

# Language Technology and Cultural Contents - Final Project

*Use of NLP for analysing French, Spanish and Korean Music*

Fall Semester 2024

Project by Alexis Dhermy and Pablo Picó Salort

Professor: Youngeun Koo

## I. Introduction

The following research project presents an approach to the current music industry and trends through the use of Natural Language Processing (NLP). Through an analysis of song lyrics trending in France, Spain and Korea, the aim is to provide an analysis on what makes songs successful in different countries, as well as to bridge the intercultural gap between them through a global phenomena as is music.

The motivation to start this project comes from the personal experience of the authors as natives from France and Spain, but studying abroad in Korea as exchange students. Given the limited time and language skills they possessed, they encountered some challenges in communicating with their Korean classmates, but eventually realised they all shared an enthusiasm for music. Thus, they decided to create a project that facilitated this communication as well as allowed them to share their passion with their classmates. Moreover, the project evolved to serve as support to understand the music industry better and what makes songs successful. One of its applications might be as aid for aspiring musicians that might struggle with the language barrier but are interested in the music of other countries.

Therefore, the following objectives were formulated for the project:

1. Providing a deeper understanding of the popular music in three different countries using their linguistic characteristics.
2. Examining the differences and similarities between countries through the lens of music.
3. Applying NLP tools to provide a structured and accurate analysis.
4. Facilitating cross-cultural communication and understanding.

## II. Methodology

### 1. Data Collection

For an analysis of the current music, it was decided that the currently trending music would serve as a representative sample. While it might not serve as representative for the entire industry, the study doesn't aim to provide a detailed analysis on the industry as a whole, but rather of the comparison between countries, which can be done in a fixed moment in time.

For the selection of data, the top 20 songs of each country in a set date (December 6th) were selected from the Spotify list. As the most popular music streaming service in the three countries, the Spotify trending page is an accurate reflection of the most popular music among the population of a country, so it was determined as the indicator for the song selection.

The names, authors and lyrics of the songs were manually crawled from the Genius website and indexed in a text file for each one of the countries. It was decided that all songs should be in the same language to facilitate an easier comparison than and not have to deal with the language connotations. The language selected for this was English, given that it is not native to any of the three countries (thus reducing bias) and given its status as a global language for music. Translations for all songs were gathered either from the Genius website or created manually by using the DeepL software. All DeepL translations were checked by the authors to correct any minor errors.

## 2. Data Cleaning

To ensure the quality and consistency of the analysis, a thorough data cleaning process was conducted on the song lyrics:

- **Text Normalization:** All lyrics were converted to lowercase to standardize the text. Structural metadata, such as text between brackets (e.g., [Chorus], [Verse 1]), was removed, as it does not contribute to semantic analysis. Non-alphabetical characters were excluded, except for common punctuation marks (., ,, !, ?) to retain meaningful linguistic elements.
- **Tokenization:** Using NLTK's tokenizer, lyrics were split into individual tokens, transforming each song into a sequence of words for further analysis.
- **POS Tagging and Filtering:** Part-of-speech (POS) tagging identifies the grammatical roles of tokens. Only nouns, adjectives, and verbs were retained, as these are the most semantically informative word categories for analyzing similarity.
- **Lemmatization:** Words were reduced to their base forms (e.g., "running" → "run") to group variations of the same word and enhance consistency in analysis.
- **Stop Word Removal:** Common stop words (e.g., "the," "and") were removed using NLTK's stopword list. A supplementary list of stop words sourced from CountWordsFree.com was also included to address frequently used words in English. And an additional custom list of repetitive and non-informative words specific to song lyrics (e.g., "uh," "yeah," "ahaha") was created to further reduce noise.
- **Limiting Repetition:** Duplicates of words were capped at three occurrences per song to address repetitive elements, particularly in choruses, while preserving lyrical meaning.

This cleaning process significantly reduced noise in the dataset and prepared the lyrics for further semantic and sentiment analysis.

### 3. Data analysis

Having done the data preprocessing, the proper analysis was carried out. The tools used were the following:

- **Word frequency across countries (nouns, verbs, adjectives).** This was done using the counter tool imported from the collections library, given its higher accuracy when dealing with lists. By extracting the word frequency, a comparison of the most used words in each country is possible, allowing for an analysis of the general line of each one of the countries and a comparison between them. For an easier visualisation, the words were displayed in word clouds using the matplotlib and the wordcloud libraries.
- **Common words across countries.** Building on the word frequency and by transforming the data into sets, all the words in the lyrics were compared to extract the ones that overlap. This was done in order to get a general idea of what topics might be shared based mainly on substantives, adjectives and verbs.
- **Sentiment analysis.** A sentiment analysis was also carried out in order to figure out if there is a predominant sentiment in each country (positive, negative or neutral) and to assess the overall tone of each language, so that a comparison between them could be established.
- **Semantic similarity across countries** was also carried out using SpaCy and BERT. Given the information that semantics can give about speakers and context, this was done to complement a purely semantic approach and also focus on how the sentences are constructed, their complexity and interrelation.
- A **Recommendations Song Service** is the final output of this project. It is specifically designed for Korean users. Starting with the top 20 Korean songs, users can select one song for which they would like to receive recommendations. The system then provides the most similar songs (based on a combination of sentiment analysis and semantic analysis) from the French and Spanish playlists.

## III. Results

### 1. Word frequency across countries



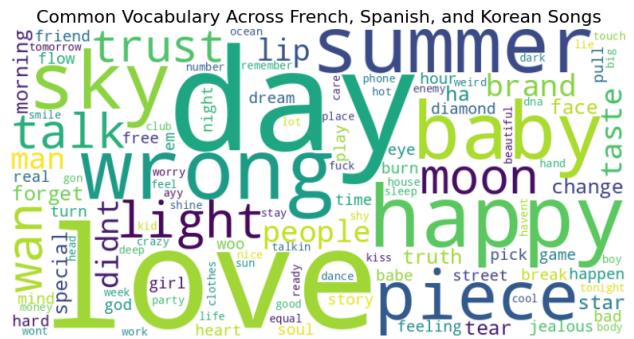


*Figure 1. Top words in French, Korean and Spanish*

As we can see, the predominant theme in all three countries is love, a word that is featured in the three countries. However, whereas Korea has a more platonic approach to love (signified by words such as "heart", "feel" and perhaps "hear"), France and Spain shared a more physical view of love (highlighted by words such as "bed", "kiss" or "fuck" in their respective word bubbles). Baby is particularly prominent in France and Spain (while also being featured in Korea), which taken the affective connotation of the word hints once again towards the predominance of love as a theme.

## 2. Common words across countries

For the shared word analysis, besides certain common-use words being repeated (verbs such as "won't", "haven't", "talk" and nouns such as "house" or "heart"), the majority of the words used correspond with the ones extracted before in the most common word section. This further emphasises how love is a central theme in the music shared between the three cultures and languages.



*Figure 2. Common Vocabulary Across French, Spanish, and Korean Songs*

### 3. Sentiment analysis

Overall Sentiment for Korean Songs: POSITIVE, 0.55  
Overall Sentiment for French Songs: NEUTRAL, -0.03  
Overall Sentiment for Spanish Songs: NEUTRAL, -0.19

*Figure 3. Overall Sentiment (Average sentiment of each song by playlist)*

The sentiment analysis among cultures proved to be rather unexpected. Despite the three languages apparently sharing similar themes as hinted by the previous analysis, the tone in which they do so is radically different. Whereas Korea maintains a strong positive tone, Spain seems to be driven more towards negative undertones, whereas France remains close to neutral. Applying it to the themes explored before, Korea sings to a vibrant and hopeful love, whereas Spain seems to center more on heartbreak and negative emotions. This could also be true of other topics covered in songs.

#### 4. Semantic similarity across countries

Top 10 Similar Songs (Korea vs France):					
Korean Songs	Artist Korean	French Songs	Artist Foreign	Playlist	similarity
3D	Jungkook ft. Jack Harlow	Gata Only	FlooyMenor & Cris Mj	French Songs	0.608439
Luther	Kendrick Lamar ft. Sza	Die With A Smile	Lady Gaga & Bruno Mars	French Songs	0.603856
Number one girl	Rosé	Die With A Smile	Lady Gaga & Bruno Mars	French Songs	0.601615
number one girl	ROSE	Die With A Smile	Lady Gaga & Bruno Mars	French Songs	0.601615
Be Mine	Jimin	Gata Only	FlooyMenor & Cris Mj	French Songs	0.598081
HOME SWEET HOME	G-Dragon	CRF	Yorssy	French Songs	0.590143
Luther	Kendrick Lamar ft. Sza	CRF	Yorssy	French Songs	0.589266
Running Wild	Jin	Die With A Smile	Lady Gaga & Bruno Mars	French Songs	0.575067
Number one girl	Rosé	CARTIER SANTOS	SDM	French Songs	0.574877
number one girl	ROSE	CARTIER SANTOS	SDM	French Songs	0.574877

Top 10 Similar Songs (Korea vs Spain):					
Korean Songs	Artist Korean	Spanish Songs	Artist Foreign	Playlist	similarity
Number one girl	Rosé	Qué pasaría	Rauw Alejandro ft. Bad Bunny	Spanish Songs	0.595968
number one girl	ROSE	Qué pasaría	Rauw Alejandro ft. Bad Bunny	Spanish Songs	0.595968
Who	Jimin	Se Fue	Rauw Alejandro ft. Laura Pausini	Spanish Songs	0.594376
Number one girl	Rosé	Los Días Contados	Quevedo ft Rels B	Spanish Songs	0.577198
number one girl	ROSE	Los Días Contados	Quevedo ft Rels B	Spanish Songs	0.577198
Smeraldo Garden Marching Band	Jimin ft. Loco	Ohnana	Kapo	Spanish Songs	0.567297
Slow Dance	Jimin ft. Sofia Carson	Ohnana	Kapo	Spanish Songs	0.567114
Who	Jimin	Shibatto	Quevedo	Spanish Songs	0.560775
3D	Jungkook ft. Jack Harlow	Duro de verdad pt. 2	Los sufridos ft. Bad Gyal	Spanish Songs	0.559588
Smeraldo Garden Marching Band	Jimin ft. Loco	Si antes te hubiera conocido	Karol G	Spanish Songs	0.556942

Figure 4. Top 10 Similar Songs between Korea and France, and Korea and Spain.

For the semantic analysis, we chose BERT-based models over SpaCy due to BERT's superior ability to capture contextual meaning. We used 'all-MiniLM-L6-v2' for its balance between accuracy and efficiency, as 'all-mpnet-base-v2' was too slow in terms of computation.

The comparison between songs showed a maximum of 0.6 similarity between Korean and French, and 0.59 for Korean and Spanish, with the other most similar songs adopting slightly lower values. While the Korean song that appears the most is "Number One Girl", whose lyrics are in English and that could be labeled as more influenced by Western music, we also find a reasonable parallel between Western songs and the more Korean ones.

We also experimented with SpaCy's model, but the results were unexpected, with all song similarities exceeding 0.97. This high similarity seemed too good to be true, indicating that the model might not have been sensitive enough to the nuanced differences between the lyrics of the songs.

#### 5. Recommendations Song Service

As mentioned earlier, the ultimate goal of this project is to create a song recommendation service. By combining all the results from previous analyses (particularly sentiment and semantic analysis), we offer Korean users a service that recommends French and Spanish songs.

How it works: The user selects a song from the Korean playlist for which they would like recommendations. The system then calculates the three most sentimentally and semantically similar songs and presents them to the user.

-----RECOMMENDATIONS SONG SERVICE-----

안녕하세요! Hello! ¡Hola! Bonjour!

Welcome to our Recommendations Songs Services!

Here is the list of top songs in Korea:

1- Who, Jimin  
 2- HOME SWEET HOME, G-Dragon  
 3- Seven, Jung Kook  
 4- Running Wild, Jin  
 5- APT., ROSE & Bruno Mars  
 6- Be Mine, Jimin  
 7- Whiplash, aespa  
 8- Winter Ahead, V & Park Hyo Shin  
 9- I'll Be There, Jin  
 10- POWER, G-Dragon  
 11- Smeraldo Garden Marching Band, Jimin ft.Loco  
 12- 3D, Jungkook ft. Jack Harlow  
 13- Number one girl, Rosé  
 14- Standing next to you, Jungkook  
 15- Slow Dance, Jimin ft. Sofia Carson  
 16- Luther, Kendrick Lamar ft. Sza  
 17- Skrrrr, Haon ft. Giselle  
 18- Cherish (My love), ILLIT  
 19- toxic till the end, ROSE  
 20- White Christmas, Bing Crosby  
 21- number one girl, ROSE

Enter the number of the song you want recommendations for: 20

Top 3 recommendations for White Christmas by Bing Crosby:  
 1- Last Christmas by Wham! from French Songs  
 2- Die With A Smile by Lady Gaga & Bruno Mars from French Songs  
 3- Ohnana by Kapo from Spanish Songs

Enter 'quit' if you want to stop the process  
 Enter 'quit' to exit or press Enter to continue: quit

Thank you for using our service!

Good Bye!  
 안녕히 가세요!  
 ¡Adiós!  
 Au revoir!

Figure 4. Recommendations Song Service

#### IV. Conclusions

In conclusion, through the NLP analysis of the top songs in Korean, Spanish and French, we are able to glimpse into the world of contemporary music through the lens of linguistics. This analysis is not only interesting for understanding each song individually, but also at a comparative level, seeing how songs in the same language share similar characteristics, how the topics for popular music often overlap no matter the language the songs are written in, and how we can deduct aspects related to themes and song creation.

Our final service offers a unique opportunity to discover new music and observe how trends can align across different cultures. It highlights how similar themes can be explored in songs while maintaining distinct rhythmic and stylistic differences.

The applications of NLP in the field of music, besides the song recommendation system discussed in the essay, are many and yet to explore. As mentioned at the beginning, work such as the one done in this project can be revised and expanded in order to facilitate the creative process for artists, as well as for letting listeners understand better music that's not in a language they speak nor might have been considered by them before.

#### V. Limitations

- Language-Specific Challenges:** The analysis was primarily conducted using English-centric NLP tools, which might miss significant nuances inherent to Spanish, Korean, and French. Translating all lyrics into English may lead to the loss of important meanings and cultural context. A more accurate approach would involve analyzing each song in its original language to preserve the unique syntactic and semantic characteristics specific to each language.

- **Limited Dataset:** The dataset consisted of only the top 20 songs per language, which may not fully capture the diversity of the broader musical landscape. By increasing the number of songs per playlist, our system could become more accurate and reveal new patterns, leading to more comprehensive and insightful recommendations.

## VI. Future Work

- Increase the number of songs and genres to enhance the diversity and accuracy of the recommendation system.
- Utilize specialized NLP tools for each language to improve analysis precision. For instance, employing KoNLPy for Korean songs can provide more accurate linguistic insights.
- Deepen the exploration of sentiment analysis to better understand the emotional nuances of lyrics.
- Integrate tempo analysis, which is a key element in music analysis.

**Thank you!** 감사합니다!

### Contact

Alexis Dhermy : [a.dhermy@live.com](mailto:a.dhermy@live.com)  
Pablo Picó Salort : [ppico.ieu2021@student.ie.edu](mailto:ppico.ieu2021@student.ie.edu)