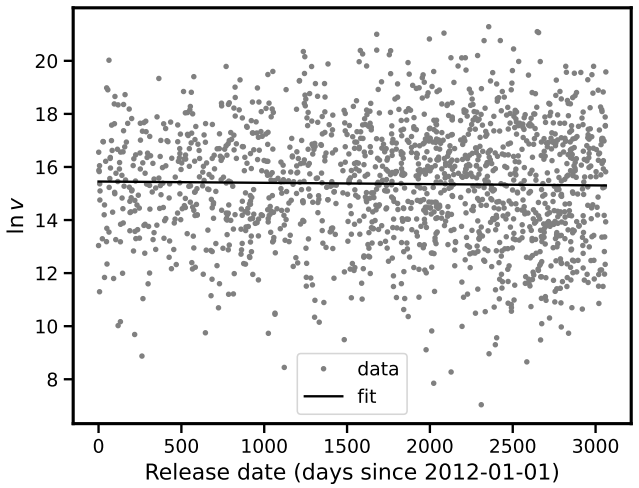

FIG. 5. The natural log of the video views plotted against the song release date. The release date is measured in days since Jan. 1, 2012. The least squares line of best fit has a slope of -4.83×10^{-5}.

"some" English. Of course, such songs should be more successful on average due to their higher budgets, regardless of English content. Another possible confounding variable would be the music genre. If the most popular K-pop subgenres tend to have an intermediate number of English words for cultural and artistic reasons, then song success could, again, be erroneously attributed to the level of English content.

Neither of the above potential confounding variables were examined in this analysis, though they might be considered in future work. One possible confounding variable that was dealt with was song release date. It is easy to imagine, say, that as time has gone on, K-pop videos have become more popular, while simultaneously the amount of English in the songs has increased. Surprisingly, however, the views are quite independent of release date, as seen in Fig. 5, where the natural log of the views is plotted as a function of release date. The Pearson correlation coefficient comes out to -0.0185, and the line of best fit has a slightly negative but extremely shallow slope of -4.83×10^{-5}. Apparently, the effect of the older music videos having had more time to garner views is well balanced by the increased number of releases and increased popularity of many of the more recent songs, at least over the range of dates considered. Release date can thus be discounted as a potential confounding variable.

IV. CONCLUSION

It was determined that K-pop songs with only Korean lyrics fare better on average than K-pop songs with only English lyrics at a 95% confidence level. A 95% confidence interval for the mean number of views of the pure Korean songs was found to be $(5 \times 10^6, 18 \times 10^6)$, while the equivalent confidence interval for the pure English songs was found to be $(1 \times 10^6, 9 \times 10^6)$. In light of this fact, it is surprising that so many songs are produced with entirely English lyrics. This might be a result of producers trying to target particular niches.

Despite the apparent favouring of Korean lyrics, it turned out that songs with an appropriate mix of Korean and English performed better still. A 95% confidence interval on the fraction of English words in K-pop songs (assuming English and Korean words only) that maximizes average views was calculated to be $(0.37, 0.43)$.

It is possible that the reason success seems to drop off for relatively high English content might be that K-pop is perceived to lose something intrinsically valuable when an insufficient number of words are in Korean, even in the ears of international audiences. Alternatively, it could be that South Korea still drives a commanding level of engagement with K-pop, so songs with roughly 40% English words strike an acceptable balance between being sufficiently understandable for the average Korean, while also giving international audiences something to latch onto. Future work might include splitting the views by country to see if this might be the case.

[1] Yonhap, "Hybe tops W1tr in annual sales, first in K-pop industry." http://www.koreaherald.com/view.php?ud=20220222000965, 2022.

[2] P. Sajnach, "The Korean wave: From PSY to BTS -the impact of K-pop on the South Korean economy." https://www.asiascot.com/news/2021/01/22/the-korean-wave-from-psy-to-bts-the-impact-of-k-pop-on-the-south-korean-economy/, 2021.

[3] M. Stassen, "Despite the pandemic, Big Hit Entertainment generated revenues of $436 in the first 9 months of 2020." https://www.musicbusinessworldwide.com/despite-the-pandemic-big-hit-entertainment-generated-revenues-of-436m-in-the-first-9-months-of-2020/, 2020.

[4] M. Takahashi and D. Calica, "The significance of English in Japanese popular music: English as a means of message, play, and character," *Proceedings of the 21st Annual Meeting of the Association for Natural Language Processing*, 2015. https://www.anlp.jp/proceedings/annual_meeting/2015/pdf_dir/D6-4.pdf.

[5] Datartist, "K-pop database (1992-2020): K-pop database from dbkpop.com." https://www.kaggle.com/kimjihoo/kpopdb/metadatal, 2020.

[6] K. Anderson, *Essentials of Linguistics*. McMaster University, 2018. https://ecampusontario.pressbooks.pub/essentialsoflinguistics/.

[7] G. Cowan, *Statistical Data Analysis*. Great Clarendon Street, Oxford, UK: Oxford Science Publications, 2002.