EXPLORING THE RECEPTION OF FOREIGN LANGUAGE MUSIC IN THE ENGLISH-SPEAKING WORLD WITH THE USE OF DIGITAL HUMANITIES TECHNIQUES

By

CHARLIE ARMSTEAD

A thesis submitted to the University of Birmingham for the degree of

MASTER OF ARTS BY RESEARCH

Department of Modern Languages

College of Arts and Law

University of Birmingham

May 2022

Abstract

This thesis presents an analysis of the reception of popular foreign language music in the English-speaking world with the use of a corpus of YouTube comments. This analysis is carried out with the use of a variety of computational methodologies and other Digital Humanities techniques. We initially situate ourselves within the wider authorship of scholars employing similar corpora to us and also provide a framework for the collection and curation of a corpus of YouTube comments. Then, in the pursuit of our analysis, we narrow our focus to three case studies: an investigation into the role of fandom amongst English-speaking K-Pop fans, an examination of language-attitudes in the comment sections of reggaeton songs, and finally an analysis of the role of the meme in viral songs alongside a novel methodology to this end.

Throughout these case studies, methodologies of varying complexity are employed and encompass fields of computer science ranging from machine learning, computational linguistics, and computational discourse analysis. With this in mind, in order to comply with Digital Humanities ethics and values, a secondary aim of this project is the careful documentation and justification of methodological techniques employed throughout. This is to ensure maximal reproducibility for future research and most of the scripts and datasets produced during this project can be found at https://ursidaeic.github.io/.

In each of the case studies, we succeed in elucidating aspects of the digital culture of YouTube with regards to the respective focus of the study. However, we conclude that, as this is such an under-studied area of digital culture, there still remains much scope for future research which will be greatly facilitated with the methodologies and datasets produced throughout this project.

Acknowledgements

First and foremost, I would like to thank my supervisors Dr Caroline Ardrey and Professor Helen Abbott. Not only have they provided me with invaluable academic guidance and critique throughout the project, but without their constant support and advocation this project would have been completed in double the time and at half the quality. In Dr Ardrey's case, this support and guidance extend back to my undergraduate degree during which her Digital Humanities module ignited my passion for the field and who ultimately supported me with my successful applications for the Woltmann scholarship and the Master's course itself.

I also would like to thank all those computer scientists close to me who I pestered with questions throughout my development as a programmer. Chief amongst them is Elliot, who always found time to guide me through any difficulties I was having, no matter how basic a problem. The technical skills I possess at the end of this project are in no small part due to his kindness and patience, and for that I give him a special thanks.

I extend my thanks to the Paul and Henry Woltmann Bequest for awarding me a scholarship.

Finally, my love and thanks to Emily. Your daily encouragement, love, and support has been an inspiration and I would not have been able to do this without you.

Contents listing

Contents

List of tables

List of figures

List of abbreviations

United States of America = USA

United Kingdom = UK

Support Vector Machine = SVM

Part Of Speech = POS

Hidden Markov Model = HMM

Introduction

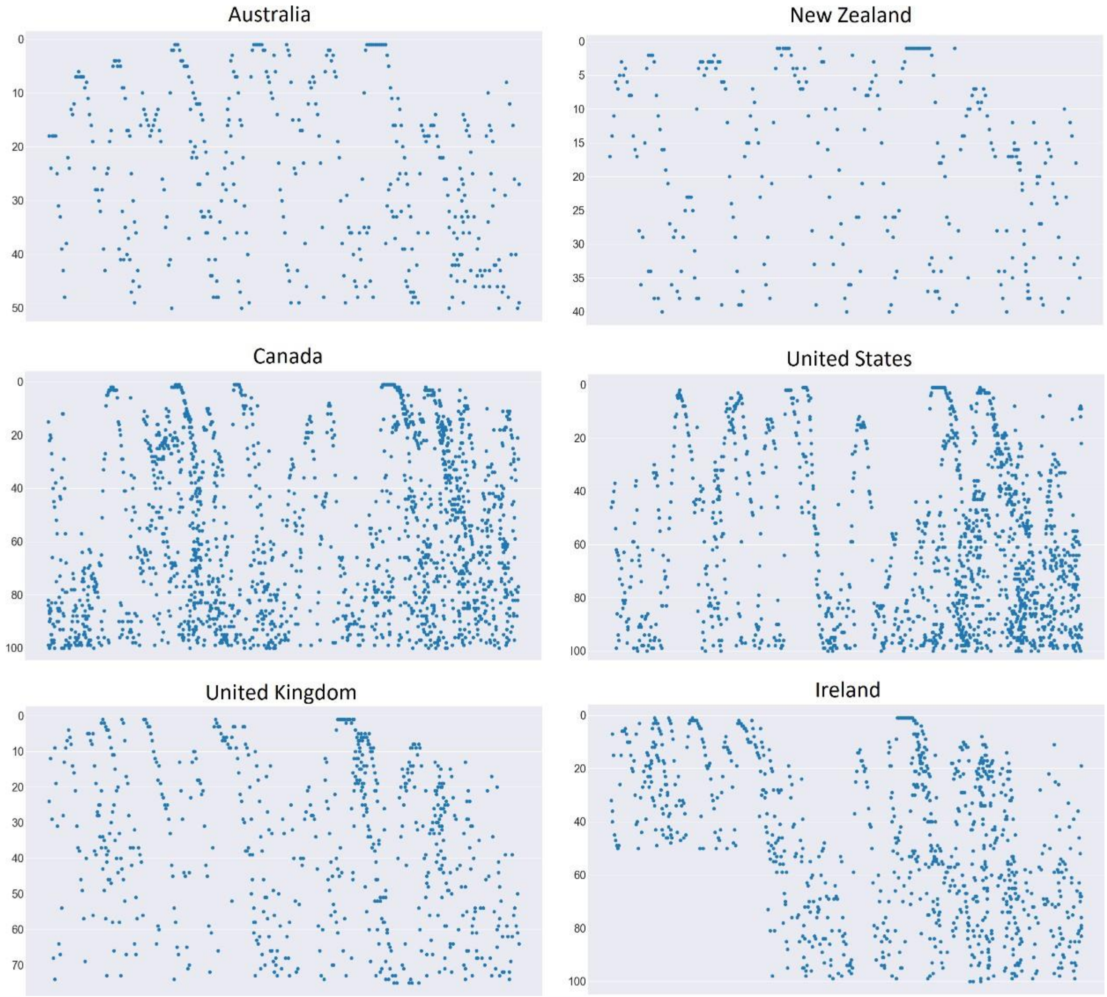


Figure i: Placement of foreign language songs in the charts of 6 English-speaking countries from week-to-week.

In 2017, 'Despacito' by Luis Fonsi and featuring Daddy Yankee exploded onto the international music scene, becoming the most-viewed video on YouTube, garnering billions of listens across several platforms, and receiving multiple accolades at the Latin Grammy Awards. Raisa Bruner of *Time* magazine remarked that "in a year where xenophobia reared its head worldwide", the fact that such a universal and multicultural hit topped the charts "inspires hope" (2017). In light of this song's success and the subsequent "Latin Renaissance" in the following three years (Acevedo, 2019), Joshi claims that this song heralded a new age of international music which is eroding the grip of English-language music on the worldwide music industry (2020). In further evidence of this, O'Connor points to the fact that, prior to the release of 'Despacito', foreign language No. 1s tended to be so-called 'one-hit wonders' (2017). That is not to say that foreign language music has never had a presence in the charts of English-speaking countries. Figure *i* displays the distribution of the weekly charting positions for foreign language songs in English-speaking countries from October 2007 to November 2020. The countries are Australia, New Zealand, Canada, The United States of America (hereafter referred to as the USA), the United Kingdom (hereafter the UK), and Ireland– left to right, top to bottom. Cursory visual analysis makes it clear that foreign language music has begun taking more of a foothold in the English-speaking world, most prominently in the USA. If, as Bruner claims, xenophobia across the word is soaring, then why are we seeing an increase in multiculturality in popular music, a medium that allegedly mirrors the flows of global culture in a particularly noticeable fashion (Marc, 2015; 2017)?

One undeniable cause, and the ultimate focus of this research project, is the advent and subsequent soaring popularity of music streaming services. In an interview with the *Independent*, Enrique Iglesias remarked how "contrary to the past, music today has no boundaries" as a result of the growth of music streaming services (O'Connor, 2017). According to the IFPI's 2019 report titled "Music listening in 2019", 89% of the global population made use of streaming services that year (2019). In his analysis of the report, Ingham highlighted that the music streaming platform that services the largest portion of the population is YouTube, despite the platform's primary focus being video sharing and viewing (2019). He adds that, in spite of Spotify being commonly cited as the market leader in music streaming, it attracts just 12.5% of YouTube's 1.5 billion monthly listeners. Thus, it follows that YouTube must be a principal object of study when tracing the flows of popular music in the age of streaming services.

As a result of the high levels of engagement and traffic on the platform, the nature of the project is one that will necessarily involve grappling with large datasets and statistical analyses. However, we will be doing so in order to examine cultural artifacts, i.e., foreign language songs. Consequently, the field of Digital Humanities and the tools and techniques associated with it are the obvious choice when designing a methodology to this end. The broad structure of the project is as follows: we will first use computational techniques such as web-scraping in order to identify the foreign language songs that have made an appearance in the charts of several English-speaking countries. Then, we will download the maximum number of comments from the YouTube videos associated with these songs and filter them for opinion with the use of machine learning techniques. This is to gain insight as far as possible into what English-speakers are saying about the songs and also why they are saying it. The data that is gathered from this endeavour will be used to support three case studies of separate focuses. The first in chapter 3 will examine the difference in reception between BLACKPINK and BTS, two K-pop groups. The reason for this is that the genre of K-pop presents unique modes of fan engagement and interaction within the broader spectrum of responses to foreign language music. The second case study will examine the role of language attitudes in opinions towards 'Despacito' and other reggaeton songs. The third and final study will be an exploration into the role of the meme in the reception of viral foreign language songs; this chapter will propose a novel methodology aimed towards the identification of these memes using syntactic features.

Whilst these chapters will have their own specific aims, through investigating these different aspects of foreign language songs we will also gain insights into the digital culture of YouTube as a platform. The large datasets, in-depth and varied methodologies, and multilingual space in which discussions play out provide grounding upon which to draw conclusions that, up until now, have not been in reach for Modern Language studies and more broadly the Humanities.

Chapter 1: Literature review

1) Why build a corpus of YouTube comments?

If, in 2004, Chad Hurley, Steve Chen, and Jawed Karim had been told that, in just under two years, their idea for a video-based online dating site would sell for $1.65 billion, a few eyebrows would have been raised. Of course, by the time of its acquisition by Google in November of 2006, YouTube had evolved somewhat from its original conception to a more general video sharing platform (Nieva, 2016). Despite the nearly $2bn price tag, it is clear that Google spotted early on the enormous potential that YouTube represented and it remains one of their most profitable purchases to date (Lipton, 2014). In addition to the immense monetary value of the platform, many scholars have noted YouTube's value as a space of unique socio-linguistic interaction. Official music videos now form a key part of YouTube's vast video offering; this, coupled with its global and multilingual reach render the platform particularly well suited as the object of the present study.

Some of the first scholars to recognise the value of this space were Jones and Schieffelin (2009). They propose that the anonymous commenting feature on YouTube facilitates the formation of an egalitarian nexus of metalinguistic views. In explanation, Jones and Schieffelin liken YouTube comments to bathroom graffiti, whereby a platform is made available to a wide array of voices whose value is not determined by the author, but the content of the message. They suggest that, unlike bathroom graffiti, however, one can find instances of back-and-forth dialogue within YouTube comments, although often this does not extend beyond more than a few replies, which ultimately leads to a lot of repeated opinions and questions. From a methodological standpoint, Jones and Schieffelin's approach is not especially relevant to the that of this project beyond the general approach of content analysis of the comments themselves. This is because Jones and Schieffelin deal with a dataset that numbers in the hundreds to the thousands and so could conceivably be parsed and analysed by hand. This project's corpus size numbers in the millions and thus requires digital techniques that were unnecessary for Jones and Schieffelin.

In spite of the relatively straightforward methodological techniques, this work by Jones and Schieffelin, as Coleman points out, remains a “rich microanalysis” of “user-generated metalinguistic data” and demonstrated to the wider academic community the potential in analysing interpersonal interactions within this unique communicative space (Coleman, 2010, pg. 494). However, in 2012, Bou-Franch et al. contended that there still remained a dearth of research examining the “YouTube text facility”, or in other words the comment section (pg. 502). The overarching goal of these authors’ work was to fill this perceived gap by launching an investigation into the ways in which Spanish-speaking users on the website strive to make sense of other contributions. Much like Jones and Schieffelin, Bou-Franch et al. laud the YouTube comment feature as a unique form of interpersonal communication (2012). However, they define the interactions as “polylogal”, meaning between 3 or more people (Dobs and Garcés-Conejos Blitvich, 2013), whereas Jones and Schieffelin claim that it is “inherently dialogic” (pg. 503; 2009, pg. 1075). Bou-Franch et al. support their claim by describing how YouTube comments encompass both one-to-many interactions and also intergroup discussions. Departing further from Jones and Schieffelin, these authors also employ a more complex methodological approach and specifically outline the challenges that arise when attempting to analyse YouTube comments. The principal challenge that they encounter is the fact that “all individuals involved in a polylogal interaction” are considered participants “regardless of whether they are message senders or readers” (pg. 504). Consequently, Bou-Franch et al. propose that it is not possible to actively capture the full scope of interaction on the platform as there still remains the “imagined ‘mass’ of ordinary users” who do not leave comments on videos, yet passively participate in the interaction and whose reactions are unknowable (Burgess and Green, 2008, p. 8., in *Ibid.,* 2012). A potential solution to this problem would be to make use of the ‘like’ feature. At the time of publication of the above article, research shows that YouTube had implemented a ‘like-counter’ for comments, but either it was not regularly used or (more likely) it gave comments a score based on the number of likes minus the number of dislikes, the result being that the top comments on videos that garnered millions of views would only display a very low number of likes (Wayback Machine, 2020). Consequently, using the number of likes that a comment gained was not a tenable approach in 2012, but at present comments on highly viewed videos are showing an expected number of votes. Thus, this could be used as a metric to try and gauge the general reaction of these invisible masses.[1] Through a series of quantitative techniques, Bou-Franch et al. conclude that "communication over the YouTube text-commenting facility was far from incoherent" and that "the postings of YouTube polylogues are sufficiently connected so as to constitute a space for online interaction rather than a series of disconnected comments" (pg. 515).

Whilst Bou-Franch et al. do refer to the multilingual nature of the platform, it is not a primary focus in their research. Despite this, it is self-evident that a platform that both lends itself to polylogal interactions and also boasts an international userbase would be a locus for attitudes regarding language. Thus, in 2013, Dejan Ivković carried out a specific analysis of language attitudes from a corpus comprised of YouTube comments. Citing Walters and Chun (2011), Ivković describes how YouTube presents an ideal space for contesting dominant linguistic ideologies as users from varied and relevant linguistic backgrounds who have a strong interest in song performance can present their own opinions on specific language-related issues. Ivković elects to conduct a corpus-based approach to discourse analysis so as to make use of computational tools. As justification, Ivković draws on work by Upton and Ann Cohen, who state that computers allow "much more interesting and comprehensive linguistic analyses" in comparison to those done by hand (2009, pg. 601). Ivković does touch on the drawbacks of a corpus-based analysis, remarking that text-based data excludes visual and non-linguistic markers that can provide secondary data. However, this claim fails to take account of the use of non-linguistic markers used in this medium - notably, emojis and emoticons. Emojis and emoticons have long been recognised as a way of 'making up' for non-linguistic cues in a text-based communicative environment (Schneebeli, 2017; Gawne and McCulloch, 2019). Schneebeli

[1] Note: during the course of this project, YouTube removed the dislike counter from videos. This would naturally impact on our ability to employ this metric for examining the responses of the 'invisible masses', however not completely exclude it as a possibility.

notes that, particularly in YouTube comments, emojis and emoticons are used as more than just an affective expression; they can provide an argumentative function, or reaction, to a given video or comment. As with Jones and Schieffelin and Bou-Franch et al., Ivković has a relatively small sample size in comparison to the one that will be collected by this project; his corpus was comprised of 769,846 individual tokens (words) (2013), whereas this project will collect over 100 times that many full comment strings. In spite of this, the techniques that he employs still provide a useful methodological basis on which to build and will thus be discussed in chapter 4. The conclusion that Ivković draws is that corpus-based analysis can indeed be used to study language attitudes on social media. With regards to their specific findings, he discovers that, within the context of the Eurovision Song Contest on YouTube (which is a unique nexus of contesting language attitudes in and of itself), online commenters tend to favour performances sung in the native language of the singer. Conversely, songs in which a singer employs several languages/accents attract controversy. This would offer an explanation as to why foreign language songs, at the very least, are not rejected outright by the English-speaking world. However, given that the Eurovision Song Contest is, as previously mentioned, a unique "locus of intensive language contact and language contestation", it is necessary to conduct further, broader research in order to verify that these results hold true in other contexts (2013).

To date, there have been very few investigations into language attitudes on YouTube beyond that of Ivković and none that employ computational methods in order to draw conclusions. Thus, this is one gap that this project will aim to fill.

2) Challenges posed by streaming services.

Streaming services have revolutionised the way in which the public consumes music (Mangold and Faulds, 2009). The impact that these services have had on the industry is both profound and wide-reaching. It ranges from altering the specific musical structure of songs released in a world dominated by streaming (McCormack, 2020) to potentially opening up mainstream popularity of songs to manipulation by proprietary, corporate algorithms (Goldschmitt and Seaver, 2019). Consequently, authors have been quick to point out the inherent irony behind the nomenclature of digital media consumption over the last decade; whilst vocabulary such as ‘stream’ or ‘cloud’ imply the potential for a freedom or weightlessness to music consumption, the truth is that these services are still very much entrenched in the gravities of the music economies that exist today (McCourt and Zuberi, 2016; Morris and Powers, 2015). Since their genesis, streaming services have relied on algorithmic recommendation in order to “contour the music flows they provide” (Goldschmitt and Seaver, 2019, pg. 66). As such, it is necessary to explore the issues and ramifications associated with these tools with regards to music consumption in the 21st century and consequently within the framework of this project.

Morris and Powers remark that, whilst the advent of streaming services might ostensibly have realised Goldstein’s concept of a “celestial jukebox” – a device through which it is possible to access any song available instantly —if we look a little deeper, this is far from being the case (2015, pg. 106). In reality, as consumers we are limited to a handful of commercial platforms through which we can access one of many ‘taps’ into the conceptual stream of music, each of which have their own styles and limitations. Morris and Powers propose that the reason for this is the need for brands to distinguish themselves from competitors in an increasingly crowded market. This impetus leads to streaming services offering “branded musical experiences” that invite consumers to see their identities and habits reflected in their choice of service (pg. 107). Morris and Powers identify a number of ways in which these services differentiate themselves, a key one being the particular qualitative features that they offer over their competition. For example, Spotify, who as of last year lay claim to 37% of the worldwide streaming market (Iqbal, 2020), promote how their service will algorithmically tailor the listening experience to the consumer. However, the authors demonstrate that Spotify has started using their data analytics, detailed user profiles, and complex algorithms to increase the exposure of specific Spotify partners on the platform. Furthermore, these promotional messages are often presented not as advertisements but as grassroots discoveries tailored to a given user based on their listening habits.

Eric Drott examines the impact of such recommendation algorithms from a psychoanalytic standpoint (2018). Drott agrees with Morris and Powers' argument that the "musical cornucopia" of streaming has been repudiated by the services that once promoted it, and that the need to carve out a brand identity for a platform is a key factor in this (pg. 332). However, he also attributes the decline in access to the companies' need to reduce user turnover rates that result from the "so-called 'paradox of choice'" (pg. 333). The argument follows that access to an overabundance of song choice means that users are far less inclined to want to seek out new music for themselves. As a result, these services have taken the emphasis away from the areas of their platforms that focus on access in favour of recommendation and curation features. Drott claims that these features promote a kind of "artificial scarcity" by offering their clients a "range of subject positions" that they can adopt without having to do the work of creating these relationships themselves (pg. 335). For instance, services may offer playlists based on mood or activities (e.g., workout, happy, party) or alternatively claim that recommendation services such as Spotify's "Discover Weekly" selection are akin to the age-old tradition of an older brother or sister passing down music to a younger sibling, thereby attributing the music with a sense of status. For the user, the perceived scarcity, and therefore desirability, in these platforms lies in finding the "perfect song" (pg. 336), and that to do so it is necessary to divide life and oneself up into moments and moods for which there is always the perfectly algorithmically curated playlist, the next song on which may well be this holy grail. For proof of this, Drott cites the advertised ethos of these companies themselves; Spotify founder Daniel Ek has reportedly declared that Spotify is "not in the music space – [they're] in the moment space" (Seabrook, 2014). Drott remarks that moments such as 'working out' or 'party' also happen to be ideal marketing segments for Gatorade or Bacardi (Drott, 2018). He concludes that the eventual outcome, and one that is encouraged by these companies, is the fragmentation of the listener into distinct selves, and that there develops a kind of dependency on this automated recommendation features.

YouTube music recommendations present much of the same algorithmic recommendation systems as their competitor platforms (Airoldi, Beraldo and Gandini, 2016). Consequently, the ideas proposed by Drott and Morris and Powers undermine the reliability of the data that we will be working from throughout this project. With regards to the specific questions of this research project, can we claim that foreign language songs are making their way into the mainstream English-speaking media due to a cultural shift towards acceptance of non-English music, or are there external influencing factors, i.e., corporate interests, that play a role in the charting success of these songs? An answer may lie in the content analysis of the YouTube comments on music videos of such songs. If the results display an overwhelmingly negative reaction from the public despite the popularity of the songs, it could signal artificially inflated streaming numbers. Thus, when discourse analysis is undertaken during the case studies, this is a dimension that will be firmly kept in mind as it represents the potential for large-scale skew for any conclusions drawn.

3) The politics of the language of music

It is important as academic researchers of foreign language music to examine critically the way in which we discuss music originating from what is commonly referred to as 'Western' cultures. Susam-Saraeva remarks that researchers are often influenced by the hegemonies present in the global music industry, that is to say, the global dominance of Anglo-American musical products (2019). The result of this is that the popular music from the rest of the world is frequently relegated to the domains of ethnomusicology or local branches of popular music. As a Digital Humanities scholar whose undergraduate training is in Modern Languages, it is important that I avoid playing into colonial biases and the first step towards doing so is to examine the nomenclature that I will be employing throughout the project.

In order to ensure that our study into global music is effectively decolonised, we must examine how we discuss the dichotomy between the native and the foreign. Born and Hesmondhalgh express some of the difficulties that we can encounter when conceptualising English and non-English language pop music as Western and 'other' (2000). According to Born and Hesmondhalgh, the label of 'Western' to refer to more affluent areas of the world has been disused as of recent in favour of a North/South conceptual division. However, within the context of musical studies there still remains the longstanding concept of 'Western music' (see: Burkholder, 2019). Born and Hesmondhalgh remark that at the time of writing, the term Euro-American emerged in lieu of 'Western' music, and since then it has seemingly been employed by musical scholars writing more recently (e.g.: Adeogun, 2018; Damodaran, 2016). It is perhaps unhelpful for this project specifically to discuss this concept of the West in purely geographical terms as we will be investigating foreign language songs in English-speaking locales in Oceania: Australia and New Zealand. Additionally, we can also observe that 'Western' is still employed today both for encyclopaedic works (Burkholder, 2019), and for critical works specifically examining East-West dichotomies in contemporary musical studies (Everett, 2021). The principal problem, however, with framing English-language music as 'Western' music in this project lies in the inevitable usage of the inverse: non-Western. In the words of Born and Hesmondhalgh, the term 'non-Western' carries with it the implication that "the rest of the world is a kind of residue of the West" (2000, pg. 47). This is especially problematic when we consider that what we really mean by Western music is specifically anglophone music heavily influenced by the culture and music industry of North America. By defining Western music solely as English-language music, we are necessarily relegating the rich and diverse languages and accompanying musical traditions of the European continent to the level of "residue" (Born and Hesmondhalgh, 2000, pg. 47). Additionally, it is not so simple to delineate 'Western' and 'non-Western' music simply with reference to the language in which a given song is sung. For example, the genre of reggaeton, as will be discussed further in chapter 4, is a genre whose song's primary sung language is Spanish and whose origins lie in Puerto Rico. Over the course of the last 20 years, however, it has found great success in the United States due in part to its ability to speak to younger Hispanic audiences (Rivera-Rideau, 2015) and eventually lead to the 'Latin Renaissance' marked by the success of 'Despacito' in 2017 (Acevedo, 2019). Indeed, many popular Reggaeton singers such as Becky G and Nicky Jam are born in the USA (Aguila, 2014; Raiford, 2021), and it is certainly not the place of this project to make assertions as to the cultural identities of these artists (or indeed any listeners and propagators of the genre). Consequently, it is no longer possible to assert to what degree reggaeton can be considered a North American (or 'Western') product. With this in mind, for the purposes of this project the non-English music will be referred to as just that – non-English or, given this is an English-language thesis, foreign-language music.

4) Positionality

In recognising the potential impact of the terminology used in this project, the value of recognising my own positionality as a researcher has been made clear. As I am writing in a domain that necessarily straddles quantitative and qualitative research, exactly pinning down the effects that it may have on the project is a complex task. In their paper on the way in which positionality should be expressed in quantitative studies, Jafar remarks that the value of positionality is well documented for qualitive subject areas (2018). In many cases, a researcher's positionality serves to contextualise both the individual and the results. Jafar contends, however, that the applications of positionality statements are less clear cut for quantitative research (2018). According to Jafar, whilst positionality is a "positive and integral element" of qualitative research (pg. 323), the element of bias that positionality introduces implies distortion and unreliability when considered from a more quantitative, statistical standpoint. Jafar urges us, however, to not consider the quantitative research process a purely positivist one, that is to say, one that produces an absolute true or false result. The following passage from the aforementioned paper sums the issue up well:

> *Ultimately, we all direct our research based on innumerable factors, most of which never make the page. This absence of positionality does not provide*

> *opportunity for the audience to decide how important these factors might be and as a consequence this reduces the validity of the research conclusions.*
>
> (Jafar, 2018, pg. 324)

Jafar's claim, here, certainly rings true for this project. Whilst I have consciously attempted to document as many of the relevant and most impactful decisions as possible, every script, research direction, and abstract concept employed was the result of thousands of decisions, some known to me but many others likely unconscious or influenced by unknown external factors. The most obvious example of where my positionality may have had the greatest tangible effect on the 'positivist' data is the machine learning model trained to identify opinion in YouTube comments in chapter 2. The data produced by this model has been used in chapter 3 and 4 to form the basis of the investigations therein. This was a supervised model, which meant that I had to supply it with trained data from which it could learn to differentiate between types of YouTube comment. As the sole tagger of the relatively large training set, much of the data was included or excluded on the basis of my own split-second, instinctive decisions. To claim that my own personal biases will not have influenced this model in any way would not only be inaccurate but also unethical. It is for this reason that I include a positionality statement in this project, not only to "define the boundaries" in which the qualitative aspects of the project were produced, but also to provide context to the emotional, social, and academic milieux within which the quantitative results of this investigation have been generated (Jafar, 2018, pg. 232).

I am academically trained as a modern linguist. My native language is English, but my undergraduate degree is in three European languages (French, Spanish, and German). This brings both advantages and disadvantages when examining the reception of foreign language songs - a large part of which will be investigating language attitudes from English-speakers towards foreign languages. The main benefit of my academic background lies in my own experience of immersion in various foreign languages. This has afforded me the opportunity to step back and examine my own preconceptions and value judgements towards language and language use which has informed my approach to this project. However, my academic training and personal commitment to languages do influence an emotional reaction to the expression of negative attitudes towards foreign language use, which may colour both the data produced and the conclusions drawn from these. My development as a computer scientist was entirely independent and, initially, focused on skills needed for this project. Consequently, the development of these skills was influenced by what I set out to achieve in this project as well as particular techniques and approaches which I saw as apt to integrate into digital modern languages studies.

Within the context of Digital Humanities principles, it also worth noting the implications of my position in terms of diversity and representation in research. I am a white male, educated in a British university; this brings with it particular scholarly conventions and privileges which, whether consciously or unconsciously, will affect the methodological approach taken, the samples drawn, and the analyses which derive from these. The multilingual and cross-cultural focus of my research is naturally underpinned by an emphasis on diversity when acknowledging my own position. Ultimately, there are myriad values and biases – both academic and social - that accompany that position, many of which are not known to me. The computational methodologies undertaken seek to partially mitigate the implications of such biases and it is my hope that the acknowledgement of my positionality and context will help furnish academics who read this thesis with the necessary insight into why I have drawn the conclusions I draw beyond the hard facts of the data itself.

Chapter 2: Methodology

1) Corpus collection and curation: Methodology

1.1) Initial considerations

A core tenet of one of the key strands of the Digital Humanities is the idea of moving from distant, computational reading to close reading (otherwise referred to as macro- and microanalysis), with the latter being able to be informed by conclusions drawn from the former (Jänicke *et al.*, 2015). Another primary goal of the field is to provide robust, flexible tools that can be reused or built upon by future researchers (Honn, 2014). The specific methodological needs of distant reading for this project become somewhat complex and invoke more advanced computer science concepts. It is for this reason that much of this project will be devoted to the clear and careful documentation, justification, and explanation of the methodological techniques employed. This section will be devoted to the process of the collection and curation of the corpus of YouTube comments. It is my hope that future researchers and practitioners will be able to closely follow the trajectory of the methodologies employed in this thesis so as to be able to replicate and improve upon them.

One of the initial, broad decisions when starting this project was selecting which programming language to use when creating these tools. As a novice to the field of computer science, it was key to select a programming language that was not only as beginner friendly as possible but that also fulfilled the aforementioned criteria of robustness, flexibility, and possessed the capacity to serve as a foundation for further scholarship. With this in mind, Python was the obvious choice. As Romano states, Python promotes readability, coherence, quality, and logicality (sometimes at the expense of efficiency; see chapter 6) which makes the language ideal for beginner programmers (2018). For example, provided a developer has taken steps to render their code readable and clear and a beginner programmer has some understanding of the broad mechanics of coding, they are able to very easily understand the intention of a script and the ways in which it can be manipulated. Romano also describes how Python also places an emphasis on integration with other languages, which is of primary importance to Digital Humanities projects with a focus on extensibility and reapplication. Finally, Python possesses an extensive library of third-party tools and packages. This is useful for novice programmers as it allows us to employ more advanced computer science techniques without having to write much of the complex, underlying code. For example, due to the many libraries dedicated to machine learning, I was able to perform AI-driven sentiment analysis within the first few months of my development as a programmer. This will be outlined further in section 2.3, however, it important that I first detail the methodological decisions leading up to the creation of the comments corpus upon which this project is built.

1.2) Creating the corpus

We are aiming to examine popular foreign language music in this project. Consequently, it was vital that we examined what we meant by 'popular' and also developed a metric through which to quantify this popularity. Following the example of Meindertsma, I elected to do so with the use of the singles charts (2019). As Stewart points out, using singles charts data as a representative metric of popular culture is not flawless (2012). Singles charts represent just one facet of the music industry and also have a history of being manipulated - we have already discussed the potential for manipulation by for-profit entities in chapter 1 in the age of streaming, however, according to Stewart chart data still remains a "crude but effective" method for measuring the impact of a song. Meindertsma remarks that previous studies have elected to use subsets of the charts, e.g., top 10, and are consequently limited in their scope. He addresses this limitation by collecting data for the entire Billboard Hot 100 for each week he examines. As we are investigating the reception of foreign language music across the broader English-speaking world, we need to take this a step further and collect data from multiple countries. For methodological reasons that will be outlined in section 1.3, I elected to collect data from the USA, Canada, Ireland, the UK, Australia, and New Zealand. In spite of the expanded scope of this project, I followed Meindertsma's example once again and collect the entire length of a given chart (2019). This ensured we would get a maximally holistic representation of the public attitude towards foreign language songs, albeit just in terms of commercial success. As the goal of this scraping was specifically to collect YouTube comments on popular songs, I decided to only collect charts dating back to August 11th, 2007 as this was the date from which the Billboard began incorporating streaming data into their weekly top 100 (Billboard Staff, 2007). The specific methodology of this step evolved drastically throughout the project, and I will detail exactly how and why this was in section 1.3.

After having scraped the charts data, the next step was to ascertain which of these songs contained lyrics in foreign languages. I originally considered querying databases in order to collect metadata about a given song, however after some research it became evident that (at the time this project was carried out) sources such as the Spotify developer API (an interface through which developers can officially make requests to access the company's database without using the Spotify app or website) or the iTunes catalogue did not publicly provide information about the language(s) used in the song. Furthermore, while there have been novel approaches to automatically identifying lyric languages by making use of Spotify user generated metadata such as song and playlist associations (Roxbergh, 2019), this was proprietary information to which I had no access. As a result, it became clear that the only way to ascertain what language(s) a song was sung in was to query a lyrics site and then perform language identification on the results. Thus, I designed another scraper for the popular lyric site https://www.azlyrics.com/ to this end.

I elected to source the lyrics from AZlyrics for sake of ease. The HTML of a given lyric page on this site contains a hidden, plaintext set of the lyrics so it was not necessary to parse too much of a page's raw HTML. Additionally, it contained a comprehensive database of over 300,000 lyrics (RapidAPI, 2021). Initially, 95 foreign language songs that have appeared on the charts in the last 13 years were identified. However, after an examination of the results it became clear that the script was not identifying the language of a large portion of the songs passed to azlyrics.com. This was due to the way that the website handles an incorrectly formatted URL. The URL of a lyrics page on azlyrics.com follows the format of https://www.azlyrics.com/lyrics/*artistname*/*songtitle*.html.

Consequently, the script would take the artist's name and song title for each entry in the data taken from the charts, remove spaces and punctuation, and insert them into the relevant sections in the URL. Naturally, it would be naïve to believe that both artist and song strings on acharts.com would be a perfect match for those pointing to the relevant lyrics page, especially as I was attempting to identify foreign language songs that would contain special characters. Thus, logic was added to the script to catch a 'page not found' error for those incorrectly encoded songs. However, when presented with an incorrect URL of this sort, azlyrics.com does not return a 404 error, and instead returns a blank page. It is unclear whether this is an intentional feature to combat web scrapers, but as a result the approach needed to be altered. I am not the first developer to scrape AZlyrics, and there exists a Python library designed to this end that contains a function to use search engines to navigate to lyrics pages. However, use of a search engine to make automated requests violates their terms of use. Thus, it was deemed that it would be inappropriate to do so for an academic research project such as this. Making use of the lyrics site's own search function does not violate any user agreements, however, and provides a solution to this particular problem. This workaround is not without drawbacks – due to the aforementioned steps taken to make my scraping more ethical, each song would now take between 30 to 40 seconds to process, and the entire script would take around 4 to 5 days to fully complete assuming that no IP bans were issued or any other exception arose. To counter this problem, I decided to follow the example of Meindertsma and query multiple lyric sites (2019). Further emulating Meindertsma, I chose https://www.lyrics.com/ as the target of my requests. I initially dismissed this site as my own research proved their database to be less extensive than AZLyrics and they have much less stringent moderation of their user-provided lyrics. This meant that there can be multiple versions of the same set of lyrics which can confound an automated approach and lead to less overall reliability. However, their website is much more receptive to scrapers, meaning shorter delays between requests. Consequently, I elected to use AZlyrics as a 'backup' source when a song did not exist on lyrics.com. As a result, the overall runtime was cut down to around 6 hours and a number of additional foreign language songs were identified.

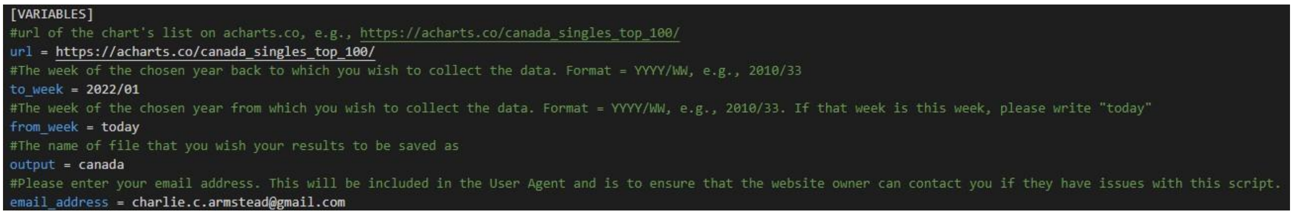
There were about 1000 songs that were not found in either AZlyrics or Lyrics.com archives. A simple language analysis was conducted on the titles of these tracks to identify the non-English ones. Naturally, there still exists songs containing a foreign language due to a non-English speaker featuring on them but still have a title in English, however, it is safe to assume that if a song does not have enough presence to have a lyrics page dedicated to it, it is unlikely to have a corpus of YouTube comments of any meaningful size associated with it, and thus it is not worth the time to comb through the 1000 unclassified songs to identify manually the language of the lyrics.

With the above revisions to the methodology, we were able to identify 291 foreign languages songs in the charts of the six chosen countries. After this, the next step was to collect YouTube comments on the videos of these songs. Initially, I attempted to scrape the comments from the HTML of the YouTube page themselves, however this proved to be an extremely time-consuming endeavour that often ended abruptly with errors. Thus I elected to use the YouTube API V3 and a script written by a GitHub user named hobogalaxy (2020). My reluctance to use the API stemmed from the need to supply an API key with each request that carried a fixed limit on how many comments one can download in a day (1 million) and also the fact that it does not support the downloading of replies to replies within a comment thread. These drawbacks, however, were outweighed by the ease and speed of the service; the YouTube API returns a simple response containing all the information requested so there is no need to parse any HTML and also provides many hundreds of comments per second. The main modification to hobogalaxy's script was the addition of the capability to retrieve the replies to a given comment. Additionally, there was also some refactoring regarding changing the Python package used to make the HTTP request from *urllib2* to the more updated *Requests* package and also changing the output file format from a plain .txt file to a nested JSON object.

Once these comments were downloaded, they needed to be pre-processed in order to become compatible with a machine-learning sentiment analysis approach. This will be discussed in section 2, however first it is necessary to detail some of the changes to the above approach and, most importantly, how doing so ensured that we were in line with the ethos and ethics of the Digital Humanities.

1.3) Fulfilling the goals of the project – flexible tools and FAIR data

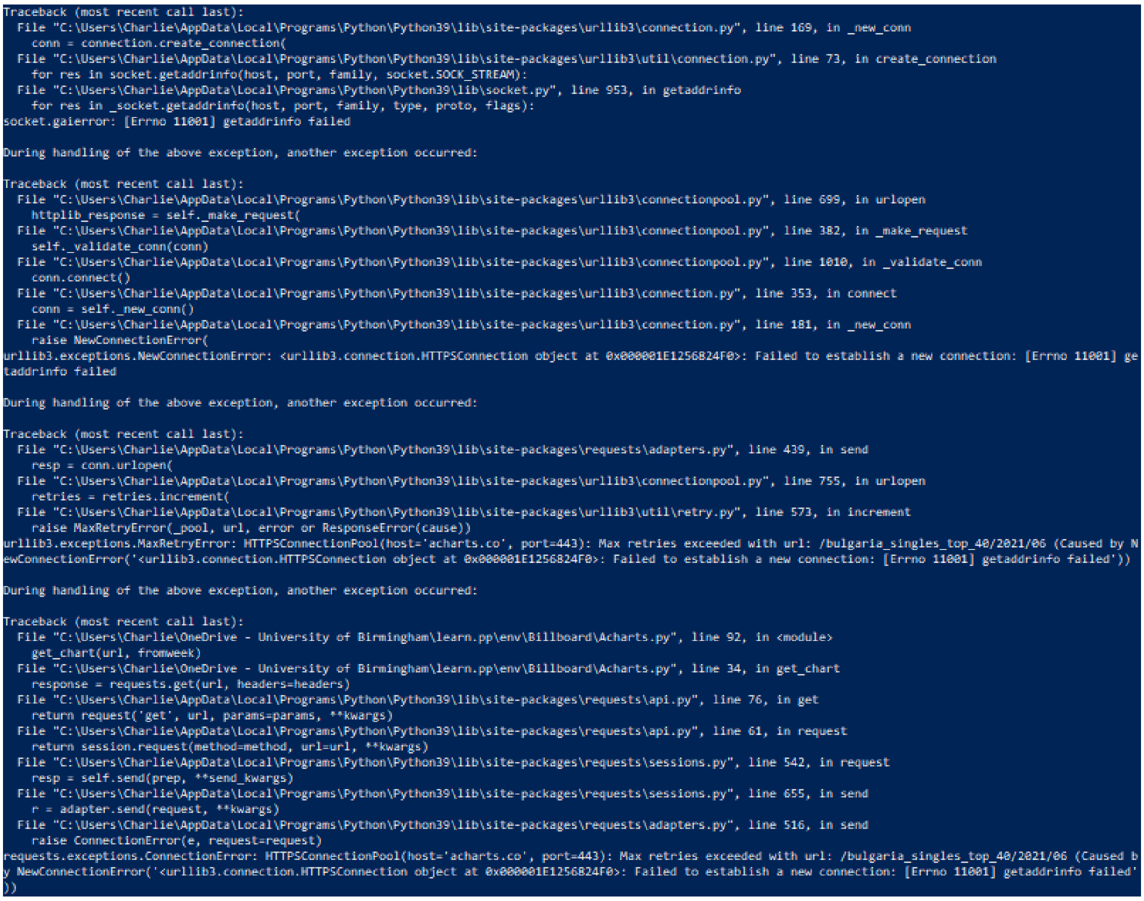
Martin Fowler remarks that “any fool can write code that a computer can understand. Good programmers write code that humans can understand” (1999, pg. 15). Given that I developed my Python skills through an exploratory process, much of the ‘foolishly’ designed groundwork of the tools I have created needed to be refactored thoroughly prior to the completion of the project. Refactoring is a process through which the structure of the code is altered (without affecting the external functionality) with the aim of improving its readability and general design (Fowler, 1999). This is important as, broadly speaking, refactoring renders code both more maintainable (Martin, 2009) and more extensible (Kerievsky, 2005). This is naturally desirable for a project that aims to provide tools for future researchers to either employ themselves or to repurpose for similar uses. For the same reasons it was also a key step in fulfilling a secondary goal of the project: my own personal development towards becoming a “good programmer” and also a good Digital Humanist, as a core component of both is the ability to produce work that is easily extensible and has a wide array of reapplications.



```
[VARIABLES]
#url of the chart's list on acharts.co, e.g., https://acharts.co/canada_singles_top_100/
url = https://acharts.co/canada_singles_top_100/
#The week of the chosen year back to which you wish to collect the data. Format = YYYY/WW, e.g., 2010/33
to_week = 2022/01
#The week of the chosen year from which you wish to collect the data. Format = YYYY/WW, e.g., 2010/33. If that week is this week, please write "today"
from_week = today
#The name of file that you wish your results to be saved as
output = canada
#Please enter your email address. This will be included in the User Agent and is to ensure that the website owner can contact you if they have issues with this script.
email_address = charlie.c.armstead@gmail.com
```

Figure 2.1: Example usage of config file for acharts.co scraper

Key scripts that needed to be refactored were the scrapers used to collect official charts data from the corresponding websites. In their initial format, they posed several problems regarding accessibility and extensibility. Firstly, the scripts could not be taken and reapplied to other charts websites as the specificities of the HTML would not be the same. With this in mind, I decided that I would write a script that could scrape charts data from several different countries with the use of the website https://acharts.co/. Whilst not exhaustive, at present Acharts displays data from 21 different countries as well as global charts data. As a result, a scraper of this website would be a much more powerful tool. The advantage of scraping official charts websites is that the data is reliable. However, given that I already possessed an official dataset thanks to earlier efforts, I was able to conclude that the data on Acharts can be considered reliable for future research use through comparative analysis. Ease of use was another key consideration when designing this script. So that future users would not have to edit the variables in the script manually, I included a config file with clear instructions within the repository. Figure 2.1 displays an example usage of this config file for clarity. This ensured that even users with minimal programming knowledge would be able to access and manipulate the tools. Secondly, a robust system for catching errors that may occur was included so that these less-experienced users who find the technical language inaccessible when an exception is raised may understand what issues have arisen if the script is does not function. As a novice with no experience of error messages in the command-line, I found it challenging to determine the reason for the error message displayed in figure 2.2. The reason is simply that the internet connection failed while the script was running. By including clauses that caught different types of errors, this would allow users with less-developed technical skills to troubleshoot problems themselves if they arose.

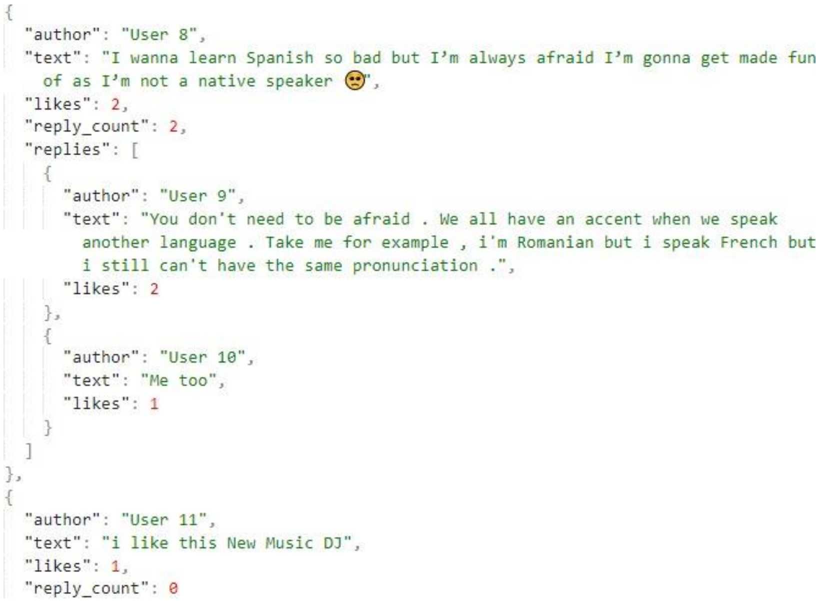


```
Traceback (most recent call last):
  File "C:\Users\Charlie\AppData\Local\Programs\Python\Python39\lib\site-packages\urllib3\connection.py", line 169, in _new_conn
    conn = connection.create_connection(
  File "C:\Users\Charlie\AppData\Local\Programs\Python\Python39\lib\site-packages\urllib3\util\connection.py", line 73, in create_connection
    for res in socket.getaddrinfo(host, port, family, socket.SOCK_STREAM):
  File "C:\Users\Charlie\AppData\Local\Programs\Python\Python39\lib\socket.py", line 953, in getaddrinfo
    for res in _socket.getaddrinfo(host, port, family, type, proto, flags):
socket.gaierror: [Errno 11001] getaddrinfo failed

During handling of the above exception, another exception occurred:

Traceback (most recent call last):
  File "C:\Users\Charlie\AppData\Local\Programs\Python\Python39\lib\site-packages\urllib3\connectionpool.py", line 699, in urlopen
    httplib_response = self._make_request(
  File "C:\Users\Charlie\AppData\Local\Programs\Python\Python39\lib\site-packages\urllib3\connectionpool.py", line 382, in _make_request
    self._validate_conn(conn)
  File "C:\Users\Charlie\AppData\Local\Programs\Python\Python39\lib\site-packages\urllib3\connectionpool.py", line 1010, in _validate_conn
    conn.connect()
  File "C:\Users\Charlie\AppData\Local\Programs\Python\Python39\lib\site-packages\urllib3\connection.py", line 353, in connect
    conn = self._new_conn()
  File "C:\Users\Charlie\AppData\Local\Programs\Python\Python39\lib\site-packages\urllib3\connection.py", line 181, in _new_conn
    raise NewConnectionError(
urllib3.exceptions.NewConnectionError: <urllib3.connection.HTTPSConnection object at 0x000001E1256824F0>: Failed to establish a new connection: [Errno 11001] ge
taddrinfo failed

During handling of the above exception, another exception occurred:

Traceback (most recent call last):
  File "C:\Users\Charlie\AppData\Local\Programs\Python\Python39\lib\site-packages\requests\adapters.py", line 439, in send
    resp = conn.urlopen(
  File "C:\Users\Charlie\AppData\Local\Programs\Python\Python39\lib\site-packages\urllib3\connectionpool.py", line 755, in urlopen
    retries = retries.increment(
  File "C:\Users\Charlie\AppData\Local\Programs\Python\Python39\lib\site-packages\urllib3\util\retry.py", line 573, in increment
    raise MaxRetryError(_pool, url, error or ResponseError(cause))
urllib3.exceptions.MaxRetryError: HTTPSConnectionPool(host='acharts.co', port=443): Max retries exceeded with url: /bulgaria_singles_top_40/2021/06 (Caused by N
ewConnectionError('<urllib3.connection.HTTPSConnection object at 0x000001E1256824F0>: Failed to establish a new connection: [Errno 11001] getaddrinfo failed'))

During handling of the above exception, another exception occurred:

Traceback (most recent call last):
  File "C:\Users\Charlie\OneDrive - University of Birmingham\learn.pp\env\Billboard\Acharts.py", line 92, in <module>
    get_chart(url, fromweek)
  File "C:\Users\Charlie\OneDrive - University of Birmingham\learn.pp\env\Billboard\Acharts.py", line 34, in get_chart
    response = requests.get(url, headers=headers)
  File "C:\Users\Charlie\AppData\Local\Programs\Python\Python39\lib\site-packages\requests\api.py", line 76, in get
    return request('get', url, params=params, **kwargs)
  File "C:\Users\Charlie\AppData\Local\Programs\Python\Python39\lib\site-packages\requests\api.py", line 61, in request
    return session.request(method=method, url=url, **kwargs)
  File "C:\Users\Charlie\AppData\Local\Programs\Python\Python39\lib\site-packages\requests\sessions.py", line 542, in request
    resp = self.send(prep, **send_kwargs)
  File "C:\Users\Charlie\AppData\Local\Programs\Python\Python39\lib\site-packages\requests\sessions.py", line 655, in send
    r = adapter.send(request, **kwargs)
  File "C:\Users\Charlie\AppData\Local\Programs\Python\Python39\lib\site-packages\requests\adapters.py", line 516, in send
    raise ConnectionError(e, request=request)
requests.exceptions.ConnectionError: HTTPSConnectionPool(host='acharts.co', port=443): Max retries exceeded with url: /bulgaria_singles_top_40/2021/06 (Caused b
y NewConnectionError('<urllib3.connection.HTTPSConnection object at 0x000001E1256824F0>: Failed to establish a new connection: [Errno 11001] getaddrinfo failed'
))
```

Figure 2.2: Example of a complicated error message arising from a failed internet connection.

Finally, I ensured that the naming conventions used in both this program and other programs written throughout this project were as intuitive as possible and that they adhered to the wider conventions of the programming community outlined by Romano (2018). I also provided comments in the code itself to elucidate the roles of specific functions and objects within the scripts so that if I or other researchers wished to refactor the programs further, the process would be as simple as possible.

This is just one of the web scrapers designed throughout the course of this project, and since they played such a foundational role, it was vital to consider the ethics of the practice. This was to ensure that it reflected that of a 'good digital humanist' but also to ensure that any future researchers using these tools would not be implicated in unethical scraping practices. When starting this project, I was unaware of the implications that unethical web scraping carried such as placing strain on physical and financial resources of the target website. Consequently, I invested time in trying to circumvent countermeasures to automated scrapers by rotating user agents to make it appear like my requests were coming from different, human users and additionally employing proxies to hide my IP address. However, as my general knowledge of the field developed further, it became clear that these techniques were unfair not only to the websites whose data being accessed but also any future researchers who may wish to employ tools created in this project but perhaps lack the knowledge to understand exactly how they work. As a result, I refactored the scripts to take into account the ethical considerations outlined by Densmore (2017). This principally involved providing a User Agent that makes my intentions clear by stating a contact email address (see figure 2.1) and also using highly conservative delays between requests so that the script is not confused with a DDoS attack (an attack on a service that attempts to flood its bandwidth with automated requests).



```
{
  "author": "User 8",
  "text": "I wanna learn Spanish so bad but I’m always afraid I’m gonna get made fun
    of as I’m not a native speaker 😟",
  "likes": 2,
  "reply_count": 2,
  "replies": [
    {
      "author": "User 9",
      "text": "You don't need to be afraid . We all have an accent when we speak
        another language . Take me for example , i'm Romanian but i speak French but
        i still can't have the same pronunciation .",
      "likes": 2
    },
    {
      "author": "User 10",
      "text": "Me too",
      "likes": 1
    }
  ]
},
{
  "author": "User 11",
  "text": "i like this New Music DJ",
  "likes": 1,
  "reply_count": 0
```

Figure 2.3: Excerpt from the JSON of comments downloaded from the music video of "La Isla Bonita" by the cast of Glee.

A cornerstone of Digital Humanities’ data sharing principals are the “FAIR Guiding Principles for scientific data management and stewardship” (Wilkinson et al., 2016). These guidelines are designed to “improve the **F**indability, **A**ccessibility, **I**nteroperability, and **R**euse of digital assets” (Go FAIR, 2021). To this end, I elected to store all data as JavaScript Object Notation (JSON). The reasoning behind this decision was that JSON has been used by developers for two decades and has been adopted by virtually all programming languages (Romano, 2018), and as a result the data is maximally compatible. The data produced would also still be human-readable and analysable when expressed in the hierarchical structure supported by the JSON format (Wasser and Joseph, 2017). For instance, in figure 2.3 it is easy to parse at a glance the information contained within the data structure. Thus, future researchers who wish to carry out more traditional analyses in lieu of employing computational techniques may still employ the tools I have created.

Peter Meindertsma displayed the results of his investigation in the form of an interactive graph on his website (Meindertsma, 2020). This is an effective way to ensure maximal accessibility, and whilst it would be ideal to create such user-friendly tools out of the scripts and programmes I have written, it is out of the scope of this project. It is still possible, however, to host the datasets and scripts at a dedicated web address like that of Meindertsma. Thus, future scholars can access all the resources associated with the project at https://ursidaeic.github.io/.

2) Pre-processing of comments

2.1) Initial steps

Any quantitative study involving textual data requires the strings to be converted into numbers in order to be computer-readable (Denny and Spirling, 2018). Consequently, it is vital that we process this data prior to transforming it into numbers so that the data we wish to extract becomes maximally salient. One of the primary concerns regarding this pre-processing of the corpus was whether to limit the length of the comment to a maximum character limit. Statistical analysis showed that the mean comment length was 50 characters with a standard deviation of 360. This standard deviation was inflated by some anomalous comments that were hundreds of thousands of characters long and were nonsensical; for example, the longest comment collected was 934431 characters long and was just a string of repeating numbers. When the longest 100 comments were removed, this standard deviation reduced from 360 to 189, however I decided that a 440-character limit was still short enough that it would exclude meaningless data such as that described above or simple postings of the lyrics, which is also a common practice and unhelpful data for the goals of this project.

After having filtered comment strings by length, the next step was to remove non-English comments from the corpus as this project is interested in English-speaking audiences. From the outset, it was necessary to consider what qualifies an English-speaker in this context. In 2011, it was reported that 60% of YouTube users selected a language other than English as their interface language (Roettgers). Given that these are foreign language songs, it is therefore surprising that only 64.2% of the comments were in a language other than English (the methodology through which this was uncovered is outlined below). This could partially be due to imprecisions in the methodology that will be addressed later, but also it is highly likely that many non-native speakers are making the conscious choice to write in English. Without access to propriety back-end user data, it is impossible to gain an insight into where a user is posting from beyond the information that they divulge themselves. Furthermore, excluding data that contain serious grammar or spelling errors on the basis that they are written by non-English speakers would ignore the fact that these are both hallmarks of informal 'text speak' or alternatively that they could be written by native English speakers with language and writing related challenges. Besides, even if the author claims to be from a non-English-speaking country in the comment and thereby identifies themselves as a non-native speaker, if they have made the choice to post in English, they are necessarily part of the English-speaking world. Not only would it be impossible to create a quantitative classification system of what it means to be an English-speaker, it implies that there is a direct link between national and linguistic identity that, although an important object of study, does not then pertain to the goal nor the domain of this project. As a result, the only course of action is to include all English-language comments, regardless of any assumptions we may make about the author's native language.

The broad script I used to detect language in comments made use of the PyCld2 library (aboSamoor, 2021), however the design of this library was not optimised for shorter strings and manual surveying proved that many foreign language comments had returned a false positive. Consequently, if a comment contained less than 10 words after emojis were removed, the program would iterate through each word and run a spell check on it to see if it was either found in the English language or was a close enough misspelling of one with the use of the PySpellchecker library (Barrus, 2021). Then, if 50% or more of the words in the string returned a positive match, the comment was deemed to qualify as an English-language comment. The reason that only a 50% match was needed to qualify was due to the fact that many comments mention the artists' names, which, as proper nouns, were not included in any dictionary. That, in conjunction with the fact many of these comments use specific dialects and slang that may not be included in the dictionaries of the spell-checker (an issue that reoccurs in later steps), required the parameters to have some level of flexibility. Whilst this method was not 100% effective and there still remained some foreign language comments in the corpus, the number was vastly reduced, and it is likely that some alterations to the above method such as more stringent parameters or perhaps reversing the order of the two steps would yield improved results. Overall, with the above techniques the corpus size was reduced from 68,055,630 comments to 24,416,623.

2.2) Spell correction, lemmatisation, and stopping.

YouTube comments represent a unique resource for sentiment analysis. This is firstly because they are they very short texts and thus tend to be highly focussed on a particular topic. Secondly, the quality of the written English is often lower than would be found in other media or makes use of a dialect highly specific not just to YouTube, but to a fanbase on the platform.

Before the data could be tokenised and stopped, it was key to normalise spelling between the comments, not just by correcting misspellings, but by ensuring that abbreviations and contractions remained consistent between comments. Clark and Araki demonstrated that spell-checking alone is inadequate when processing social media posts that make use of casual English due to the range of variations in spelling (2011). These authors propose a methodology through which different misspellings are placed into categories - for example abbreviation (‘laugh out loud’ being shortened to ‘lol’) represents a distinctly different form of ‘misspelling’ to the omission of an apostrophe. Satapathy et al., take a more general approach of performing phonetic analyses on misspelled words in casual English so as to identify the intended word through the use of more advanced computational techniques (2017). These are both ideal solutions and are valid options for future researchers, but the constraints of this project meant that a lower resolution technique needed to be implemented. Apostrophe of contraction such as “dont” and “Im” were corrected and a list of common abbreviations was downloaded from webopedia.com and then expanded out (Beal, 2021). On top of this, a Python port of Wolf Garbe’s famously rapid Symspell algorithm was employed to correct other misspellings (Garbe, 2019). Next, I lemmatised and removed stop words from the texts. Stop words are words that are deemed to add no value to a sentence and are a removed by “stopping” a text, and lemmatisation is a technique whereby the different inflections of words are normalised (e.g., “runs”, “ran”, “running” all become “run”). Finally, the text was tokenised – a term which described the breaking up of a sentence into individual words or ‘tokens’. Throughout this process, I had elected not to remove emojis as they form a key part of expression of emotion in informal text. However, the number of emojis in a row was reduced to a maximum of 3 as frequently commenters would use upwards of 10 of the same emoji, and if they were not truncated it would lead to a noisier text.

2.3) Machine learning (ML) approaches

For case study 1 and 2, we wished to investigate the opinions that users are expressing towards foreign language songs. This is so we could uncover different facets of the ways in which English-speakers receive these songs – namely concepts around fandom and language attitudes. To this end, we needed to further filter our corpus of comments by sentiment. The most effective way to do this was with machine learning (ML) techniques - specifically through sentiment analysis.

Due to the extensive use of slang and dialect in YouTube comments, it was not viable to use pretrained models for sentiment analysis. For example, no positive or negative sentiment was detected in the following comments when the open source TextBlob sentiment analyser was applied to them (Loria, 2018):

“do not speak it but its fire” - anon

“QUEEN OF SLAYING MY ENTIRE EXISTENCE” - anon

“This is a bop ″- anon

Pre-trained models simply have not been exposed to the type of language typically found and employed within this specialised corpus, and thus it was necessary to create and train my own classification model.

Kharde and Sonawane outline two main machine learning approaches for sentiment analysis: unsupervised and supervised algorithms (2016). The former function by basing an algorithm upon rules or conditions from which to draw conclusions regarding the polarity of a given text. The latter work from ‘perfect’, pre-annotated datasets from which the algorithm can build predictive models. Kharde and Sonawane conclude that supervised learning algorithms are preferred for sentiment analysis. Consequently, I initially decided that I would make use of a Naïve Bayes classifier – specifically the pre-designed model found in the Natural Language Toolkit package (Bird et al., 2009). By providing this algorithm a manually tagged dataset of positive and negative comments, we would be able to divide up the corpus in the same way for the case studies to come. A Naïve Bayes classifier functions differently to other models as, when a given string is converted into a numerical format that the algorithm can parse, the order of the words is preserved, and thus so are grammatical and syntactical features. My initial hypotheses that the addition of these sentence-level features would prove to be effective were supported when tests on a modestly sized training set of 500 comments (50:50 positive/negative split) produced accuracy scores of around 70%. However, while tagging the training set it became clear that there was a huge discrepancy between the respective numbers of positive and negative comments; for every comment expressing a negative sentiment toward a song, there were roughly 13 positive ones. This revelation required some reformulating of the aims of the methodological process. This was because there now seemed little value in attempting to investigate the differences between the polarity of reactions of YouTube users if there were 13 times more positives than negatives. Thus, I decided to proceed by trying to answer the broader question of 'what are users saying about these songs' instead of being constrained by the binary positive vs negative.

An advantage of this change of focus was that dividing the data into two classes– opinion-carrying vs opinion-free – as opposed to positive/negative/neutral allowed me to experiment with different ML models given I was now dividing my data into two classes. A Support Vector Machine is a commonly used tool in sentiment analysis and broadly functions by converting a set of strings into human-unreadable datapoints and plotting them in a 2d space, whereupon it draws a hyperplane between the two classes, as shown in figure 2.4. Then, any new data is transformed and plotted in the same way and, depending on which side it falls, it can be classified. After retagging a new dataset of 3000 comments, I compared the Naïve Bayes classifier and an SVM from the Scikit-learn package (Pedregosa et al., 2011) and, despite my earlier hypothesis regarding the advantage of syntax-preserving algorithms, the SVM performed on average 7% better with an accuracy score of 0.82. This could be because strings tend to be too short for syntactical features to become salient within the classification framework. In sum, however, with the use of the SVM an average of 42.5% of the comments on a given video were identified as containing an opinion regarding the song to some degree.

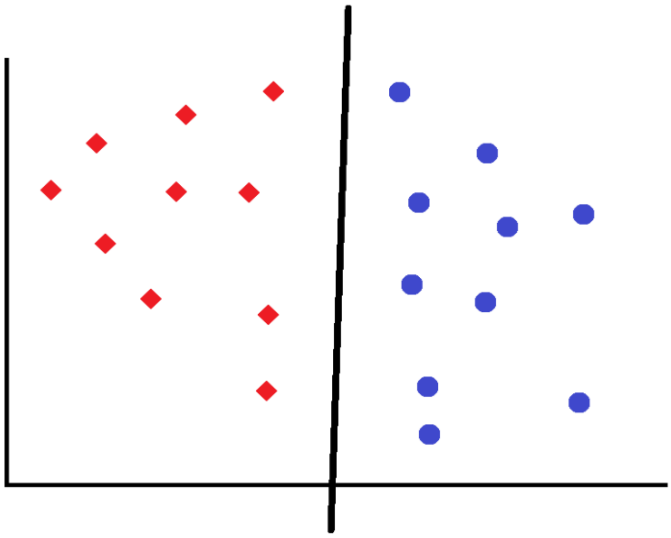


Figure 2.4: Simplified representation of how an SVM would draw a separating hyperplane between two classes.

We are now left with three different corpora of comments. The first is the raw comments from the videos, none of which had been spellchecked or filtered from language. This corpus is not helpful for the aims of this project as we are only interested in the English-language comments. The second is the English-language comments filtered from the raw comments. Finally, we have a corpus of comments containing opinion. These latter two corpora form the base upon which the following case studies are built.

Chapter 3: Exploring the reception of BTS and BLACKPINK through the lens of fandom

1) Song selection, context, and rationale

The term *Hallyu* - meaning Korean wave - refers to the rapid expansion and spread of Korean cultural artifacts that has been occurring since the 1980s (Farrar, 2010). Korean pop, or K-pop, is one such artifact and has found extreme success across Asia, Eastern Europe, and South America (Russell, 2012). However, it is not until recently that K-pop has found mainstream success in English-speaking countries (see table 3.1). It is widely documented that the idea of fandom is central to this success (Han, 2017; J. O. Kim, 2021; Otmazgin and Lyan, 2014), particularly on YouTube (Swan, 2018). To explore this concept of fandom further within the borders of the English-speaking online space, this case study will examine the work and reception of certifiably (see below) the two most popular K-pop groups in the English-speaking charts: BLACKPINK and BTS.

1.1) K-pop: conventions of the genre and initial hypotheses

According to Rousse-Marquet, K-pop is characterised by three key features: a) careful curation and selection of singers, often referred to as ‘idols’, b) specific choreographic, musicological, and sociological features, and c) the formation and curation of large fandoms (2012). Firstly, K-pop idols are intensively trained by huge record labels from their late teens in order to become masters of their craft and “multi-purpose stars” who are capable of acting, singing, and dancing. These idols are then formed into groups of around 4 members with each member picked with the intention of fulfilling a role within that group (Kim, Hwang and Kim, 2021). For example, one idol may be the leader of the group while another the specialised dancer. Rousse-Marquet also states that the music itself is typically pop in style that features blends of synthesized music, hip-hop and rock sounds (2012). Rousse-Marquet continues that the songs are typically accompanied by strong visual elements such as fashionable outfits or sharply choreographed dance routines. This extends to the singers themselves who are chosen because they are deemed to be ‘conventionally attractive’, whether naturally or, often, as a result of cosmetic procedures. K-pop has also been packaged with the express intention of creating an international product; whilst initially aspiring to break into the Japanese market during its nascent stages in the 1990s and early 2000s, more recently many K-pop labels have attempted to find success in Western markets, in particular in the USA. To this end, many of these songs feature a section of rap or occasional lyrics in English. In spite of this, many acts have still struggled to find success in Western anglophone markets due to the target audience's reticence to engage with foreign language music (Boman, 2019). Finally, and most importantly for this case study, a core feature of K-pop is its creation and curation of fandoms focussed on the bands and idols. Elfving-Hwang points to efforts on the part of entertainment management companies to foster parasocial relationships between fans and their idols (2018). At the very extreme end of the spectrum, these relationships can result in fans undergoing cosmetic surgeries in order to look like their favourite artists or in some cases strings of 'copy-cat suicides' following celebrity suicides.

Elfving-Hwang focuses on Korea for the majority of her analysis, however, K-pop fandoms have been seen to possess significant presence in Latin America (Han, 2017) and the Middle East (Otmazgin and Lyan, 2014). Thus, I hypothesise that this idea of fandom must also constitute a core part of the way in which English-speaking listeners of K-Pop songs receive these foreign language songs. While the aforementioned international appeal of K-pop songs undoubtedly contributes to their success in English-speaking markets, I contend that the driving force behind the expansion of this *hallyu* is the expansion of fandoms devoted to the idols in the bands.

1.2) BLACKPINK and BTS: Rationale and initial analyses

It was not until 2012 that Korean-language music reached the English-speaking markets with PSY's 'Gangnam Style'. However, 'Gangnam Style' is widely considered by fans not to be 'real' K-Pop as PSY does not fit with the aforementioned norms and styles of the genre (Boman, 2019). PSY is considered to be "chubby" and "unattractive" in stark contrast to other more conventionally attractive young male singers (Lie, 2014, pg. 60). Thanks to the steps taken to identify charting foreign language, we have access to detailed data pertaining to song charting positions. Table 3.1 shows the year and week of the year in which Korean-language artists debuted in the charts. After PSY, BTS and BLACKPINK were the next earliest charting Korean-language artists. Unlike PSY however, both these groups adhere much more closely to the aforementioned conventions of the K-pop genre (Boman, 2019). Additionally, BLACKPINK and BTS both boast a much greater level of English-language engagement on YouTube: even if the incredibly popular songs by PSY are counted, the total corpus size of comments collected through the course of this project for all other Korean-language songs is just 5,683,389 English-language comments, whereas BLACKPINK and BTS have accrued 7,091,067[2] and 11,959,084 comments respectively[3]. Furthermore, Jennie, J-Hope, Jungkook, and Agust D -are all members of either BLACKPINK or BTS and their solo success in English-speaking markets could be attributed that of their group.

Table 3.1: Table showing the week in which Korean-language first charted in English-speaking countries according to data scraped from acharts.co

ARTIST	YEAR/WEEK
PSY	2012/36
BTS	2016/44
BLACKPINK	2016/47
Red Velvet	2018/07
Jennie	2018/48
J-Hope	2019/40
TWICE	2019/40
SuperM	2019/42
Jungkook	2020/10
ITZY	2020/13
Agust D	2020/22

The concurrent emergence of BTS and BLACKPINK and the fact that they have enjoyed a much greater level of success when compared to other groups within the K-pop genre suggests that there lies potential value in comparing the ways in which the public receives songs by these two artists. This is because there is no need to account for diachronic factors when the reception of the artists is compared. To expand, as both bands entered the charts (and therefore ostensibly the sphere of mainstream, popular music in the English-speaking world) simultaneously, it cannot be argued that differences in reception between the two bands is due to the English-speaking public's general increase in receptiveness to foreign language music. Thus, differences in public reactions to these two artists may then be attributed to other factors such as musicological or contextual features of

[2] This number has attributed to it the comments on the song Sour Candy by Lady Gaga on which BLACKPINK was a featured artist. As a note, this song will be included in some of the musicological analyses of BLACKPINK because, as will be observed later, it follows a similar pattern of engagement as pure BLACKPINK songs.

[3] These figures may differ slightly from actual numbers due to the fact that, as outlined in chapter 2, the method used to collate this corpus only permitted the downloading of top-level comments and its direct replies. At present, replies to replies remain inaccessible

the songs. It should be noted that songs that have made it to the charts by BTS and BLACKPINK but whose the lyrics do not contain any Korean language have been excluded from the following analyses. Comparing the reception of pure English-language songs to those that contain foreign language would be a useful way of exploring opinion towards the latter, however, within the context of this case study and its aims, including such songs would undermine the reliability of any conclusions drawn from the data. This chapter tackles the question of fandom and its influences on a K-pop song's success and a purely English-language song is more likely to outperform a song with a large quantity of Korean in it. As a result, there will be an anomalously high number of YouTube comments from users who do not participate in the fandoms associated with these songs. Additionally, those that do ascribe themselves to the fandom may engage with the media in a different way if it differs from the norms of the respective band in terms of language use, thereby further confusing the results.

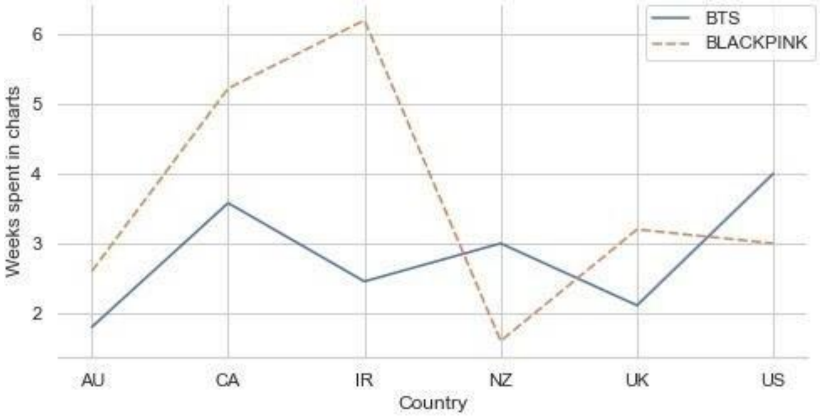


Figure 3.1: Line graph visualising the mean number of weeks BTS and BLACKPINK songs spent in the charts by country

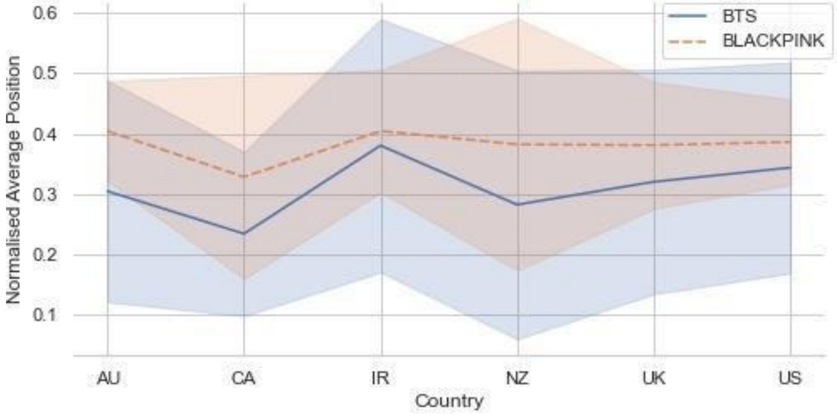


Figure 3.2: Visualisation of the average charting position for a song (line) and standard deviation of song placements (shaded area) by country. Chart position has been normalised to a decimal between 0 and 1 as different countries employ different chart lengths.

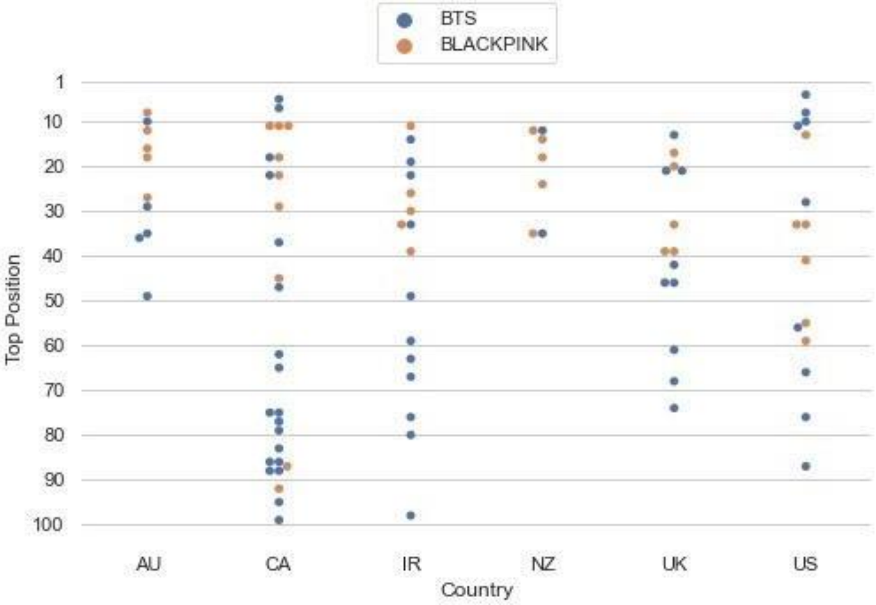
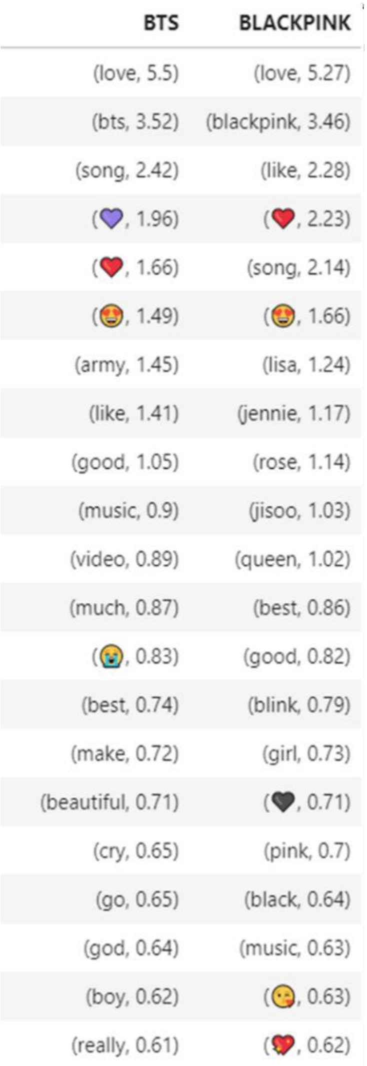


Figure 3.3: Scatter plot showing highest charting positions of individual BLACKPINK and BTS songs by country

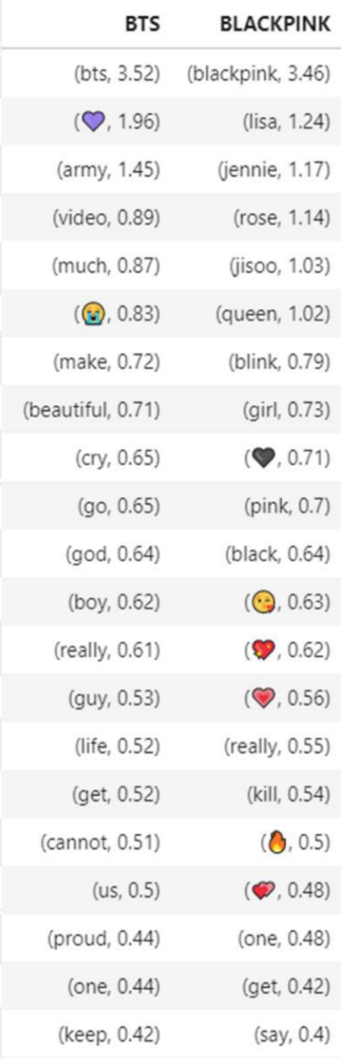
Cursory, low-resolution analysis of official charts data collected during the project gives a broad indication of the general public's attitude toward these two artists in contrast. Figure 3.1 shows comparatively the average of the number of weeks that songs by BTS and BLACKPINK spent in the charts. In four countries (Australia, Canada, Ireland, and the United Kingdom), BLACKPINK songs tend to spend more time in the charts than those of BTS - the most pronounced instance of this being in Ireland where BLACKPINK songs chart for over twice as long as BTS singles. This preliminarily suggests that BLACKPINK is generally better received than BTS. Figure 3.2 elucidates the matter further. The data visualised demonstrates that, while BLACKPINK on average charts higher with their songs, BTS has a much greater standard deviation in their positions and thus a greater range of chart positions. This conclusion is visualised in figure 3.3 - a scatter plot of each song's highest chart position, again sorted by country. With the exception of Canada, BLACKPINK songs tend to have a peak chart position in the top half of a given chart, whereas there is no particular trend in where a BTS song places. However, out of the seven K-Pop songs by the two bands that managed to break the top 10 across all countries, only one belonged to BLACKPINK.

The obvious skew that needs to be considered when trying to draw conclusions from this statistical data is that the number of BTS songs found on the charts, and indeed in their discography, vastly outnumbers those of BLACKPINK. This adds complexity to the data as, of BLACKPINK's 20 released songs (Ailin, 2021), 8 have managed to enter into the English-speaking musical sphere and while BTS have done the same with 21 singles, their discography contains 157 songs (Big Hit, 2021). Thus, if one were to use the proportion of an artist's songs that have reached the charts as a metric for popularity, it would be indisputable to claim that BLACKPINK is the more popular artist. However, in reality the fact that around three times more BTS songs have appeared in English-speaking charts suggests that BTS is the Korean-language artist that has most entered the public consciousness of English-speakers. We can assert, however, that when a BLACKPINK single reaches the charts, it performs on average better than a BTS song, but it is not possible to draw any qualitative conclusions about these results from quantitative charts data alone.

In order to use qualitative analysis to examine more deeply the ways in which English-speaking listeners receive these two bands, it is helpful to employ the more precise technique of discourse analysis, specifically by examining the word-frequencies across the corpus of YouTube comments on these charting songs. By doing so, it will be possible to perform some manual topic modelling of the two corpora by observing which tokens occur with the greatest frequency in the respective comments sections.



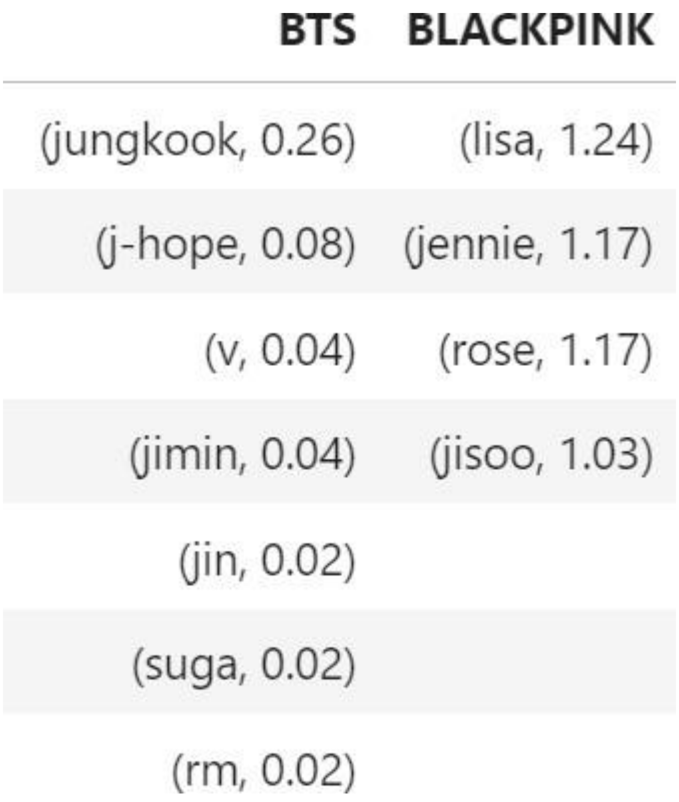
BTS	BLACKPINK
(love, 5.5)	(love, 5.27)
(bts, 3.52)	(blackpink, 3.46)
(song, 2.42)	(like, 2.28)
(💜, 1.96)	(❤️, 2.23)
(❤️, 1.66)	(song, 2.14)
(😍, 1.49)	(😍, 1.66)
(army, 1.45)	(lisa, 1.24)
(like, 1.41)	(jennie, 1.17)
(good, 1.05)	(rose, 1.14)
(music, 0.9)	(jisoo, 1.03)
(video, 0.89)	(queen, 1.02)
(much, 0.87)	(best, 0.86)
(😭, 0.83)	(good, 0.82)
(best, 0.74)	(blink, 0.79)
(make, 0.72)	(girl, 0.73)
(beautiful, 0.71)	(🖤, 0.71)
(cry, 0.65)	(pink, 0.7)
(go, 0.65)	(black, 0.64)
(god, 0.64)	(music, 0.63)
(boy, 0.62)	(😘, 0.63)
(really, 0.61)	(💖, 0.62)



BTS	BLACKPINK
(bts, 3.52)	(blackpink, 3.46)
(💜, 1.96)	(lisa, 1.24)
(army, 1.45)	(jennie, 1.17)
(video, 0.89)	(rose, 1.14)
(much, 0.87)	(jisoo, 1.03)
(😭, 0.83)	(queen, 1.02)
(make, 0.72)	(blink, 0.79)
(beautiful, 0.71)	(girl, 0.73)
(cry, 0.65)	(🖤, 0.71)
(go, 0.65)	(pink, 0.7)
(god, 0.64)	(black, 0.64)
(boy, 0.62)	(😘, 0.63)
(really, 0.61)	(💖, 0.62)
(guy, 0.53)	(💗, 0.56)
(life, 0.52)	(really, 0.55)
(get, 0.52)	(kill, 0.54)
(cannot, 0.51)	(🔥, 0.5)
(us, 0.5)	(💕, 0.48)
(proud, 0.44)	(one, 0.48)
(one, 0.44)	(get, 0.42)
(keep, 0.42)	(say, 0.4)

Figure 3.4: Frequency tables of the BLACKPINK and BTS comments corpus. Frequency is expressed as a percentage of the total words in corpus. Right hand table contains only unique words from each corpus.

Figure 3.4 displays two frequency tables of the most common 20 terms in the opinion-filtered corpus as a percentage of the total terms. The data for this table was pre-processed using the same script outlined in chapter 2. The right-hand table displays data from which terms that occur in both datasets over a threshold of 0.5% have been removed, or ‘stopped’, so at to identify with more precision the terms that occur specifically to either group. The first conclusion that we can draw from the stopped table is that, much like with the initial charts data, listeners tend to react to these two artists in a similar fashion with similar vocabulary. Commenters often express how much they “love” or “like” the song or artist, with the addition of emojis so as to reinforce this response aesthetic response. It should be noted that similar emojis such as a different coloured hearts have not been stopped as there may be socio-semantic differences in how these emojis are used and what meaning they carry. For example, the use of the purple heart within BTS comments has been adopted by the fanbase as an expression of fan identity and has its roots in specific artist-fan interactions (Williams, 2019). It is probable that the black heart found in the BLACKPINK corpus has similar connotations and usages. It should also be mentioned that the same could apply to the pink heart, given the group’s combination of BLACK and PINK to form their name. However, as this is also the default colour for the heart emoji, as evidenced by its frequency within the BTS corpus and thus its stoppage in the right-hand table, it cannot be used as evidence for or against the contention that BLACKPINK fans are just as fandom-focussed in their reception as BTS. Even if the more advanced computation techniques that will be discussed below were employed, it is unlikely that the nuance of such usages would be elucidated.



BTS	BLACKPINK
(jungkook, 0.26)	(lisa, 1.24)
(j-hope, 0.08)	(jennie, 1.17)
(v, 0.04)	(rose, 1.17)
(jimin, 0.04)	(jisoo, 1.03)
(jin, 0.02)	
(suga, 0.02)	
(rm, 0.02)	

Figure 3.5: Table depicting the frequency with which commenters use the band members names ion the respective corpuses. Frequency is expressed as a percentage of total words in corpus

The stopped table displays a clear trend within the BLACKPINK corpus; listeners who comment on these songs tend to do so with a focus on the four singers of the group: Lisa, Jennie, Rosé, and Jisoo. Additionally, as can be inferred from the data in figure 3.5, the singers of BLACKPINK are mentioned at a much higher rate than those of BTS. The lead vocalist of BTS and most talked about of the septet, Jungkook, is still only mentioned in comments roughly four times less than the least-mentioned member of BLACKPINK. Also, the remaining 6 members' names make-up less than a thousandth of the total corpus' terms each. The question that now arises is whether there is a link between the proclivity amongst BLACKPINK users to discuss the singers and the concept of fandom.

There is a convention amongst K-pop fanbases to coin a moniker referring to its collective members (Latestly, 2020). Examining the frequency with which commenters mention this name for the collective fanbase, Blinks and Army for BLACKPINK and BTS respectively, is telling. The term 'Army' occurs just under twice as often as 'Blinks' (1.45% of the respective corpus as opposed to 0.79%)

which shows that, in contrast to BLACKPINK fans who tend to express opinions about the individual singers, BTS fans are much more community focussed. This idea is further supported by the fact that the aforementioned purple heart is the second most mentioned term in the stopped list. This implies that, for BTS, the principal manner of reception in English-speaking countries is through the lens of fandom. However, by drawing Elfving-Hwang's observations about Korean fans' tendency to fixate on a particular idol (2018), I contend that the commenters' disposition towards discussing the four singers of BLACKPINK is also an expression of fan identity. A potential avenue for exploring this idea computationally would be through topic modelling such as that employed in chapter 5, however given the scope of this project and the fact that it is principally foregrounded in the Humanities, this case study will instead do so through a close musicological reading of specific songs. There are limitations to this method of analysis, however. By its very nature, musicological analysis ignores a core feature of K-pop which is that it blurs the lines between an audio and visual product. Fashion and dance are as much a part of a BLACKPINK or BTS song as the music and vocals (Rousse-Marquet, 2012). Additionally, without more advanced computational techniques, it is impossible to claim with absolute certainty whether those who are talking about Rosé are doing so with reference to her vocal performance, her dance moves, her outfit, or another non-musical feature entirely. Thus, as the comments corpus was downloaded from music videos is also worth considering the effect which visual aspects of the music videos will have upon the experience of the commenters. Given that 'video' is one of the most frequent tokens in the BTS corpus, it is undeniable that this is a dimension that commenters are expressing opinions about. Therefore, it would be naïve to assume that BLACKPINK fans are not taking this visual dimension into account when expressing their opinions about the singers. This is a factor that will have to be continuously considered and evaluated in the following section.

2) Song analysis: methods

This section seeks to examine the extent to which BLACKPINK commenter's focus on the singers of the band is a feature of fan expression. To this end, this case study will attempt to gauge whether there is a musicological difference between BTS and BLACKPINK songs that would lead to such a divergence in focus within the comments. In order to answer this question, three methods of analysis will be employed:

1. A statistical analysis, to investigate distributions of vocal features across the wider corpus of BTS and BLACKPINK songs.
2. A comparative schematic analysis, to present and elucidate differences and similarities in pertinent musicological features of specific songs by BLACKPINK and BTS.
3. An analysis of sung-time by each singer to support this schematic analysis, visualised through a colour-coded waveform.

The features that will be examined are style, diatonicism, language-use, vocal range, and syncopation all with an aim toward finding an explanation for particular distributions of singer-tokens. As previously mentioned, it could be argued that the answer to this question lies with YouTube music videos being audio-visual texts that, especially in the case of K-pop, incorporate visual messages such as dance and body language into the viewer's experience. However, with the use of these three analyses it will be possible to ascertain the degree to which the music itself influences singer-token distributions.

3) Song analysis: findings and discussion

3.1) Statistical analysis

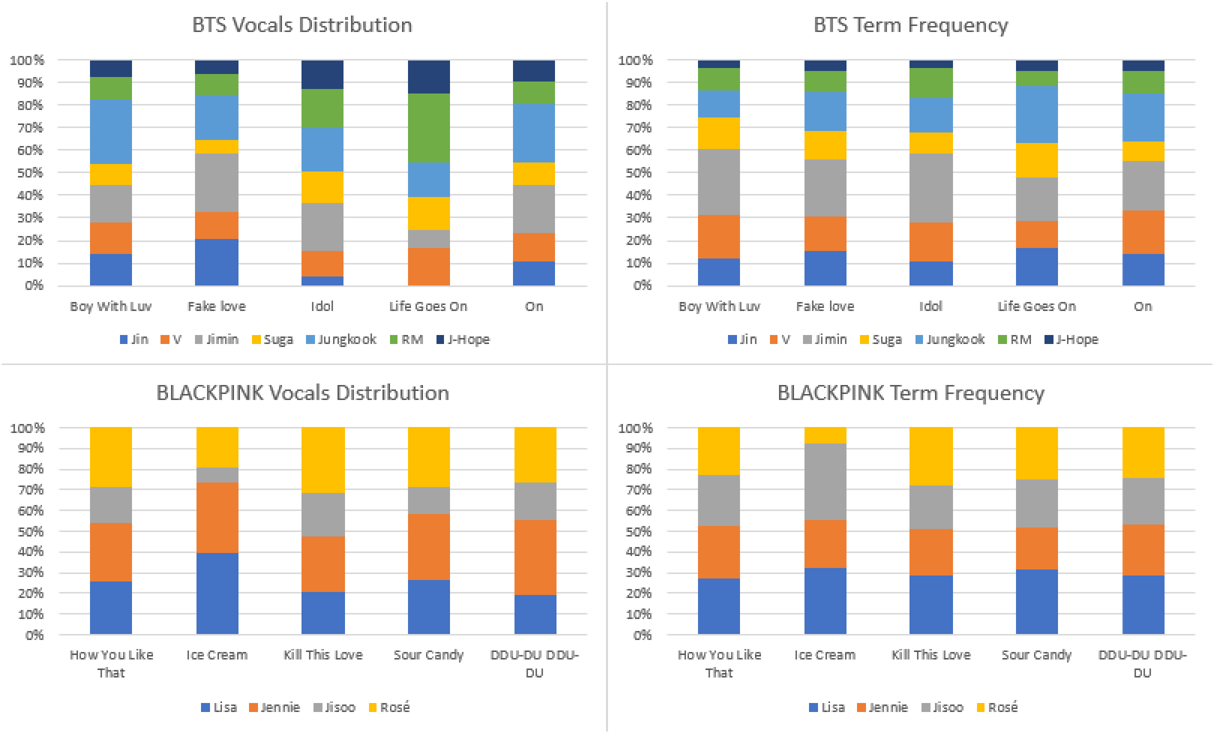


Figure 3.6: Composite bar charts for BTS (top) and BLACKPINK (bottom) visualising the percentage of vocals in a song sung by each member (left) and term-frequencies of tokens belonging to those artists (right)

A statistical analysis is helpful as it allows us to form generalisation for the broader corpus. To this end, the 5 most popular songs based on chart position and/or English language engagement on YouTube were selected for analysis. Figure 3.6 visualises data pertaining to the songs comparatively as a composite bar chart. The charts on the left display how much time each vocalist sings out of the total sung-time of the song, whereas the right-hand charts are the previously referred to singer-token frequencies for each song. It should be noted that the sung-time only refers to instances where the vocalist is singing by themselves as opposed to as backing vocals or when the whole group sings together. From this data, we can draw one principal conclusion; the distribution of how much of the vocals in a BLACKPINK song that a singer occupies does not influence the frequency with which commenters discuss that singer. We make this claim, for two reasons:

1. Although initial uniformity between the datasets for BLACKPINK imply correlation, analyses of specific singers demonstrate that this is incidental.
2. Furthermore, comparison with the data pertaining to BTS highlights that it is possible to predict distributions of singer-name tokens on the basis of how much time in the song that singer occupies vocally; for BTS songs, there is a correlation between sung-time and singer-token frequencies, but not for BLACKPINK.

From cursory analysis, it would seem that term-frequency and vocal distribution are at least superficially correlated; in ‘Kill This Love’ and ‘How You Like That’, for example, the difference in term-frequency and vocal distribution for each singer does not vary more than 8.5%. However, by examining broader trends, specifically for the singer Jisoo, it becomes clear that this superficial correlation is incidental. The most notable example of this is in the song ‘Ice Cream’, in which, despite only occupying 7% of the sung time in the song, Jisoo is the most discussed of the singers as she accounts for 32% of all singer tokens. Additionally, despite the fact that Jisoo sings for consistently less time overall within the songs, this is in no way reflected by the term frequencies; other than the aforementioned song, commenters tend to mention Jisoo with approximately the same frequency as the other singers with a maximum variation of just under 9% between all the singers. The inverse is also true; Jennie tends to occupy the most time vocally in the songs but does not do so with term frequencies.

However, when the BLACKPINK results are compared to those of BTS, it is clear that for the latter there is a strong correlation between vocal and term-frequency distributions. When the data from ‘Life Goes On’ are excluded as around 50% of the vocals are sung by multiple band members at once and thus skews the results, the distributions between the term-frequency and sung-time do not differ more than 4.8% percent on average. Thus, across the five songs that were analysed statistically, there was a strong correlation between the two least and most mentioned band members for each metric, namely J-Hope and Suga, and Jimin, and Jungkook respectively. It is thus undeniable that, for BTS songs, there is a link between the term-frequency of a given singer's token within the YouTube comments of a song and how much time they occupy vocally in the song.

We can conclude that for BLACKPINK, there must be an alternative explanation as to why there is such uniformity between singer-token distributions; the results for BTS demonstrate that YouTube commenters are prone to being influenced by the amount of time a singer occupies in a song and it is probable that if BLACKPINK singers were mentioned with the same frequency of the BTS members that the distribution would follow a similar pattern. However, as previously mentioned, BLACKPINK members are discussed up to 50 times more frequently than BTS members and thus there is clearly an additional feature of these songs - be it audio-visual, socio-cultural, musical, or a combination of the three - that not only causes fans to discuss the singers at roughly the same comparative frequency as each other but also at an increased rate. With this in mind, the following schematic analysis will explore the differences between specific songs musically so as to ascertain to what extent this dimension influences the singer-token distribution.

3.2) Schematic Analysis

Whilst BLACKPINK songs broadly display much greater levels of uniformity between the singer-token distributions, there are a number of songs whose corpus of comments varies in how much they mention the singers. By comparing the vocal lines in two songs, one which displays very little variation in singer-token distributions and another for which there are noticeable differences in how often commenters mention the vocalists, we will be able to ascertain which factors, if any, contribute towards one singer being favoured over another in the comments of that song.

Figure 3.7 shows how often singer-tokens occur in BLACKPINK songs relative to one another as a percentage. It is important note that these percentages are based off the real number of token instances and not a percentage of the total corpus as in figures 3.5 and 3.6. This is because, if a percentage is used, popular songs can seem to display a more uniform distribution of singer-tokens. The reason for this is self-evident; popular songs garner more views and therefore comments from listeners who have never encountered the band before and as a result would not know the names of the singers. Thus, the percentage of the corpus occupied by singer-tokens is lower and the differences in distributions between them is less pronounced.

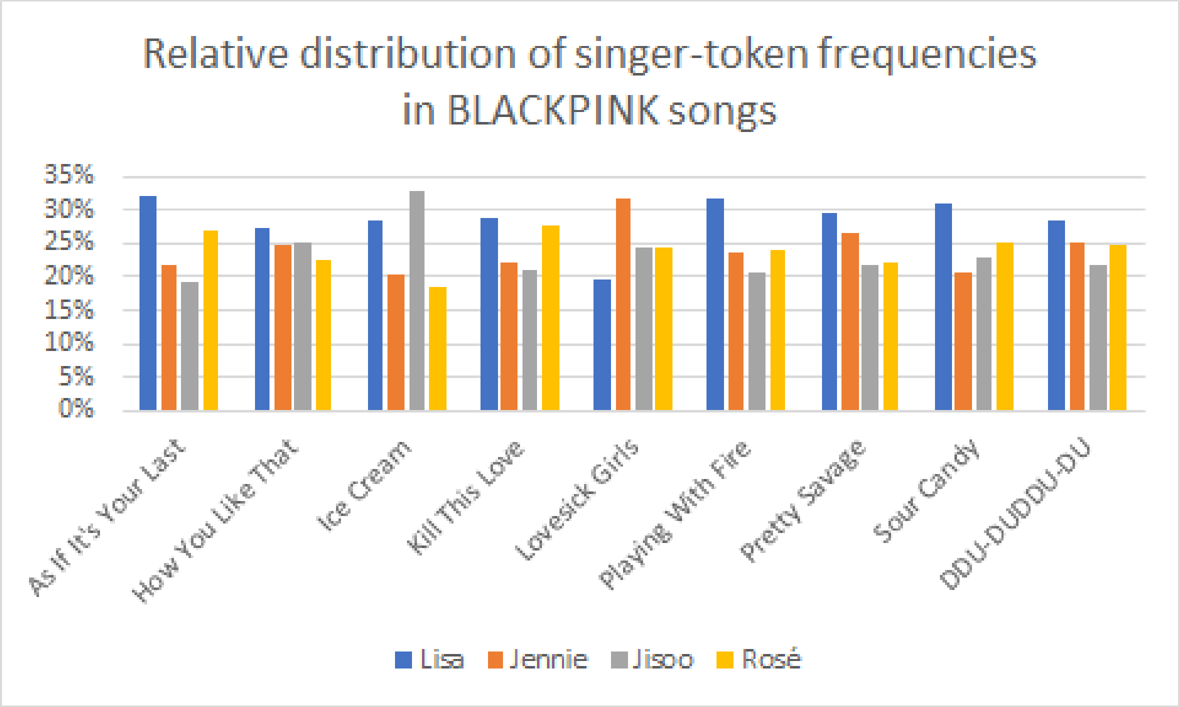


Figure 3.7: Bar graph displaying comparatively the relative percentages of singer-tokens on BLACKPINK songs

It is clear from the visualisation that 'How You Like That' has the most uniform spread of tokens, but analysis proved that the values for both 'Ice Cream' and 'As If It's Your Last' varied greatly with around 8% and 7% average variation respectively. In spite of the fact that 'Ice Cream' displays slightly more variation than 'As If It's Your Last', the latter will be chosen to be a subject for this schematic analysis due to the fact that in 'Ice Cream' the band member Jisoo only sung for an anomalous 6.2 seconds, which will render it more difficult to draw meaningful, comparative conclusions between her and the other singers. Additionally, 'Ice Cream' heavily features Selena Gomez (who is not part of the band) during the chorus and thus provides even less material with which to contrast vocal melodies and features.

Table 3.2: Schematic presentation of the vocal lines in 'How You Like That'

Feature	Lisa	Jennie	Jisoo	Rosé
Style	Mostly rap with some spoken word	Rapped refrain in chorus	Sung melody	Sung melody
Language	Mix of Korean and English	Majority English	Majority Korean	Majority Korean
Rhythmic variation	n/a	Frequent syncopation	Frequent syncopation	Frequent syncopation
Diatonicism	n/a	Occasional non-diatonic notes	Occasional non-diatonic notes	Very few non-diatonic notes
Solo section	Y	N	N	Y
Vocal Range	n/a	Lowest/highest note: B5/B6	Lowest/highest note: B5/B6	Lowest/highest note: B5/B6

Table 3.3: Schematic presentation of the vocal lines in 'As If It's Your Last'

Feature	Lisa	Jennie	Jisoo	Rosé
Style	Mostly rap	Rapped refrain in chorus	Sung melody	Sung melody
Language	Korean	Infrequent use of English	Majority Korean	Majority Korean
Rhythmic variation	n/a	Some use of syncopation	Some use of syncopation	Some use of syncopation
Diatonicism	n/a	Some use of non-diatonic notes	Some use of non-diatonic notes	Very few non-diatonic notes
Solo section	Y	N	Y	Y
Vocal Range	Lowest/highest note: E^{b}3/D^{b}4	Lowest/highest note: C4/D^{b}5	Lowest/highest note: C4/C5	Lowest/highest note: C4/D^{b}5

Tables 3.2 and 3.3 display schematically the results of musical analyses for 'How You Like That' and 'As if It's Your Last' respectively. By comparing figure 3.7 to these tables, the immediate conclusion that we can draw is that there is a strong correlation between Lisa contributing the most unique section stylistically to the songs and being the highest frequency singer-token. With the exception of a rapped refrain by Jennie in 'How You Like That', none of the other members of the band have any rap or spoken sections in the songs. Furthermore, when Lisa does sing, she does so in a range much lower than the other vocalists, thus lending her a more distinct vocal sound. Additionally, the data

pertaining to Jisoo and Jennie follow a similar pattern when comparing the schematic analyses to figure 3.7. To expand, figure 3.7 demonstrates that distributions of tokens belonging to Jennie and Jisoo closely mirror each other. Likewise, in both songs the vocal lines of the two singers contain non-diatonic notes and syncopation at roughly the same frequency. Additionally, the schematic analysis for 'How You Like That' with reference to these two singers provides a useful insight; the usage of English does not have any bearing on the fans' preference for one singer over another. If it were the case, we would expect some variation between the distributions for Jennie and Jisoo. Finally, Rosé is separated from Lisa and the Jennie-Jisoo grouping in figure 3.7 for the two songs, and this is also reflected in the schematic analysis. Rosé does not rap in these songs; thus, her vocal lines do not resemble those of Lisa. Furthermore, her sung melodies differ from those of Jisoo and Jennie with regards to diatonicism. Thus, it follows that her singer-token distributions would differ from Jennie/Jisoo and Lisa as she presents a unique sound within the context of the other vocalists in the band.

Whilst the variables examined in the schematic analysis above are able to predict groupings (or lack thereof) in singer-token distributions, Lisa is the only singer who has vocal lines with musicological features that can directly lead to her being mentioned at a higher frequency than the other members of the group. The fact that Lisa presents a stylistically unique sound when she sings correlates directly with commenters expressing an opinion towards her. Furthermore, there is evidence that the musicological feature of diatonicism does not have any bearing on how often an artist is discussed in the comments. In the comments for 'How You Like That', Rosé is mentioned less than Jisoo and Jennie and her vocal lines contain fewer non-diatonic notes than those two singers. This is the pattern that one would expect as non-diatonic notes typically make a melody present a richer sonic variation. My emotional, aesthetic response to this is I find the melody to be more interesting. This is undoubtedly due to my musical training and tastes being influenced by my white, British, anglophone background. However, it is possible that a large majority of the commenters will also share some of these demographic characteristics, and thus we might be able to apply this personal response to the wider corpus. However, in spite of singing no non-diatonic notes in 'As If It's Your Last', Rosé-tokens appear with a much greater frequency than those of Jisoo and Jennie.

This schematic analysis exposes a core feature of these two songs - the reason why the sung ranges and the usages or syncopation are so similar between the singers is because there is a tendency for different members of the group to share sections of the songs and the repeated melodies thereof, hence the addition of 'solo section' as a metric into tables 3.2 and 3.3. Additionally, as with 'As If It's Your Last', whilst Rosé may sing pre-chorus 1, Jennie and Jisoo will share pre-chorus 2 and repeat the melodies and rhythms of pre-chorus 1; this means that looking at the songs schematically in a single dimension is an imperfect method. Figures 3.8 and 3.9 are waveform diagrams of 'How You Like That' and 'As If It's Your Last' respectively created with the use of Sonic Visualiser (Cannam et al., 2010). Each soundwave has been segmented into part of song (e.g., verse, chorus) and also by singer. By presenting the songs schematically in this manner, we are presented with a potential explanation for the even and uneven distributions of 'How You Like That' and 'As If It's Your Last' respectively. In figure 3.8, there is a higher level of incongruence between sections of the song with regards to how much and which parts of a given section one of the vocalists sings. For example, in chorus 2 the singer changes rapidly compared to chorus 1. Additionally, when comparing the songs in figures 3.8 and 3.9, 'How You Like That' presents a more complicated song in terms of singer distributions and also with regards to the song structure. Not only does 'How You Like That' contain a breakdown wherein multiple vocalists are singing at the same time, it also has a distinct middle 8 sung by two of the vocalists. Furthermore, there is a roughly even distribution of singers in all of the sections in 'As If It's Your Last' and at no point does a vocalist return to sing in a section after they have stopped, which contrasts heavily with 'How You Like That'.

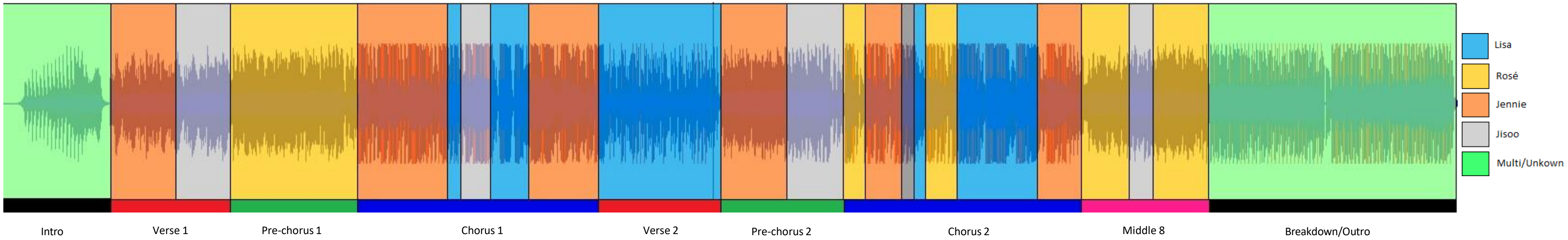


Figure 3.8: Waveform diagram of 'How You Like That' created using Sonic Visualiser segmented by singer and song structure.

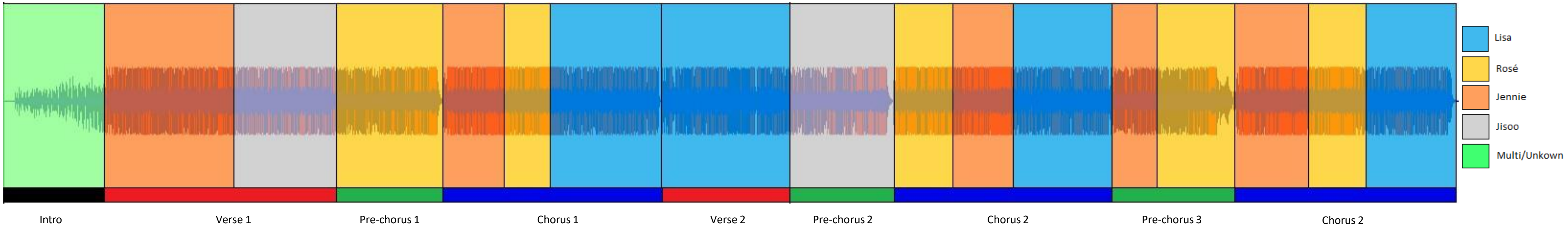


Figure 3.9: Waveform diagram of 'As If It's Your Last' created using Sonic Visualiser segmented by singer and song structure.

The distributions of the vocals in the two songs serve as an explanation for the difference in relative distribution of singer tokens seen in figure 3.7. The more complicated vocal structure of ‘How You Like That’ could mean that listeners/viewers of the song would have less uninterrupted time with which to form an opinion regarding a given singer’s vocal performance. Not only is there a middle 8 and a breakdown that both present unique vocal lines, but during the chorus the main vocal line rapidly switches between the different singers. With less time devoted to focussing on a given singer, commenters may themselves be unable to process specific features of said singer’s vocal contribution to a section. This could be further compounded the fact that, as evidenced by table 3.2 and 3.3, there is a degree of homogeneity between the specific vocal qualities of the singers, in particular Jennie, Jisoo, and Rosé. This could mean that, without obvious visual clues, a rapidly changing main vocalist may impede a commenters ability to identify clearly which of the group is singing at a given moment, thereby contributing towards a more even spread of singer-tokens. By contrast, the much more formulaic and potentially ‘uninteresting’ structure of ‘As If It’s Your Last’ could paradoxically result in more defined opinions towards the vocalists’ parts in the song. For example, the other vocalists’ sections could act as a musical foil for Lisa’s musically and stylistically distinct section, thereby causing it to be remarked upon with a much greater frequency. Simply put, the fact that Rosé raps whilst the others (generally) do not would mean that she would stand out more in the minds of the commenters. As previously mentioned, these musicological techniques necessarily ignore the visual dimension of these videos in their analyses and in the conclusions they draw. However, cursory investigation of this dimension with regards to BLACKPINK music videos when coupled with the above hypothesis can help draw some broad conclusions. K-pop videos are characterised by sharp choreography and fashionable outfits. It thus follows that these are the areas about which commenters concerned with the visuals are going to express their opinions. A typical BLACKPINK video will feature closeups of the vocalist that happens to be singing at a given moment interspliced with wider shots of the whole group performing choreographed dance routines, once again with the singer at the forefront. Given this, we infer that conclusions drawn from the musicological analysis of the structure of the song regarding singer-token frequencies also holds true for, and are supported by, the sartorial and choreographic elements of these songs. However, without either more advanced computational analyses on the corpus of comments or in-depth visual analyses, it is not possible to ascertain with a degree of certainty whether the visual or musical elements hold precedence. The only conclusion that can be drawn is that the structure of the song itself likely influences commenters to remark on both these dimensions.

It is clear that musicological factors such as structure and style of vocals can at least partially influence singer-token frequencies between songs by the same artist. However, there still remains the question as to why singer-tokens appear in much greater quantities within the BLACKPINK corpus compared to that of BTS. By comparing musicological features of BLACKPINK and BTS songs, we can ascertain to what degree Elfving-Hwang's assertions about the singer-centric nature of K-pop fandom hold true for Western anglophone fans of BLACKPINK (2018). If there is no musicological basis for a much higher propensity amongst commenters to focus on the individual band members, then we can claim with a degree of certainty that this idea of fandom is a core part of the way in which K-pop is received in the English-speaking world. In order to investigate this, it is necessary to investigate the degree to which the vocals of individual singers feature more heavily over the other musical elements of the songs. In doing so, distributions of singer-tokens can be compared to elucidate the effect that emphasising individual singers has on their corresponding token frequency. To this end, it is necessary to select BTS and BLACKPINK songs that have attained similar levels of success in the charts. This is because, as previously mentioned, songs by both bands which place higher in the charts also tend to display fewer instances of singer-tokens relative to the corpus size for the obvious reason that first-time listeners/commenters are not going to know the names of the band members. Four songs present average chart positions close enough together for popularity of the song to be considered a controlled variable:

a) Life Goes On (BTS) and How You Like That (BLACKPINK) with average chart positions of 19 and 19.3 respectively

b) Fake Love (BTS) and Lovesick Girls (BLACKPINK) with average chart positions of 35.3 and 38 respectively

Table 3.4 displays the instrumental arrangements of 'Life Goes On' and 'How You Like That', and table 3.5 displays 'Fake Love' and 'Lovesick Girls'. The basis for examining these songs using this particular metric is that a broader instrumental arrangement will necessarily draw listeners' attention away from the main vocalist. It would be thus expected that a song with a simpler arrangements would present an increased level of singer-token distributions amongst the corpus of comments.

Table 3.4: Table displaying the different instrumental elements of 'Life Goes On' and 'How You Like That'.

Life Goes On	How You Like That
Drums	Drums
Guitar	Bass synth
Bass	Melody synth
Layered Synths	Occasional overdubbed vocals
Frequent overdubbed vocals	Occasional backing vocals
Backing Vocals	

Table 3.5: Table displaying the different instrumental elements of 'Fake Love' and Lovesick Girls'.

Fake Love	Lovesick Girls
Drums	Drums
Guitar	Guitar
Bass synth	Bass synth
Multiple layered rhythm synths	Melody synth
Multiple layered melody synths	Occasional overdubbed vocals
Frequent overdubbed vocals	Occasional backing vocals
Backing vocals	

Both Tables 3.4 and 3.5 demonstrate that BTS songs (left column) are more complex in terms of instrument usage. The two BLACKPINK songs examined typically only have a bassline, a melody and drums supporting the vocals. Additionally, the main vocals are only occasionally overdubbed and supported by another band member singing backing vocals. Whilst both BLACKPINK songs present relatively comparable levels of complexity, the same does not fully hold true for the BTS songs. 'Life Goes On' presents a somewhat simplified instrumental arrangement on account of not containing multiple rhythm and melody synths layered over one another. This is not reflected in the distributions of singer-tokens in the corpus of comments, however. The total frequencies of singer-tokens for 'Life Goes On' and 'How You Like That' – from now on designated group A - are 1.04% of the total tokens and 4.11% respectively; for 'Fake love' and 'Lovesick Girls' – group B - they are 1.51% and 4.80% respectively. On the one hand, the fact that the singer-token frequencies for BLACKPINK are higher than those of BTS undermines the applicability of Elfving-Hwang's contentions to the BLACKPINK fanbase (2018). This is because it implies there is a strong basis to the hypothesis that the complexity of instrumental arrangement directly leads to commenters focussing less on vocalists in the group. However, if this were the case, it would be expected that the proportion of singer-tokens between the two songs in group B to be much greater due to the fact that, as previously mentioned, 'Fake Love' is much more instrumentally complex. However, the inverse is true. Commenters for the BLACKPINK song in group A discuss the singers at a rate of just under four times as often as those of its counterpart. In comparison, the frequency of singer-tokens for 'Lovesick Girls' in group B is around three times that of 'Fake Love'. The conclusion that can be drawn here is that, based on the assumption that more complex instrumental arrangements draw focus away from the vocal lines and their respective singers in a song, it is highly likely that there is no musicological basis for BLACKPINK songs to be featuring higher levels of singer-tokens than those of BTS. It is worth reiterating at this point that this method of analysis ignores the visual aspects of these songs. However, given that both these songs feature the fashionable outfits and impressively choreographed dance sequences that Rousse-Marquet states to be so endemic to the K-pop genre (2012), it is highly unlikely that the visuals of the BLACKPINK songs are so central to the overall effect that this would account for three times as many singer-related comments within the corpus. In fact, in the stopped word-frequency tables, the word "video" is the fourth most common amongst BTS commenters, whereas it does not feature in that of BLACKPINK. This could mean that, if the visual dimension was removed from these songs, BLACKPINK singer tokens could appear with even greater frequency compared to those of BTS.

4) Conclusion

The two methods of analysis offer rich data to support the argument that the phenomenon of fandom is a principal means by which English-language K-Pop fans interact with its musical artefacts. We initially hypothesised this as a result of manual surveying of token-frequency tables formed from the opinion-filtered corpus of comments. Two of the three most unique tokens written by BTS commenters referred specifically to the BTS fandom – a purple heart and the self-ascribed name for the collective members of the fandom: “Army”. However, this was not the case for the BLACKPINK frequency table; in this case, the equivalent tokens – a black heart and the term “blinks” – were only the ninth and seventh most frequent amongst comments on the videos. Instead, there was an overwhelming propensity amongst BLACKPINK commenters to discuss the four group members: ‘Jennie’, ‘Jisoo’, ‘Rosé’, and ‘Lisa’. These four singers were mentioned in total roughly four times as often as their counterparts in BTS. Based on work by Elfving-Hwang regarding the tendency of Korean fans within K-pop fandoms to focus on particular band members (2018), our hypothesis was that the habit of English-speaking BLACKPINK fans to discuss the singers so frequently is an example of this phenomenon.

As an alternative to using more advanced computational techniques, I elected to explore this hypothesis by attempting to discern whether there was a musicological basis for this focus on the singers in the band. More specifically, the aim was to examine whether musicological features could explain discrepancies between distributions of individual singer-tokens. This was an appropriate decision, not only because of the technical constraints of the project, but also because qualitative analysis is a necessary step in examining cultural artifacts such as these. For example, if I was unable to find any supporting evidence that song A exhibited an increased frequency of tokens pertaining to singer A due to particularly marked musicological features of her vocal line, it is logical to conclude that it is due to extra-musical factors. The principal confounding factor that prevents us from unequivocally claiming that fandom idolisation of the singers of BLACKPINK is the reason why such a large majority of the comments are devoted to them is the nature of K-pop songs as an audio-visual product. I have hypothesised that, if the visual aspects of K-pop songs had such a great influence we would see a concurrent raised frequency of singer-tokens in the BTS corpus, but we do not. Therefore, it is it likely that this visual dimension does not play an important role. However, it is the place of future research to investigate this matter further.

Building upon this case study, future research might examine the visual elements of the songs in the same way that we analysed the musicological features of the tracks. Instead of features of vocal lines such as diatonicism, syncopation, or range, a researcher could scrutinise outfit choice, choreography, or camerawork and compare these results to the token frequencies. Alternatively, it is possible to delve further into computational methodologies in order to build a fuller picture of the ways in which commenters are discussing these artists. Chapter 4 provides a conceptual framework for grouping comments based off the collocations of keywords identified in the corpus to the end of elucidating language attitudes, whereas chapter 5 presents a clustering methodology based off content and syntactic features. It would be relatively trivial to repurpose these methodologies for the goals of this case study by examining the collocates of the singer-tokens. In doing so, we could obtain much more specific data regarding the content of these comments, be they discussing musicological or visual, or other dimensions of these songs. The main challenge of this would be grouping of the collocates into semantic fields, e.g., ensuring that tokens such as “dance”, “outfit”, “moves”, or “makeup” were all categorised as belonging to the visual dimension. There are a variety of computational solutions for this problem, ranging from unsupervised clustering to dictionary-based methods (Klebanov et al., 2008). With this in mind, I suggest that a computationally driven approach to the aims of this chapter could be fruitful and relatively simple to implement.

We have established that K-pop presents a unique genre for English-language engagement with foreign-language music as it is specifically designed to be accessible for anglophone audiences. It is perhaps for this reason why language attitudes seemingly do not play much part in the way commenters engage with the music. However, this is not the case for all foreign-language music within the anglophone sphere of popular music. Indeed, it would be negligent for this project to examine the reception of foreign-language music without examining the role that language attitudes have to play in the process. To do so, we now move on to a very different genre stylistically and socio-culturally: reggaeton.

Chapter 4: Exploring the attitudes towards language choice in reggaeton songs.

1) Introduction: the success of reggaeton and the 'Despacito Effect'

In 2006, Wayne Marshall remarked on a new wave of Latin music sweeping the United States. He noted that, whilst in previous years use of the Spanish language on the radio had been relegated to hits such as José Feliciano's 'Feliz Navidad' (1970) and Ricky Martin's 'La Vida Loca' (1999) (both of which contain snippets of Spanish), at the time of writing, "15 to 30 minutes blocks of Spanish-language pop" had begun to appear on hip-hop stations (2006). This new genre of music was reggaeton – a style of music originating from Puerto Rico and Panama and that had managed to reach international success, in spite of efforts from the Puerto Rican government to censor and suppress it (Duany, 2010; Ilich, 2018). Marshall describes how, most importantly, this genre of music lacked the "exotic veneer" that was present in previous Spanish-language music, which found popularity in the US market; this was music that was written and performed for Spanish-speaking American audiences (2006). Ilich states that modern reggaeton music is a fusion of "electronic dance music, hip-hop elements and Spanish/Spanglish rap" which has its roots in reggae beats mixed with traditional Latin music such as salsa and bomba (2018). It is important to acknowledge that reggaeton is just one of the many sub-genres of popular music that exist in Latin America. However, given the success of 'Despacito' by Luis Fonsi and other hits such as 'Mi Gente' by J. Balvin, reggaeton and its derivatives such as the trap genre have been propelled to the forefront in representing Latin American music on the global stage, and in doing so pushing other genres to the side. This phenomenon is principally aggravated in English-speaking markets by the propensity of anglophone markets to group Spanish music under the umbrella of reggaeton even if it belongs to a distinctly different genre (Leight, 2018). With this in mind, this case study will limit its scope to songs belonging to reggaeton or related sub-genres. In practical terms, this means that music such as that by Enrique Iglesias will not be "grouped under the umbrella" of reggaeton for the sole reason that it makes use of the Spanish language (Leight, 2018). Instead, research on Google or, when results there are inconclusive or questionable, qualitive analysis using the definition proposed by Ilich will be used in order to classify the Spanish-language songs in our corpus (2018). In total, eighty-nine reggaeton songs were identified to have appeared in the charts of our six anglophone countries. For a full list, see appendix 1.

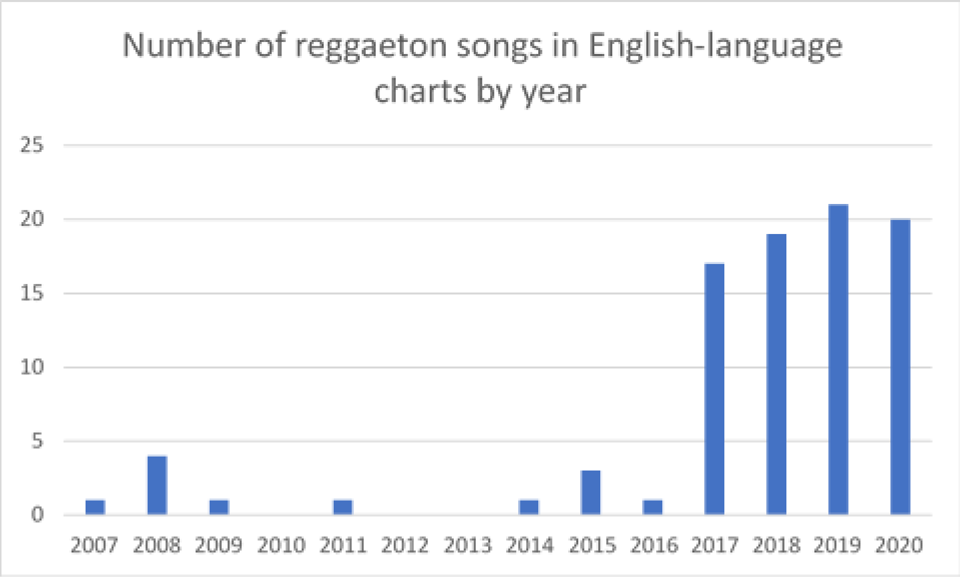


Figure 4.1: Bar chart visualising how many reggaeton songs charted in 6 English-speaking countries from 2007

Arbona-Ruiz remarks that perhaps the most famous reggaeton song is 'Despacito' (2017), originally by Daddy Yankee and Luis Fonsi but also released as a remix featuring Justin Bieber. She continue that this song has been credited as demonstrating to the world that Latin artists were starting to have a profound influence over the American music industry, an influence that is reflected in the data collected throughout this project. Figure 4.1 offers a visual representation of the number of reggaeton songs charting each year, starting from 2007. It is clear from the data presented here that 2017, the year in which 'Despacito' released, was a turning point for the reception of reggaeton songs. In 2017 alone, more reggaeton songs reached the charts than all previous years combined and this trend of success for these Spanish language tracks has continued thus far; it remains to be seen, however, whether the slight downturn in popularity in 2020 is the result of random fluctuation or whether the 'Despacito effect' is wearing off. Regardless, as the wave of reggaeton sweeping the world represents a (practically) unprecedented acceptance of mainstream foreign language music. Thus, any project on foreign language music in 2021 would be necessarily incomplete without a focus on the "Latin renaissance" and the language attitudes thereof (Acevedo, 2019).

2) Aims of the case study and theoretical backdrop.

Whilst English is commonly used in reggaeton songs, unlike K-pop, artists do not do so with the specific goal of international appeal through the use of English (Rousse-Marquet, 2012). In reality, as of 2017 popular reggaeton artists frequently receive offers to collaborate from big mainstream anglophone artists such as Post Malone and Flo-Rida (Arbona-Ruiz, 2017). This, coupled with the fact that at present reggaeton represents the most popular foreign language genre in terms of engagement on YouTube and presence in music charts, means that it is a rich resource for the study of language attitudes. As mentioned in chapter 4, limitations of the project bar the use of advanced computational techniques in order to illuminate the topics discussed by commenters. However, through the use of lower-resolution techniques such as keyword collocations, it will be possible to uncover the opinions of commenters on a specific, manually selected topic – language use in reggaeton songs.

2.1) Language attitude theoretical framework and hypotheses

Dragojevic et al. define language attitudes as "evaluative reactions to language" that can be divided into two categories: a) reactions toward different language varieties and b) reactions toward speakers of these different varieties (2021, pg. 61). There is also often a corollary relationship between these two strands; for example, opinions towards the structure of a language will correlate with value judgements about the speaker. Dragojevic et al. identify five strands of language attitude research; research concerned with the documentation, development, explanation, consequences, or changing of language attitudes (2021). This case study will concern itself primarily with the documentation of language attitude toward reggaeton songs. However, research from other strands (e.g. change and development) will be incorporated in order to specifically examine how these attitudes evolve over time through a comparison of pre- and post-2017 comments. It is worth noting that Dragojevic et al. wrote their article to be used as a springboard from which future researchers with diverse methodological skill sets will be able to collaborate in order to further common research goals. This multimodality is in line with ethos of this project and the wider goals of the Digital Humanities (Honn, 2014), and as a result is a particularly apt theoretical framework within which to situate this case study.

Dragojevic et al. contend that the collective body of research documenting language attitudes demonstrates that they hinge on a “hierarchy of prestige” (2021, pg. 63), that is to say, the attitudes are entrenched in the power relations of the societies in which they are formed. This initial contention proves problematic for the type of corpus utilised in this case study as it is multinational and multicultural by nature and with very little way to determine the particular national/cultural background of a given element of the corpus. Whilst I have endeavoured to limit the corpus to western anglophone audiences, doing so based solely on usage of English poses problems with regards to *a priori* assumptions about who can be an English-speaker (see chapter 1). One way to clear up ambiguities regarding the socio-cultural background of the commenters is to make use of the charts data collected during the initial stages of the project. If a song charts better in country A versus country B, it logically follows that the majority of the commenters on the video will be from country A. Determining what qualifies as ‘charting better’ is a question that poses some methodological challenges and will be addressed in section 3.

As outlined by Ivković (2013), a corpus-based approach necessarily ignores lexical features such as accent and other non-linguistic markers; this is one drawback of the methodology used here, as controlling for such markers provides valuable insights into language attitudes. In spite of this, YouTube remains an important space in which dominant linguistic ideologies are contested and where users “metadiscursively evaluate linguistic intelligibility, acceptability, and authenticity” (Aslan and Vásquez, 2018, pg 413; Walters and Chun, 2011). With this mind, it is still appropriate to make use of a corpus of YouTube comments to this end, in order to examine language attitudes in responses to translanguage media, particularly in the online space, where many features of linguistic and cultural identity are hidden behind text-based communication.

As previously stated, the focus of this case study will be the documentation of how language attitudes toward Spanish-language reggaeton songs have changed as a result of the so-called ‘Despacito Effect’ (Arbona-Ruiz, 2017), i.e., the rapid increase in the popularity of reggaeton music in 2017. Dragojevic et al., contend that the socio-economic groups with which a language variety is associated directly affects the “hierarchy of prestige” in which that variety finds itself (2021, pg. 63) This is influenced by media portrayals that play a role in the creation and dissemination of language attitudes towards speakers of nonstandard language to a region (Lippi-Green, 2012). To expand, overrepresentation of standard speakers and subsequent underrepresentation of nonstandard speakers can undermine the societal status of the latter, thereby placing it lower in the prestige hierarchy. However, in figure 4.1 we can see that, as of 2017, the representation of Spanish has increased significantly. Thus, with this increase of representation it follows that there is an increase in prestige from the point of view of the consumers of the media. The hypothesis I propose is that a direct result of the ‘Despacito effect’ is an increase in positive reactions toward use of the Spanish language.

2.2) ‘Othering’ and language attitudes

Since its conception by De Beauvoir in 1949, the concept of the ‘other’ and the process of ‘othering’ have taken root in a wide range of disciplines (Brons, 2015). Othering is a type of social representation “through which identities are set up in an unequal relationship” and that ignores the complexity and subjective experience of the target (Crang, 2013; Dervin, 2012). To expand, the self or an ‘in-group’ will set up an unequal opposition between themselves and the other/out-group on the basis of a desirable characteristic (Brons, 2015). The process of othering has been documented to occur towards foreign language users by those whose native tongue has a higher social prestige (Zabus, 1990). Additionally, González-Cruz posits that, because othering is entrenched in concepts of identity and the self and language forms an integral part of these phenomena, the ideas of othering and otherness are integral to exploring language attitudes (2020). Finally Yoon et al. demonstrate that the process of othering can occur specifically when audiences consume music in a foreign language on the basis of the language used by the target artists (2020). This demonstrates that ideas of othering are an apt lens through which to view the term frequencies that will be produced in section 3 when attempting to elucidate the language attitudes that are displayed therein.

One challenge of examining othering language attitudes in corpora is that, as Crang states, the dynamic between the in- and out-group is often left implicit (2013). Implicitness is a common problem that researchers in natural language processing must overcome, often with the use of specialised methodologies that involve advanced computational techniques (Davidov, Tsur and Rappoport, 2010). As will be discussed later, the approach taken in this case study will involve a computational pipeline in order to locate clusters of opinion in the corpus. However, the meanings behind these opinions will ultimately be decided by the human researcher through term frequency analysis such as that undertaken in chapter 3.

3) Previous corpus-based work

According to Baker et al., making use of corpora and corpus linguistics techniques for discourse analysis is a practice that dates back more than 20 years from today (2008). Nonetheless, at the time of writing Baker et al. remarked that the use of such techniques in critical approaches to discourse analysis was increasing. This is a trend that has continued up to the present day, with there being no lack of studies that not only make use of corpora and corpus linguistics in order to carry out discourse analysis but do so with the specific goal of examining language attitudes and ideologies. Writing in 2007, Leeman and Martínez represent some of the earlier efforts of corpus-based analysis of language ideologies. Despite working from a relatively large text corpus, they do not describe employing any computational corpus linguistics techniques, instead opting for a more qualitative analysis (Leeman and Martínez, 2007). Graedler, however, investigates language attitudes in Norwegian newspapers through the use of more computational techniques (2014). They apply the program Wordsmith Tools 6 to the end of generating a variety of statistical information about the corpus, including keyword frequency lists, collocates, and clusters of topics. In doing so, Graedler draws conclusions regarding the attitudes towards the English language in Norwegian media. Graedler is not the only scholar to employ keyword collocates in order to identify language attitudes. Fitzsimmons-Doolan examine collocates around what they term 'node words' in order to elucidate language ideologies present in a policy corpus (2014). More recently, Liu and Gao gather insights on the representation of the Chinese language in Irish media by inspecting the collocates of the word 'Chinese' in a corpus of 85 newspaper articles (Liu and Gao, 2020). Thus, it is clear that a tried and tested method of elucidating language attitudes is to examine the types of language that accompany key words related to languages.

Aslan and Vásquez remark that there is little in the way of studies exploring language attitudes on YouTube (2018). Some of the few authors to do so are Chun, Ivković, and Walters and Chun (2013; 2013; 2011). As mentioned in chapter 1, Ivković is one of the few authors to make use of a corpus of YouTube comments for musicological analysis to the end of examining language ideologies (2013). As with many of the authors already discussed in this chapter, Ivković does so through the use of keywords, which they refer to as 'conceptual lemma' due to the fact that they are examining comments in a variety of languages. Much like Aslan and Vásquez and unlike many of the non-YouTube based corpus analyses outlined above, Ivković employs qualitative techniques when evaluating the keyword collocates with the use of three human evaluators. It is possible that the reason these scholars have opted for qualitative methods of linguistic analyses is due to the general short length of YouTube comments. The average length of comment collected in this project was roughly 10 words, and thus it is conceivable that general shortness of lexical item lends itself to human annotation well. However, given the size of our corpus (many millions of individual comments) it is highly impractical to employ human methods of annotation. For this reason, we will forgo the example of previous YouTube scholars and instead focus on the precedent set by other discourse analysts and employ highly computational techniques.

3) Methodology

In order to elucidate particular language attitudes through collocations, the first step was to decide the keywords around which term frequencies were to be collected. “Spanish”, “English”, and “language” were the three necessary choices for keywords, however “understand” and “speak” were also included. This is because manual surveying of the corpus revealed that listeners often express opinions about the song with an introductory phrase such as “I don’t understand” or “I don’t speak”, for example “i dont understand it but i love it 😍❤️”. Although not as explicit of a language attitude as a comment such as “Wow i love spanish now thank u for getting me into spanish”, comments such as the former can still be considered an opinion relating to language use. This is simply due to the fact that the authors of such comments have specifically elected to include the fact that the song is in a foreign language to their own as part of their commentary on the music (or extra-musical features). It should be noted that “understand” and “speak” is not an exhaustive list of the possible introductory keywords that users could employ when expressing a language attitude in this manner. However, unlike with “English”, “Spanish”, and “language”, these keywords are not solely language related. Thus, I elected not to include more than two non-language related keywords in order to reduce noise in the data that could produce misleading results vis-à-vis the term-frequency analysis. For example, if other non-language related keywords that could form part of an introductory phrase to a language attitude (e.g., “say” for “I don’t know what he is saying but…”) are introduced into the analysis, there is an increased possibility that terms unconnected to language attitude may influence the token-frequency analysis at latter stages. Once the keywords were chosen, the corpus of comments for a particular song was iterated through and each string was checked for presence of the keywords. Then, the comment was processed in order to make the natural language more accessible to statistical analyses. Checking for the presence of a keyword in a comment before processing it meant that strings with misspelled keywords would not be identified. However, preliminary tests demonstrated that natural language processing was by far the most computationally demanding step of this pipeline. As a result, by opening up the results to minor statistical skew in this way, run-time for the script was decreased from hours to a matter of minutes. Additionally, in order to further speed up this process, the script was multiprocessed to run concurrently across CPU cores, as this resulted in a 3x speedup.

The script used for text processing was broadly the same as that applied in chapters 2 and 3. The exception was that the following three steps were added: 1) tokens containing the name of an artist were transformed into "artist_name, 2) the collection of ngrams (a sequence of *n* tokens combined into one), and 3) the combining of negatives and the subsequent word into one token, e.g. "not understand" becomes "not_understand". It should be noted that certain adverbs such as "really" and "quite" can split the not + verb structure, however in the majority of cases these adverbs are removed as stop words by default. The instances where this does not occur are when the adverb has a synonym, for example "even". However, given the relative rarity of these occurrences, if they are seen to be impacting the results then the researcher can include logic in the code to specifically screen for this possibility. The reason why ngrams had not been collected up to this point was due to the computational costs associated with it. Hallman writes that "the technique of processing N-grams can become prohibitively computationally expensive as n increases" (2021), however, thanks to the aforementioned efforts to render the script more efficient, I was able to include bigrams and trigrams (ngram of a length of two and three respectively). It was consequently for this reason that I elected to combine negatives and the word immediately after into one token. As previously mentioned, "understand" and "speak" were included for the reason that commenters tended to negate these verbs, combining them with a clause expressing an opinion towards the song, thereby expressing a language attitude - the example given earlier was "i dont understand it but i love it 😍❤”. In this comment, if “not” and “understand” are not combined into one token, then two of the bigrams that will be returned once stop words are removed are “not understand” and “understand love”. As we have seen the original string, it is clear to us from the bigrams that the author is expressing a positive sentiment towards the song despite not being able to speak the language. However, in analysing the comment without this context (such as in a token-frequency table of the corpus) a weakness is exposed: there is no preservation of syntactic elements and it falls on the researcher to perform keyword in context (KWiC) analysis themselves. By combining the negative “not” with “understand” into “not_understand”, this weakness is to a certain extent mitigated. Due to the nature of natural language - especially language that does not conform to generally accepted standards of correctness - there are undoubtedly permutations of comments that will confound this effort to still convey syntactic information. However, the decision to combine these tokens will still provide the opportunity for more insightful analyses.

At the end of the pipeline outlined above, a frequency table of the 100 most common ngrams and single-word tokens was generated. Whilst it was possible to have a frequency table many hundreds of tokens longer, it would have made the manual, human examination of the results much more complicated. Additionally, after this cut-off, the frequency with which tokens appear is significantly reduced and thus only serves as statistical noise. A total of 2031 unique tokens appeared across all eighty-nine frequency tables. These figures alone inform us that there is a large degree of similarity in opinions commenters express towards language use in the selected songs; only around one in five tokens in a given frequency table are unique to that table. In spite of this general uniformity, there still lay value in attempting to elucidate which groupings, if any, existed within this corpus of frequency tokens. This is because, in the words of Dejan Ivković, “dividing the corpus facilitated making potential generalizations through the analysis of [salient features] and dispersion across the sub-corpora” (2013). In order to discern which frequency tables were most similar in their contents, and as a result group them, the method of cosine similarity was employed. This technique functioned by taking the frequency with which each of the 2031 unique tokens appear in a given video's frequency table (which will necessarily be zero if a particular token does not feature in the video's comments) and plotting them as a single vector (a line with both direction and magnitude) in Euclidean (not curved) space. Then, the cosine of the angle between the vector of one frequency table and that of another is calculated. The closer this number is to one, the more similar the two tables are in their contents.

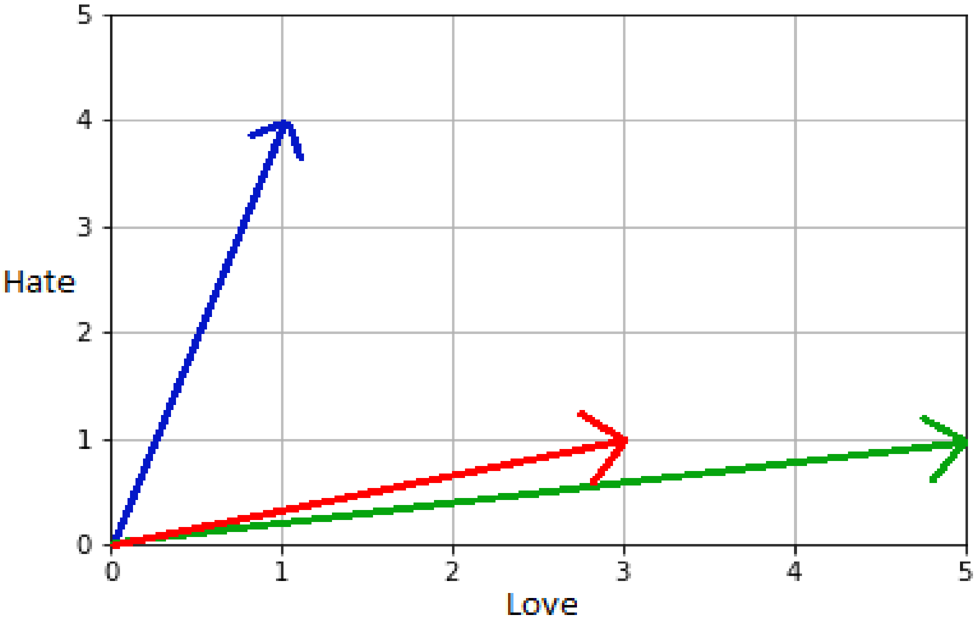


Figure 4.2: A simplified vector graph of token-frequencies

Figure 4.2 visualises a simplified example in which each of the coloured arrows represents a frequency table (red, green, and blue) and the respective values for the occurrence of "love" and "hate" therein. In the blue frequency table, the token "hate" appears four times, while "love" only appears once, whereas in the red table "hate" only appears once and "love" three times. The cosine of the angle between the blue vector and the red vector is given with the following formula:

$$\cos(\theta) = \frac{blue.red}{\|blue\| \times \|red\|} = \frac{\text{dot product[4] of red and blue}}{\text{magnitude[5] of red and blue}} = \frac{4\times1+1\times3}{\sqrt{4^2+1^2}\times\sqrt{1^2+3^2}}$$

[4] Dot product refers to the summation of the product of every *nth* element of two sets of integers of equal length. In this case the sets are vectors.

[5] Magnitude refers to the length of the vector.

This cosine is 0.537 and, as previously stated, the closer this figure is to 1 the more similar the two vectors. As a result, if this formula is applied to the red and green frequency tables, we get a figure of 0.992 for the reason that, as is clear from the visualisation, the vectorised tables are much more similar. Figure 4.3 is a visualisation of the same vectorised frequency tables, but with a third dimension included: the token "Weird" with a frequency of 3, 0, and 5 for blue, green, and red respectively. Due to the extensibility of the formula, this third dimension requires no alterations to the equation. Table 4.1 is a similarity matrix calculated from the information displayed in figure 4.3 and displays the cosine similarity between all three vectors.

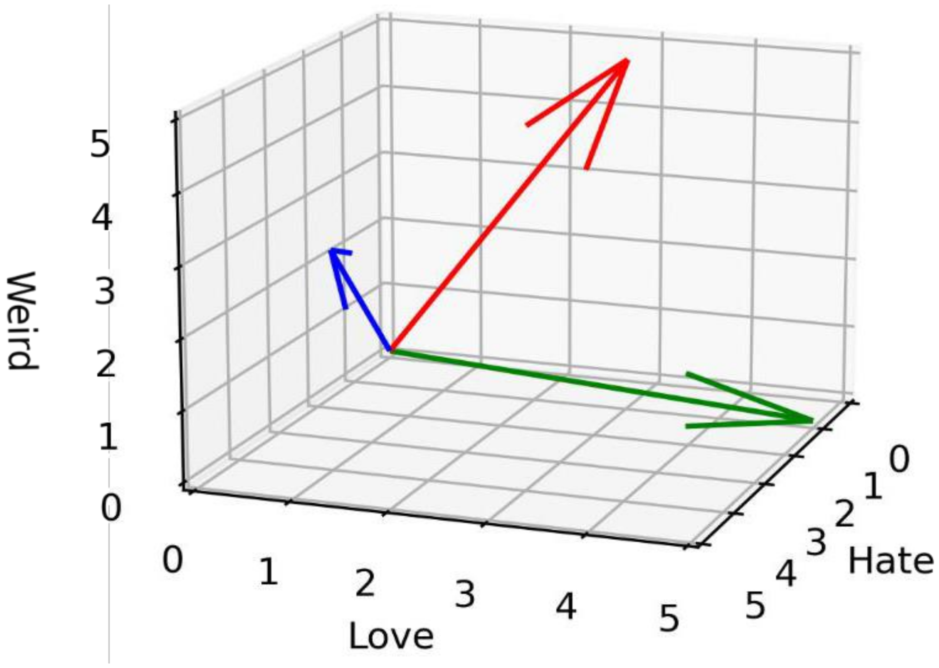


Figure 4.3: Vectorised frequency tables with an added third dimension

Table 4.1: Similarity matrix of the vectors visualised in figure 4.3

	Blue	Green	Red
Blue	1	0.346	0.729
Green	0.346	1	0.530
Red	0.729	0.530	1

The data in figure 4.3 and table 4.1 demonstrate how the introduction of a third dimension to these vectors can have an influence on the similarity between the two texts; as “weird” is not mentioned in the green frequency table, the red’s similarity with green has decreased from 0.992 to 0.530. Additionally, as “weird” appears in both red and blue’s corpora fairly frequently (5 and 3 respectively) the similarity between the two has increased from 0.537 to 0.729.

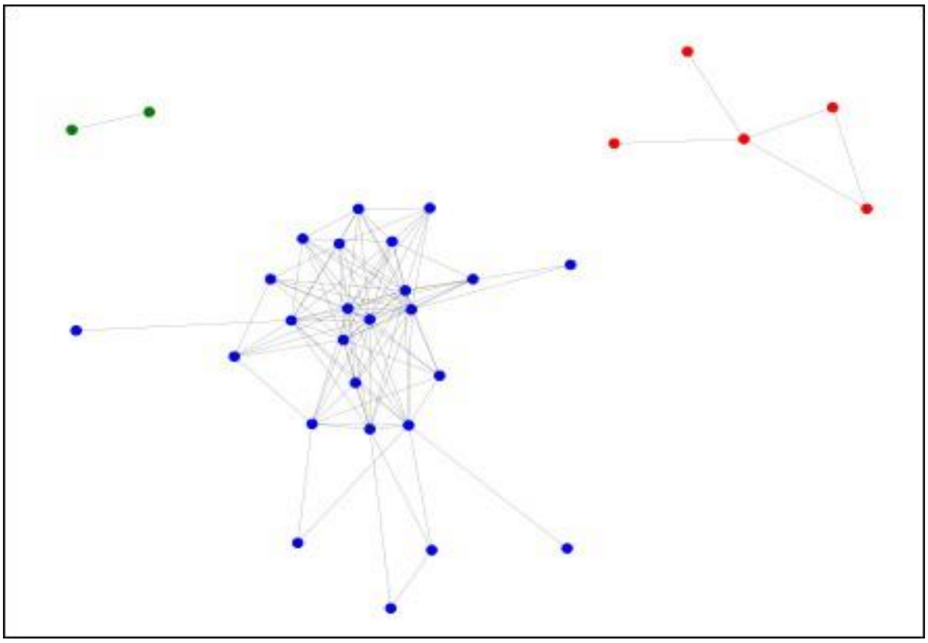


Figure 4.4: Network diagram of cosine similarities between frequency tables.

Due to the aforementioned extensibility of the formula, this measure of similarity can generalise into as many dimensions as necessary, even if we cannot physically visualise what this would look like graphically. This is what permitted the calculation of cosine angles between vectors existing in 2031 dimensions - one for each of the tokens in the total corpus vocabulary. From this, it was trivial to elucidate groupings in the corpus by identifying frequency tables whose cosine similarity with another was above a threshold of 0.925. In doing so, it was possible to create networks of songs who are related in the content of their frequency tables. This network is displayed figure 4.4. Each node in this figure represents a frequency table, and the shorter the link between each node, the higher the cosine similarity between the tables. This network diagram (and all that follow) were created with the Python package NetworkX (Hagberg et al., 2008).

Two conclusions are immediately clear from figure 4.4: a) there is further evidence that all of the frequency tables display a great deal of similarity and b), there are three groups with differing semantic topics broached in the frequency tables. We can infer the former as there are 32 nodes in figure 4.4 – approximately 35% of the eighty-nine songs examined in this manner. As around a third of the songs have a cosine similarity equal to or over 0.925, there must be a principal theme that all commenters discuss when evaluating the songs. This hypothesis is backed up by the fact that lowering the cosine similarity threshold to 0.8 increases the number of frequency tables represented in the network diagram to 70% of the total. However, by setting the grouping threshold as high as 0.925, we have managed to isolate the texts which display the most divergent semantic features and cluster them together. A list of the songs in these clusters is available in appendix 2.

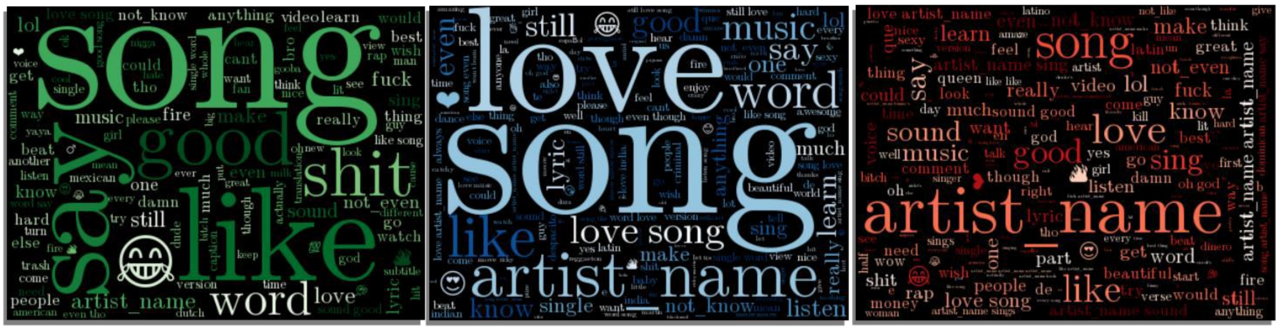


Figure 4.5: Green, blue, and red (left to right) colour-coded word clouds formed from the frequency-table grouping in figure 4.4

Figure 4.5 displays colour-coded word clouds formed from the tokens of the frequency tables in figure 4.4's red, blue, and green groups. These clouds were made with the use of the word cloud python package (Mueller, 2012). To construct these clouds, the frequency with which each token in the grouped frequency tables occurs was cumulatively added to a 'dictionary' of values. At the end of this process, this dictionary was populated by key-value pairs of the tokens that appear across all the frequency tables in a group and a frequency weighting. For example, if the token 'rubbish' appears in 3 of the frequency tables with a frequency of 3, 5, and 1 respectively then the frequency weighting for 'rubbish' in the dictionary will be 9. This frequency weighting is then translated into the size of the word in the cloud as a way of visually representing tokens that occur with greater frequency. It is worth noting that the first instinct on cursory analysis of figure 4.5 is that green (G) and blue (B) are two separate groupings due to the inability of the algorithms employed in the processing pipeline to group word-tokens semantically. To expand, in G the word “like” has been preferred when used in conjunction with the language-related keywords, whereas in B commenters have elected to use “love”. Although there is some difference in strength of emotion, both tokens belong to the same positive semantic field, which was not a dimension that was accounted for during the machine analysis of the tokens. However, as humans we are able to tell that these tokens convey roughly the same meaning, that is to say “like” and “love” express positive emotion towards something. Thus, it follows that the cosine similarity metric should be ignored and G and B should be grouped together for subsequent analyses. However, both the songs in G are by the same artist – ‘Takeshi 6ix9ine’[6]. This fact means that it is unlikely that G is just an anomalous grouping of two, and more probable that commenters are engaging with this artist in a specific manner. By examining these collocated in conjunction with one another in the following section, we will be able to draw conclusions regarding this specific form of engagement.

With the clusters and respective word clouds established, the next step was to discover the relationship between other songs in the corpus and these groupings. When the prevailing language attitudes of the groups are analysed and extracted in the following section, we can also relate songs in the wider corpus to these attitudes if we have already established what their relationship to the ‘parent’ cluster is. To determine this relationship, an average cosine similarity between a given song and cluster was taken. Whichever cluster presents the highest mean cosine similarity for a song then becomes that node’s parent. These relationships are visualised in figure 4.6 and it should be noted that the clusters are represented here by the colour-coded red, green, and blue nodes. Whilst these parent-child relationships will be further analysed in section 4, there are two things of note in figure 4.6. First, although the green group in figure 4.4 is the smallest in size, it is not the parent with the least children-nodes; the red group counts 10 members, whereas the green cluster totals 14. This

[6] For reference, the two songs are ‘BEBE’ and ‘YAYA’

affirms the earlier contention that the green grouping was not anomalous in nature – but rather similar songs did not fall beyond the required threshold of 0.925 to be included in figure 4.4. Secondly, due to the number and relative proximity of the nodes surrounding the blue group, we can conclude that the language attitude(s) that users promote in the comment sections of these videos is the most common among reggaeton songs.

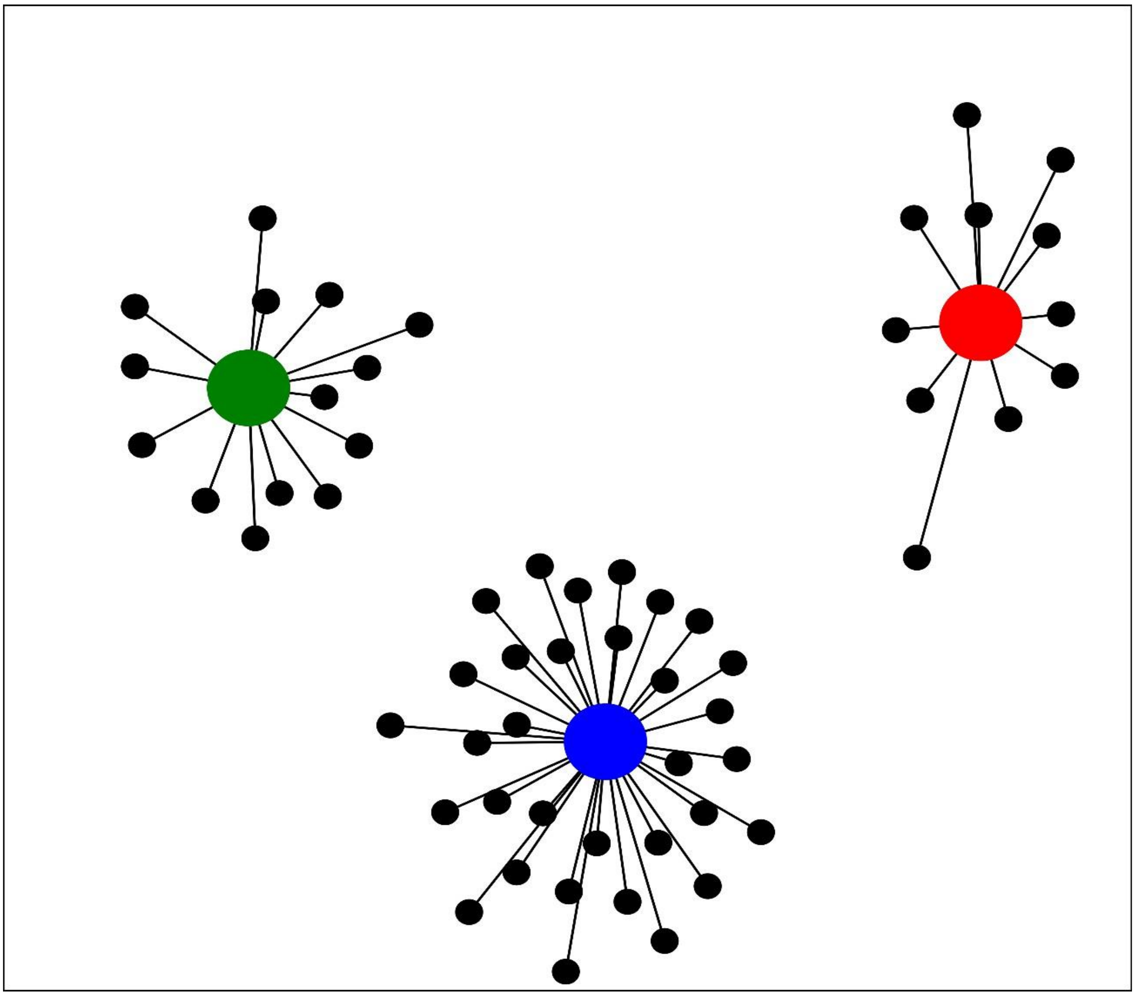


Figure 4.6: Network diagram of parent-child relationships between individual songs and the clusters displayed in Figure 4.4

As each comment was treated as a distinct text when processing the keyword collocations, this means it is likely that the tokens most present in the word cloud also appear in conjunction with each other in many comments. Thus, we are able to draw conclusions about the language attitudes that the commenters display towards these songs (and the children of the relative groupings) by analysing the array of tokens that occupy the most space in the word cloud.

4) Discussion of results

4.1) Word cloud analysis

4.1.a) Green group

As discussed above, it is unlikely that the results of the green grouping are as a consequence of the random distribution of synonymous tokens. In order for the cosine similarity of the members of the green group to be dissimilar enough from the rest of the corpus to form a distinct cluster, there must be other, lower frequency tokens that appear at distinct frequencies. The most notable of these is the token ‘shit’, which does not appear with great frequency in either of the other clusters. This expletive can have varied and sometimes contradictory usages (Terraschke, 2007). For example, it can serve as a vague noun (e.g., “this shit don't deserve 1b views at all”[7]), a negative adjective (e.g. “What the fuck is this..... !????? It's shit”), or a positive noun (e.g., “Tekashi is the shit. The best out there today”). However, by examining the other frequent collocates in the word cloud and then performing an array of computational statistical analyses, we can identify distributions of opinions that can broadly be categorised as positive or negative. With more resources and time, a model such as that used in chapter 2 to identify comments containing an opinion could be trained to do so. However, given the specificity of the language used in the corpus (as evidenced by the groupings upon which this section is based) it is unlikely that a model such as this would have the flexibility for any application beyond this one task. This would be in conflict with a main goal of this project (and the Digital Humanities): the creation and reproducibility of flexible, powerful tools.

I conducted a manual search of the corpus belonging to the two songs in the green group for instances where the five keywords (‘english’, ‘spanish’, ‘language’, ‘understand’, ‘speak’) and the

[7] As with all quotes from this project, these comments have been anonymised. This is particularly pertinent in this instance as many of the quotes in these sections express potentially harmful sentiments towards language.

token ‘shit’ occur in the same comment. By doing so, it was possible to get an overview of how and when commenters expressed an opinion in this way. As a general rule, if a commenter is to use ‘shit’ in a positive manner, they will do so with the use of an adversative subordinating or coordinating conjunction, e.g. “I can’t understand any of this shit but I like it”. In the same way a list of keywords was used as the basis for detecting language attitude, a list of adversative conjunctions was compiled: ‘but’, ‘although’, ‘nevertheless’, ‘however’, and ‘even though’. This list is not exhaustive as adversative conjunctions such as ‘yet’ and ‘still’ with common synonyms were excluded in order to reduce any false positives. In total, 47.2% of the usages of the token ‘shit’ were not accompanied by an adversative conjunction thereby meaning that the overall sentiment expressed in the comment is negative. It is true that this is a fairly small proportion of the overall corpus that expresses a negative opinion towards language use, but, as will be addressed in the following sections, this is the only clustering that has a significant element of negative language attitude.

Other tokens that appear with a high frequency are “say” and “word”. In the majority of cases (69.2%), these tokens appear in conjunction with the keywords “understand” and “speak” (at a ratio of approximately 9:1 respectively). Thus, we can conclude that another common topic for members of the green group (and the children thereof) is their lack of comprehension of the lyrics. In the case of “say” it is imaginable that commenters will express that they do not “understand” what the artist is “say[ing]” or that they do not “speak” a “word” of the language of the song. By highlighting their lack of comprehension of the lyrics, these commenters are setting clear boundaries between them, the ‘in-group’ in Brons’ terms (2015), and the foreign language singer. Taking this in conjunction with the aforementioned common negative sentiment leads us to believe that, in this particular cluster, the process of othering is taking place when commenters express their attitudes towards language use.

4.1.b) Blue group

The blue cluster is the most numerous in terms of core nodes and children nodes. The two most frequent tokens expressed across all these corpora in conjunction with the language-related keywords are “love” and “song”. Furthermore, these two tokens appear as a bigram (‘love song’) and are some of the only tokens to do so across the entire set of results. There are other tokens with (potential) positive connotations that appear with fairly high frequency, e.g. “like”, “good”, and the emojis “ ”and “🟡 ”. Thus, it can be concluded that the overall reaction of listeners to the specific use of language is highly positive in these songs. Another very frequent token that commenters use are ones related to the artist(s) of the song which have been normalised to appear as “artist_name” at the end of the pipeline (see section 3), implying that when a language attitude is expressed, it is usually in conjunction with the artist of the song.

4.1.c Red group

Of the three clusters, the red group displays the most variation in its vocabulary. With the exception of discussing the artists, there is no real prevailing semantic theme throughout the cluster in terms of what topic commenters tend to broach when displaying a language attitude. If one examines the lower-frequency tokens, however, the semantic field that is expressed is positive and contains many of the same tokens as the blue group (“love”, “good”, and “like”). There are also adjectives such as “beautiful” and “sexy” that appear at about the same frequency, which is to be expected if the principal topic of discussion is the artists involved in the music. It is not worth examining this in great detail so as to discern whether these tokens refer to the visual dimension of the music (as YouTube videos are principally an audio-visual medium) or the Spanish language as the tokens in question occur in such a relatively low frequency that drawing meaningful conclusions would be very challenging and would lack the ability to be generalised across the broader corpus. Furthermore, because cosine similarity is sensitive to the magnitude of the vectors it takes as variables (which in this case is the frequency of a given token), the lower frequency tokens contribute very little to the overall similarity scores between the two frequency lists. Thus, there is even less value in making generalisations based on these tokens.

4.1.d) Summary

To conclude, each word cloud represents a distinct grouping of ideas that commenters formulate when they discuss the use of language in the eighty-nine reggaeton songs. The green cloud is characterised by a more negative attitude towards the use of foreign language in the songs, whereas the blue cloud is an overwhelmingly positive response. The red cloud is a more disparate collection of frequency tables that are mainly linked as a result of placing the artists of the songs at the forefront when expressing a language attitude.

4.2) Parent-child analysis

As previously established, the blue cluster is the largest both in terms of core nodes and child connections. Additionally, the child nodes are mostly clustered close to the parent which demonstrates that the positive semantic field in this grouping does not diminish significantly. Thus, an immediate, key conclusion that we can draw is that English-speaking users are generally receptive to the Spanish language in the context of reggaeton songs. As with the blue cluster, the green child nodes are located in fairly close proximity to the parent and thus there is little dilution of the principal themes of the cluster among the children. The red group, however, contains the greatest variation in relationship between the parent and children. This is to be expected given the lack in topic-coherence displayed in the corresponding word cloud (with the except of the “artist_name” token). It follows that a group with a very disparate, low-frequency themes would also be more prone to topical diffusion.

With these conclusions in mind, we can make generalisations about the parent-child clusters in figure 4.6, thereby effectively partitioning the entire corpus into distinct units of opinion. From these generalisations, we can apply three different filters to the clusters and examine the way in which the artist of the song, temporal factors, and geographical success affect the types of attitude expressed by commenters

4.2a) Artist

Research has shown that musical audiences can be influenced when passing judgement onto performers by factors such as physical attractiveness (Ryan et al., 2006), perceived social class (Griffiths, 2010), and physical movements (Davidson, 1993). Dragojevic et al. also remark that language attitudes can be directed specifically towards speakers of that language, as opposed to general evaluations regarding the language as a whole (2021). Consequently, I assert that factors governing the ways in which performers are received by their audiences may also contribute to the ways in which commenters will express their attitudes toward the language used in the songs examined in this case study. To expand, in the same way that Griffiths demonstrates that percieved social class influences judgements on musical performances (2010), so do Dragojevic et al. contend that inferred social standing affects language attitudes formed by listeners (2021). Whilst in chapter 3 the audio-visual dimension of music on YouTube introduced a level of uncertainty into the conclusions drawn through the corpus analysis, for the aims of this particular section it may function to normalise results to a degree. To use the previous example of percieved social class, the added dimension of the visual presentation of the performers in the music introduce a myriad of social cues (for example sartorial choices by the performer or the environment in which the video is filmed) that audiences can use to inform judgements on the singer's social class. We would expect commenters to react in a similar manner to those artists who present themselves in a way that will cause audiences to classify them as a particular social class. The hypothesis that this all leads us towards is that the artist of a song may be one of the principal influences on the clusters observed in figure 4.4 and 4.6.

With regards to the specific data of the project, an initial supporting feature of the clusters for this hypothesis is the fact that the two songs in the green parent node are both by the artist 'Takeshi

6ix9ine'. If, for example, we observe that an anomalous proportion of the songs by the artist Bad Bunny are found in the green group (which, as we have established in 4.1.a, represents a collection of more negative language attitudes) then we can assert that commenters present a less-favourable view of the Spanish language due to their perception of the artist. It should be noted that, given the number of songs that this case study examines, it may not be possible to investigate the specific features of all the artists that affect the language attitudes of the commenters. However, if it is shown that the artist has a significant influence on the clusters, this case study will be able to function as a steppingstone off of which future researchers can base their own investigations.

One way of quantitively ascertaining the influence of artists on the groupings is by examining the distributions of songs by a given artist across the groups. If this percentage roughly follows the same distributions of songs across the groupings, then we can conclude that a given artist has no influence on the way commenters react to the language used. The distributions of the songs across the groups are as follows: the red and green group both contain 16 songs, which accounts for 18% of the total corpus each; the blue group has a total of 57 songs and thus contains 64% of the corpus. The mean percentage of songs per artist for each group was calculated, and the result of this are displayed in table 4.2.

Table 4.2: Table displaying mean percentage of songs by a given artist per group

Group	Mean % of songs by artist
Red	38.5%
Blue	82.2%
Green	60.0%

The artist distributions for the red and blue groups roughly follows what would be expected by their overall sizes provided that commenters are not influenced by the artist who sings the song. To clarify, the red group contains relatively few songs and thus the mean percentage of songs by a given artist in that group is also low, whereas the inverse is true for blue. Green, however, contains on average 60.0% of artists whose songs are associated with the cluster, which contrasts with the relatively small size of the group. This seemingly proves that commenters are influenced in their language attitudes by the singer of the song, however the methodology employed thus far is unreliable for two reasons. Firstly, this analysis only considers the principal artist of the song and ignores any featuring or collaborating artists and secondly, a given artist only authors on average 3.4 of the songs in the corpus (excluding artists who have only one song). The former is problematic as sixty-two out of the eighty-nine reggaeton songs identified for this analysis contain an average of 1.5 featuring artists (the maximum of which is 'Te Bote Remix' by Casper with five featuring artists). Excluding these featuring artists is ignoring a key dimension that could influence commenters to express a language opinion in some way. For example, in spite of the fact that a featured artist may not occupy the same amount of temporal or focal space in the song, if a commenters perceives them to be of lower social status, as per Dragojevic et al.'s contentions this may impact the language attitudes they express in the comments section (2021). Thus, it is necessary to revise our methodology in order to include the featured artists. As there is often a blurred line within this corpus between an artist featuring on another's track and one artist collaborating with another to create a song, I have elected to treat all features as co-authors for this analysis. There is the possibility that by doing so I may miss some subtle dimension of the way a commenter will use this extra-musical information (either implicitly or explicitly) when forming their language-attitude, but I have deemed the likelihood of this to be negligible enough to be a non-variable. Thus, by including featuring artists in the statistical analysis, the second problem outlined above vis-à-vis low average number of songs per artist is also circumvented. Now that there is much more metadata with regards to each song, we are able to get a more holistic analysis of the corpus, and table 3 shows the mean percentages revised to include the featuring artists.

Table 4.3: Table displaying mean percentage of songs by a given artist per group including featuring artists

Group	Mean % of songs by artist
Red	40.7%
Blue	76.1%
Green	45.2%

Now that featuring artists have been included, it is clear that there is no significant influence by the artist on the clusters of language attitudes. We can infer this from the fact that the red and green groups on average contain the same number of songs per artist – 40.7% and 45.2% respectively - whilst also containing the same number of songs of the full corpus – 18%. Additionally, the blue cluster displays the highest percentage by both metrics. One conclusion we can draw from this line of analysis is that including the featured artists is a necessary dimension when examining language attitudes in this manner. It was only by doing so that it became clear that commenters are unaffected by perceptions of the artists when forming their language attitudes. If steps had not been taken to obtain a more holistic view of the metadata of the songs, it is likely that some incorrect conclusions would have been drawn.

4.2b) Temporal analysis

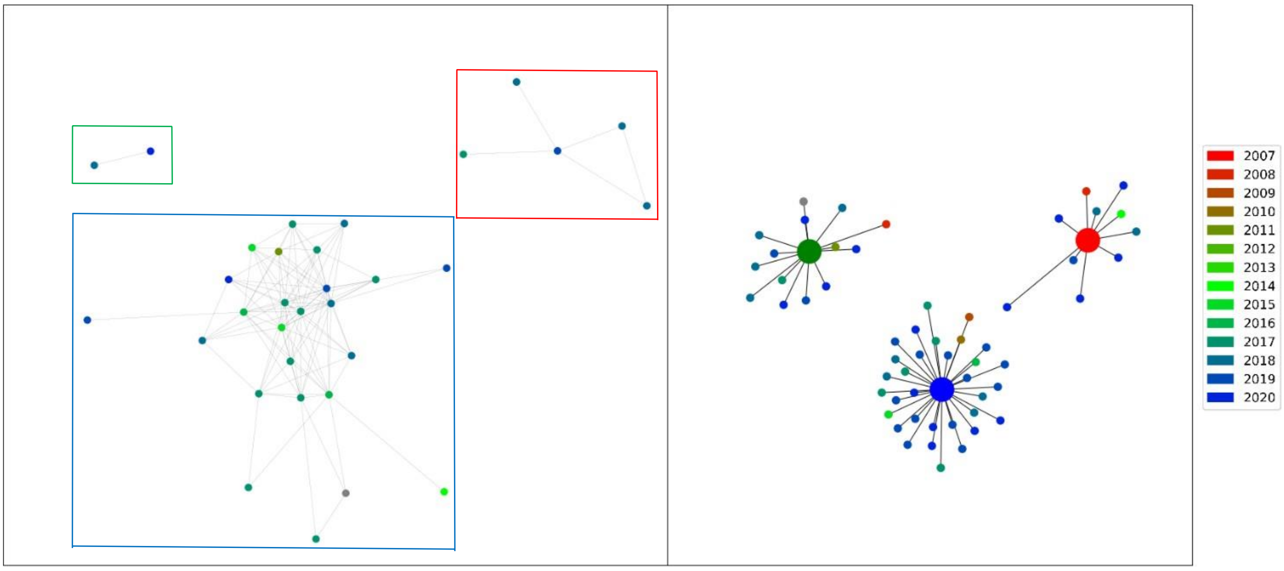


Figure 4.7: Network diagrams displayed in figure 4.4 and figure 4.6 colour-coded for date of release. Note: coloured boxes corresponding to the group have been added in the left-hand diagram.

2017 marked a turning point for the popularity of reggaeton songs (Arbona-Ruiz, 2017), however the question that this analysis will attempt to answer is whether there was a concurrent turning point in the language attitudes displayed by commenters towards the Spanish language in this context. Lippi-Green's assertion that over- and under-representation of a minority language in the media can contribute to its perception by non-speakers would imply that post-"despacito effect" songs would display more positive language attitudes in their comment corpora (Arbona-Ruiz, 2017; 2012). In terms of the language-attitude clusters identified in section 3, it would be expected that pre-2017 songs would be found in the green group as this group displays a more negative collection of language attitudes.

This is not the case, however. As can be seen in figure 4.7, there is no conclusive evidence that one group has a higher propensity to contain pre-2017 songs than another. It should be noted, however, that if average date of release of all the songs in a given group is calculated then the results are as follows: Green = 2016.8, red = 2017.5, blue = 2017.8. So, whilst there is no clear visual grouping in figure 4.7 vis-à-vis the date of release, songs in the green group tend to date earlier than the red and blue. However, the fact that these values are so close together undermines the validity of drawing any meaningful conclusions from this analysis. So, whilst Lippi-Green's assertion that increased representation of non-standard language in the media translates to more positive reception of said language may hold true for Spanish reggaeton songs (2012), it likely that there are more influential factors governing the groupings identified in this corpus.

4.2c) Geographical analysis

One problem with performing a geographical analysis corpus such as this is that, due to the fact that only Google has access to comment-specific metadata regarding where a commenter is posting from, there is no way to know the country of origin of a commenter. This is an issue for exploring language attitudes as, even if two individuals from different countries share a native language, there is no guarantee that the hierarchical schema of language prestige from which they may draw their language attitudes are exactly the same. To expand, in the United States it is well documented than the Spanish language has long been associated with lower-wage workers and working-class members of society, principally due to the influx of migrants from Latin America (Achugar and Pessoa, 2009). However, it is highly unlikely that the same view of the Spanish language will be held by a New Zealand native due to the fact that the same geo-political forces do not exist in that region of the world.

Whilst country-specific metadata of the comments is not accessible for this project, one available resource is the charts data collated and analysed when identifying the foreign language songs in the earlier stages of the project. If we can assess and compare how well one song performs across different countries, we may be able to form conclusions regarding the demographic of the commenter-base. For example, if 'Está Cabrón Ser Yo' by Bad Bunny performs better in the US charts verses the UK top 40 then it is logical to conclude that the commenters on that song are more likely to be based in the US. There are three immediate challenges posed by attempting classify the songs in this way: 1) it is difficult to define quantitatively exactly what it means for a song to perform well in the charts, 2) not all charts are comprised of the same number of songs, and 3) some countries have greater populations and thus some charts carry more weight on the YouTube corpus than others.

With regards to the first problem, it is challenging to decide by which metric to determine how well a song performs in the charts as there are several ways to interpret 'popularity'. For instance, is a song that spends 4 weeks in the top 5 and then subsequently drops out of the chart considered to have performed better than a song that spends 10 weeks at the 25 spot. To tackle this issue, it is important to understand how songs behave in the charts so as to create a general schema of song performance. By doing so we will be able to make a more informed judgement about what metric to use when discerning how well a given song performs. Figure 4.8 is a visual representation of the average trajectory of a song in the charts.

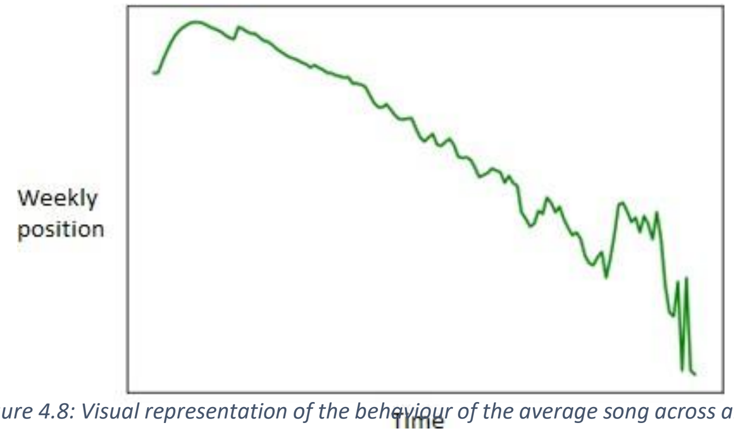


Figure 4.8: Visual representation of the behaviour of the average song across all charts

This data was generated by tracking the positions of each song to appear in the charts and then normalising starting positions to an arbitrary value. This was done as only the behaviour of the song with regards to changes in position were of interest, therefore we could discard any data regarding whether a song starts higher or lower than another one as this will affect the accuracy of the visualisation in figure 4.8. This is because the data was generated by placing the positions of a given song from week to week in a 2-dimensional array (a data structure with rows and columns) and then calculating the mean of each column. This meant that we were able to obtain the mean position of a song for every *nth* week and by normalising the starting position to an arbitrary value we are able to reduce statistical noise that would be generated by two songs that are located at different extremes of the chart[8]. It should be noted that this effort to reduce noise has not been fully successful as evidenced by the fact that the trajectory of the line in figure 4.8 becomes less stable and clear in its latter stages. This is because at there is much less data at the extreme of the X axis with which to form a representative average; whilst the total number of songs examined by this algorithm is

[8] There is great potential here for some insightful analyses into how songs behave in the charts that are tangential to the subject matter of this case study and project. For example, we could launch an investigation into the ways in which wildly popular songs are received by the public versus ones that do not attain as much chart success by examining their trajectories in the charts in this manner. As mentioned in chapter 2, all resources (including charts data) collected throughout the course of this project will be made available in an open-source format if future researchers decide to continue this line of investigation.

29051[9], the latter portion of the line only represents around 50 songs that have persisted in the charts for upwards of 100 weeks. Thus, we can easily disregard the anomalous behaviour of the line on the basis of a dearth of data with which to form accurate generalisations.

What figure 4.8 shows us is that songs tend to enter the charts quite near to their peak position and after having done so gradually move down the rankings. With this in mind it seems appropriate to make use of two metrics when comparing chart performance between two songs: a) the peak position of the song, and b) how many weeks it takes for the song to fall out the charts from this position. However, this second metric is problematic because, as outlined above, not all charts are the same length. For example, it may seem that a song has fallen much faster out of the New Zealand charts due to the fact that New Zealand only collates the top 40 most popular songs for a given week compared to 100 in the USA. One way to circumvent this would be to predictively model how a song might have performed had there been more data available for tracking its progress. However, such a technique would be highly speculative and would necessarily rely on data from other countries that, as outlined earlier, have their own unique geo-political factors influencing the way songs chart and are thus problematic for use in this way. As a result, I propose that maximum charting position is an adequate measure for quantifying what it means to chart better in one country over another. This leads us to the final challenge of using charts data when attempting to

[9] This number naturally includes duplicates of the same song in different countries' charts, however as we are only interested in the way songs behave in the charts, this is not a dimension for which I elected to control.

gauge the demographics of the YouTube comments which is the six countries do not possess equal populations.

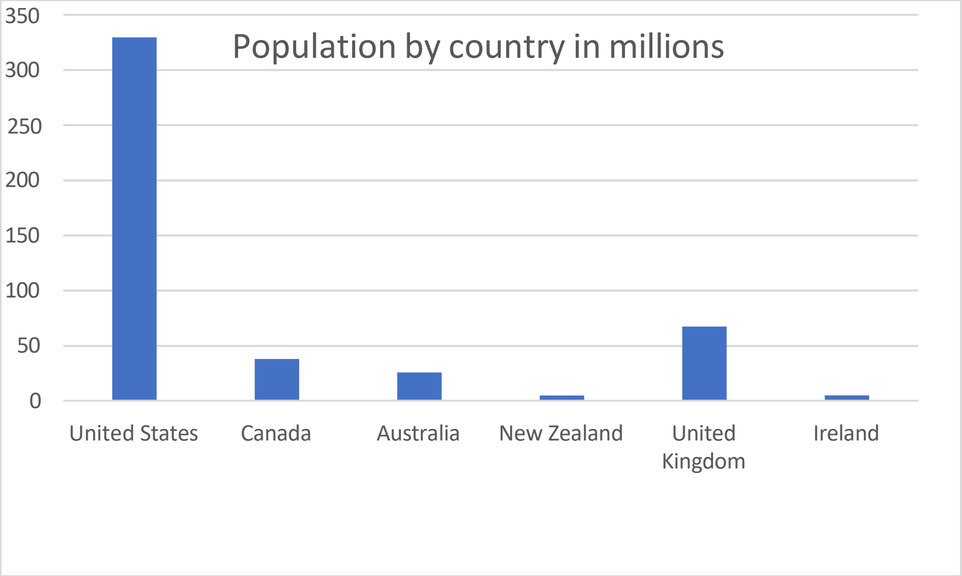


Figure 4.9: Bar chart showing populations of the six countries examined in this project as of 2020 (Worldbank, 2021). Population is measured in millions.

Figure 4.9 shows us that the United States has a greater population than every other country combined. This proves problematic as, unless a given reggaeton song does not feature at all in the United States top 100 (which is only the case for eight of the eighty-nine songs) then it will not be possible to claim that the corpus of comments of a given song is likely to be comprised of commenters who are not USA-based. Even then it is probable that the majority of the comments still belong to users who are based in the United States as a chart is not a finite list of the popular songs in a country, just a list of arbitrary length of the best performing songs in a given week. Just because a song has not reached the United States' top 100 does not mean that it is not in the public sphere of popularity – it could well be the 101st most popular song in the country that week. Thus, if we take it as a given that a large proportion of any reggaeton song's comment corpus is comprised of American audiences, what we can then do is examine the charts data and observe if charting well outside the USA has any effect on the language attitudes expressed in the comments. Figure 4.10 displays the network diagrams of language attitude clusters colour-coded for the country in which a song attained the highest charting position. If the USA was this country (as was the case for sixty-one of the songs) then the country with the next highest chart position was selected if possible.

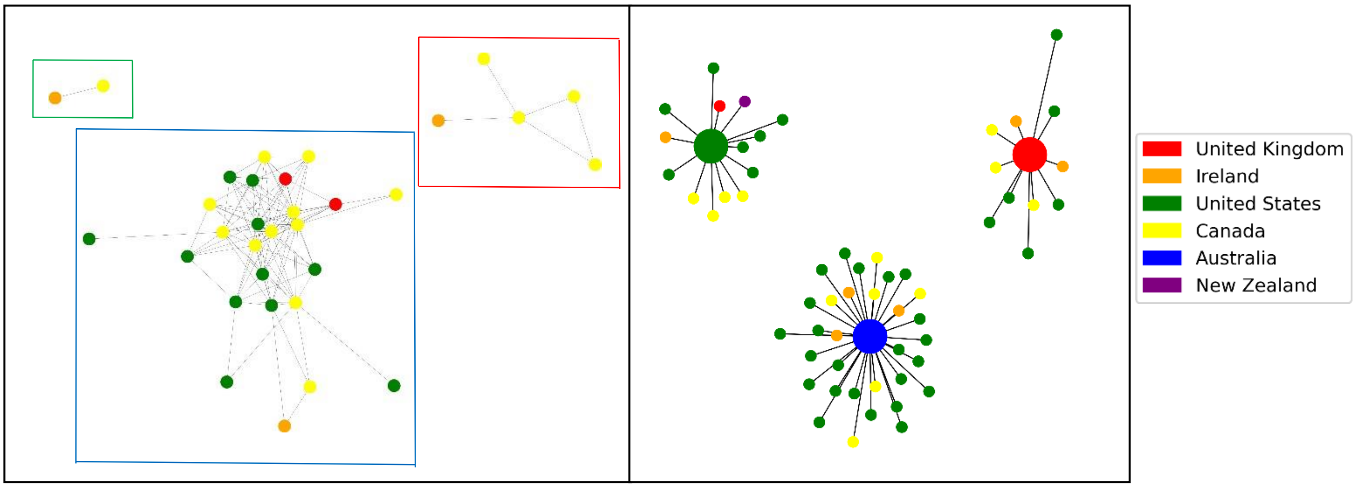


Figure 4.10: Language attitude clusters colour-coded to show the country in which the song charted the highest.

In figure 4.10, it seems that there is no correlation between the language attitude clusters and chart performance. If there were, we would expect that all of the songs that charted highest in Ireland (after the USA) for example would appear in the green cluster. Despite the rationale above, it is possible that this is because a) the technique chosen to measure chart success is too low resolution or ignores a necessary dimension of charting success, e.g., comparing the highest charting position relative to other countries as opposed to an absolute value, or b) YouTube comments are too far removed from, or are such a poor representation of, the charts for them to be reflective of thereof. To expand on this, charts companies have long been incorporating streaming services into their data, however, they are doing so from many different sources such as Spotify, Deezer and YouTube (Official Charts, 2021). Thus, it is possible that any nuance in terms of the extent to which listeners from a given country are represented in the YouTube comment section is erased when it comes to using the data from these multi-source datasets.

Another potential reason why this line of analysis did not produce any convincing results could be due to the fact that identifying the reggaeton songs in the scraped charts data proved problematic. At the beginning of the project when foreign language songs were being identified, no note of the exact location of the song in terms of chart position and date was taken. This was challenging for the purposes of this case study as it required me to iterate through the charts and identify a given reggaeton song by matching the title and artist of the song with the stored version. For the majority of the songs, this was not an issue as the stored information was identical to that contained in the chart data. For a number of tracks, however, either the title or the artist listed differed somewhat so it was necessary to run pattern-matching algorithms (specifically making use of Jaro-Winkler distance) in order to identify the foreign language tracks[10]. It is unclear to what extent this was an issue caused by using the non-official charts resource acharts.co to collect the charts data; it is equally likely that the difference in formatting was either due to idiosyncrasies in the way different charts present the tracks or as a result of inconsistencies in a non-monetised unofficial project to collect charts data. Regardless, it is possible that some data may have been excluded from the above analyses due to the fact that I was unable to locate it without running the lyrics querying algorithm to identify the songs as was done originally. This was due to ethical reasons as it is key that digital researchers not over-query databases that have made themselves publicly available and therefore I could not conscientiously a) query the site for data that I had already previously downloaded and b) complete a full scrape with appropriate delays within the time constraints of the project.

5) Conclusion

In this chapter I identified key clusters of opinion toward language attitude in the comments of eighty-nine reggaeton music videos. The three lines of analysis, geographical, temporal, and by singer, were unable to provide a solid explanation for these clusters. There are a number of reasons why this may be. The first explanation could be due to the fact that the initial methodology for identifying the clusters treated the corpus of comments of a given song as one whole text and not as distinct units of opinion posted by myriad users. The assumption on which I based this methodology was that commenters who remark on language attitudes would be highly influenced by the features of the music, and thus there would be general trends that uniformly track from video to video. This,

[10] Note: This was not an issue during the initial identification stages of the project as I made (ethical) use of search engines when searching lyrics sites for the tracks – see chapter 2.

however, was observed not to be the case. A potential clue to this was the fact that the cosine similarity threshold had to be raised to 0.95 before the clusters appeared, which demonstrated that there was a high degree of similarity between the different corpora. In light of this, a bottom-up approach that worked by identifying themes from at the level of the comment and then working up to the whole corpus may find more decisive results. One such example of this is in section 4.1a) in which a rule-based system for differentiating between the semantic usages of the token 'shit' was employed. This was proven to be an effective method for quantitatively evaluating language attitude, however, the disadvantage was that it required careful consideration and research and was only applicable to a small subset of clauses that express an attitude towards language. With this is mind, creating a rule-based methodology for tackling the problem at hand may result in a great deal more work for the researcher, but also more accurate and flexible results from which to draw conclusions

Another reason why the analyses carried out in this chapter revealed no particularly insightful conclusions could be due to the way in which commenters discuss language attitudes on YouTube. Previous scholars have had found success using YouTube as a source for investigating language attitudes (see: Aslan and Vásquez, 2018; Ivković, 2013; Walters and Chun, 2011), however it is possible that in the context of popular foreign language music language choice is not as much of a contentious issue (unlike for the Eurovision song contest, as Ivković examines). As a result, commenters may not concern themselves with such issues; on average, only 0.1%[11] of those commenters who expressed an opinion about the song did so on the topic of the language used. Furthermore, as was discussed in chapter 2, users tend express positive opinions toward these songs, and we can assume this extends to opinions towards language use. We can claim this because, as was discussed in section 4.1, there was only one major element of the three clusters

[11] This number fluctuated somewhat between songs. A fruitful line of analysis for future research may be to examine those songs whose comments boast a greater number of language attitudes and compare these opinions to those found in less attitude-heavy corpora

that could form the basis of a negative opinion – the token ‘shit’ – and even then, only around half of the usages of this token were negative. Thus, given the sparsity and uniformity of the comments, perhaps it is not appropriate to use a corpus such as this for the purposes of examining language attitude.

Whilst this case study has used ‘Despacito’ as a reference point for the “Latin Renaissance” that occurred in 2017 (Acevedo, 2019), no particular focus was placed on the song itself. As ostensibly the most popular foreign language track to have penetrated into the sphere of popular music in the English-speaking world, this project would be remiss not to examine the factors that contributed towards its success, be they musicological or sociocultural. The following chapter will concern itself with unpacking the success of ‘Despacito’ and other foreign language songs that achieved great success for their time in which they released.

Chapter 5: 'Alexa, play despacito' – examining the role of the meme in the reception of viral foreign language songs.

1) Foregrounding

1.1) Memes and virality

In recent years, social media has begun to play a defining role in the way the public consumes and disseminates information (Mangold and Faulds, 2009). Marwick contends that an example of social media-driven communication is the phenomenon of internet memes (2013). She states than a contemporary internet meme is a (typically humorous) unit of culture that gradually gains popularity and influence as it is transmitted from person to person. Marwick attributes the origin of the term to Richard Dawkins who observed that, much like biological genes, fashion and trends tend to evolve and propagate through a process similar to that of Darwinian natural selection. Marino identifies three historical stages of the internet meme: proto-internet memes in the 1990s typically propagated via emails and discussion groups, the internet meme subculture that existed from the late 1990s to 2005 that spread on message boards such as 4chan and Reddit, and finally the era of global internet memes that has existed from 2005 to the present day (2015). This latter stage is characterised by "industrial" and "self-reflexive" memes facilitated by social media and dedicated websites for the 'fool proof' generation of memes with premade templates (pg. 76). The concept of the meme is intrinsically linked to that of virality in the online space. Goel et al. (2016) describe the process of 'going viral' in terms of a rapidly spreading contagion travelling from person-to-person. Whilst Stein et al. reject the implications of the nomenclature surrounding this process as it implies a level of passivity on the part of the spreader of the meme (as opposed to the much more proactive process that Dawkins originally associated with the phenomenon) (2014), this still remains a useful metaphor for conceptualising how these memes and ideas can spread exponentially.

Defining virality in this way provides us with a useful base upon which we can select the songs that will be examined in this case study. Google trends is a tool developed by Google that allows users to visualise the relative quantity of specific Google searches over time. Two songs that have attained the level of virality are Despacito and Gangnam Style. Figure 5.1 shows the Google trends of these two songs between 1st January 2012 and 1st January 2019[12] [13].

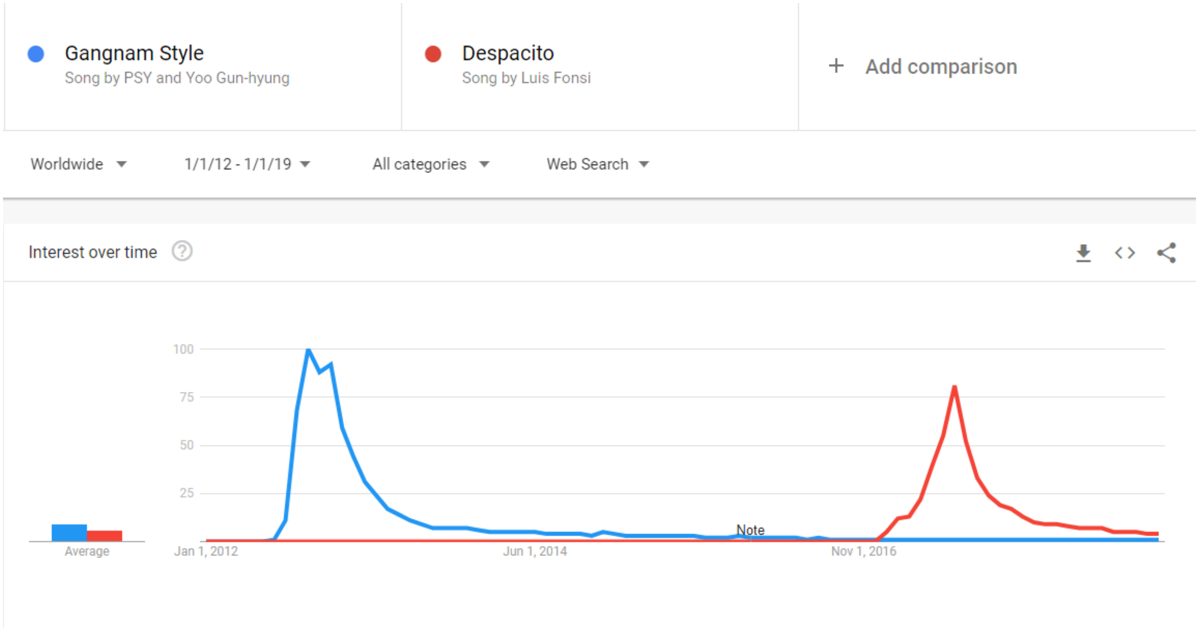


Figure 5.1: Google Trends result for the search times 'Gangnam Style' and 'Despacito' (Google, 2021).

Within the corpus assembled in this project, 'Despacito' and 'Gangnam Style' are the fifth and first most numerous songs in terms of English-language comments (including replies). These two songs excluded, the nine songs with the greatest number of comments all belong to either BLACKPINK or BTS. In spite of this clear interest and popularity on YouTube, none of these K-pop songs ever reached above a 5 on the scale in figure 5.1, thereby demonstrating that they do not have the 'virality' of either 'Despacito' or 'Gangnam Style'. In fact, when other notable songs such as 'Mi

[12] Google helpfully provides us with the ability to search their archives by topic as opposed to the raw search term. The challenges of parsing data for different string permutations have been thoroughly discussed in this project thus far – and will be revisited in this chapter – thus I elected to make use of this feature in spite of the fact that it introduces a level of imprecision to the results.

[13] This data pertains to worldwide searches on this topic as the Google Trends tool does not permit the grouping of countries in its filters. This is of course problematic as this case study is only concerned with the anglophone world, however, due to the nature of virality and its pandemic-like behaviour as a result of globalised media, I do not believe that there is much to be gained from arduous and rigorous comparison between different English-speaking countries.

Gente' by J. Balvin and 'Gentleman' by Psy were examined in the same way, they too did not present the viral success of 'Despacito' and 'Gangnam Style'. This does not mean that these songs will not be of use in the coming investigation. The reason 'Mi Gente' and 'Gentleman' were singled out is they are apt 'foils' for 'Despacito and 'Gangnam Style' respectively. In the same way that one can easily draw parallels between 'Gentleman' and 'Gangnam Style' (both are subversive examples of the K-pop genre by the artist PSY who have a focus on humour and visual storytelling), so are 'Mi Gente' and 'Despacito' comparable insomuch that they are reggaeton songs that garnered huge success in 2017 and were later remixed by highly successful anglophone artists (Beyoncé and Justin Bieber respectively). With this in mind, these songs will provide a useful point of reference when determining the extent to which memes play a role in the way the English-speaking public engages with viral foreign language music on YouTube.

1.2) Memes and YouTube comments

Xu et al. remark that there is a dearth in literature on the study of memes in the YouTube online space with scholars tending to focus more on other popular social networking sites like Facebook and Twitter (2016). Xu et al. do make use of YouTube comments in their analysis, but it only to create networks of YouTube videos via users who have commented on both, and only the 1000 most recent comments are collected for each video. Aslan and Vásquez also make use of their meme-oriented investigation on YouTube (2018), however, theirs mainly concerns itself with the meme of the video itself and does not focus on identifying or tracking memes within the comments of the video itself. If, as Xu et al. posit, there is a lack of studies on video memes on YouTube, then there is even less in the way of examining how text memes evolve within the comments of these videos. This is the task that falls to us and given that there is no precedence which we can follow, we must also provide a conceptual framework for not only tracking these memes but identifying them in the first place.

2) Methodology

2.1) Hypothesis

To investigate the extent to which English-speakers treat these songs as a meme, we will perform computational text analyses on the comments section. The object of this analysis will be to ascertain the degree to which text-based memes about the song in question make up the corpus of comments. This will be an insightful analysis as not only will it provide us with an idea of how the public treats wildly popular foreign language songs, but we will also be provided with a unique opportunity to examine the way the public behaves in the space of the YouTube comment section on a more general level.

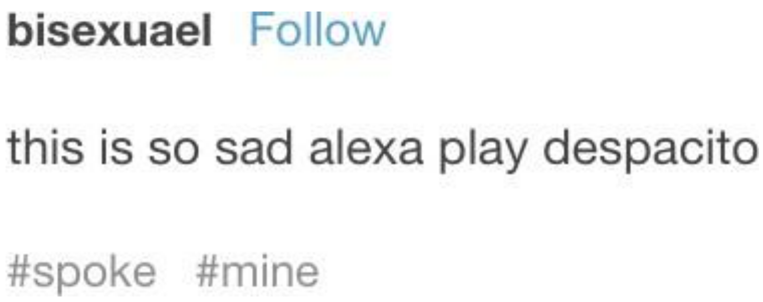
2.2) Initial considerations

There are a number of ways to approach the identification of text memes about music computationally and it is important to strike a balance between efficiency and thoroughness. The following quote from Veselovsky et al. summarises the crux of the matter (and indeed the broader challenge in the Digital Humanities of moving from distant to close reading):

"Qualitative approaches to the study of music have uncovered rich connections between music and social context but are limited in scale, while computational approaches process large amounts of musical data but lack information on the social contexts music is embedded in" (2021)

It is therefore key that the positives and potential drawback of different computational approaches are examined and evaluated so as to accurately capture both the information and social context in which it finds itself.

As established, the core features of a meme are its reproducibility but also mutability (Marwick, 2013). In this way, it is able to undergo the Darwinian process of natural selection, but this poses a challenge for the identification thereof. A particularly popular meme shortly after the release of 'Despacito' was the phrase 'This is so sad, Alexa, play Despacito' (Felderman, 2018). This meme originated from a Tumblr post by the user 'bisexuael', pictured in figure 5.2.



bisexuael Follow

this is so sad alexa play despacito

#spoke #mine

Figure 5.2: The Tumblr post that spawned the 'Alexa, play Despacito' meme (Know Your Meme, 2022).

A search of the comments for the keywords "Alexa" and "Despacito" reveal the ways in which a text-based meme can evolve and 'mutate' over time. The following are four examples displaying the ways in which a change in social context can influence memes about music[14]:

1) 'This is so sad, Alexa play despacito'
2) 'Alexa play Despacito by 6ix9ine'
3) 'is this what comes up when you tell alexa to play despacito two'
4) 'When u tell Alexa to play Despacito 2'
5) 'Alexa forget about despacito and play taki taki'

Depending on the type of commentary the author is providing through the use of the 'Alexa, play Despacito' meme (example 1 in the above list), a number of syntactic or semantic features can be altered. To expand, in example 2, a comparison between 'Despacito' and 'BEBE' by Takeshi 6ix9ine is provided with the use of the meme format from which they have removed the preceding clause "this is so sad". By removing the preface of the phrase "Alexa play Despacito", they have repurposed the meme in such a manner that provides several interpretations; it could be that the user is providing a value judgement on 'BEBE' according to their opinions towards 'Despacito'[15] or alternatively remarking that they observe musical (or linguistic) similarities between the two tracks

[14] As stated before, these comments have been anonymised for ethical reasons

[15] This is an example of a shortcoming of corpus-based analysis. Without extra-linguistic information such as body language, tone, or contextual information, it is not possible to ascertain whether this is a positive or negative value judgement.

that, to them, render the tracks comparable. Examples 3 and 4 are expressing roughly the same sentiment as example 2 but with slightly altered syntax and without specifically mentioning the song. Finally, example 5 is clearly stating that 'Taki Taki' by DJ Snake (and several featured artists) is a superior song to 'Despacito' by changing the object of the imperative 'play' to 'Taki Taki' whilst also changing the verb pertaining to 'Despacito' to 'forget'.

With the use of four (non-exhaustive) examples, this short analysis demonstrates the extreme mutability of the text meme within YouTube comments. In doing so, we are immediately able to exclude most basic of computational approaches to the identification of text memes. By far the most computationally efficient method would have been counting the number of instances of strings within the corpus. When such an algorithm had finished its iteration, we would have been able to see which whole comments occurred in the greatest quantity, thereby revealing to us the text memes that users propagate within the corpus. However, the fact that this technique attempts to exactly match the comment makes it an inappropriate methodology for identifying a syntactically and semantically mutable unit of text; were this technique used to examine the 'Alexa, play Despacito' meme on the corpus, just 3% of the total instances where 'Alexa' and 'Despacito' occurred in the same comment would be identified[16]. It is thus important that we write an algorithm that is able to account for potential syntactic/semantic changes in the text meme and the most straightforward way to do so is through the use of keywords.

2.3) Keyword approach

2.3 a) Initial thoughts

As demonstrated above, using keywords is an ideal way to search a corpus for the different permutations of a meme with which we are already familiar. However, I propose that it will also form a key part of the identification of text memes. Were we able to algorithmically conclude that a

[16] This number reduced to 2.7% when the keyword 'despacito' was replaced with 'play' in the search as clearly a small proportion of users wish to make a commentary by replacing the object of the verb.

number of text strings contained the word “alexa”, “play”, and name of a song in that order, more manual examination would inform us of the existence, propagation, and evolution of the ‘This is so sad, Alexa, play Despacito’ meme.

Ferrara et al. propose a technique of identifying memes on Twitter with the use of what they coin ‘protomemes’ (2013). It should be noted that for Ferrara et al., a meme is much closer to Dawkins original conception; a unit of information, be it an idea or concept, that is transmitted from person to person as opposed to the more humorous connotation that the term carries today. For the purposes of this case study, this is a useful definition of a meme. This is because we are examining a concept that is not necessarily tied to humour: the reception of songs. If we were to exclude repeated topics and ideas on the basis that they were not jokes, we would necessarily be ignoring an important dimension of the dynamic between English-speakers and these foreign language songs. For Ferrara et al., protomemes are ‘atomic entities’ of memes, of which only one is the actual text contained in the tweet (2013). The other elements of a tweet that these authors designate as protomemes are hashtags, mentions, and URLs to other websites. By using this metadata, Ferrara et al. are successfully able to cluster tweets by the general topic they broach and thereby identify memes. The issue with using this exact methodology in this situation is that there is no such metadata available for YouTube comments. Whilst it is possible to mention another user in ones comment, it is not only rare to do so but typically occurs in the replies. As has been previously discussed, a limitation of using YouTube comments downloaded from the API is that it does not provide access to replies of replies at the time that this project was carried out. Consequently, using such metadata may result is unreliable results that only capture a small proportion of actual usages thereof. One element of metadata that is available to us when using a corpus of YouTube comments is the number of likes that a comment receives. Hypothetically, the number of likes a text-meme receives could be a useful, quantitative measure of the Darwinian process that memes undergo when spreading from user to user. However, given that there is a dearth of research into text-memes on YouTube, the use of like-count will require careful examination and calibration as, unlike the metadata available to researchers who use tweets to form their corpus, the number of likes a comment has does not provide any semantic information given that it is a simple integer.

A common approach to natural language processing (NLP), and one of which this project has made extensive use, is the process of simplifying text for computation analysis in order to reduce its statistical noise. Some of the measures that have been employed thus far are the removal of stop words, the lemmatisation of morphological variations, and the removal of extra-linguistic text features (excluding emojis). Additionally, the subsequent techniques for the computational analysis of the processed text have also all been bag-of-words techniques that do not account for syntactic elements of the strings[17]. Given that the content of the text was just one aspect of Ferrara et al.'s successful approach toward the clustering of memes (2013), it is logical to assume that if this approach is to succeed it will be necessary to also employ some extra-semantic information when processing the strings. With this in mind, as strings are processed, the syntax of the string will also be logged and employed when trying to identify the clusters of memes – a process that will be detailed in section 2.4. The following section, however, will focus on the process and challenges of comparing vast quantities of comments based off the keywords within the strings.

2.3 b) Practical approach to keyword processing: challenges and methodology

The approach to using keywords to identify text memes is at its core a data-clustering methodology; an approach that has been widely researched and implemented in areas from computational genealogy to natural language processing (Srinivasa-Desikan, 2018; Blanke and Aradau, 2021). K-means clustering is a common clustering techniques employed today and one that comprises part of many pre-built python libraries, such as a scikit learn (Garbade, 2018). K-means is a machine learning algorithm that attempts to find a predetermined number of clusters (denoted by K) within a dataset through trial and error. Due to the aforementioned fact that there is a wide array of prebuilt tools to

[17] Two exceptions to this in chapter 5 was a) the distinction between verbs that had been negated and those that had not and b) the collection of n-grams. These measures somewhat took into account syntactic elements on an (albeit basic) level of analysis.

this end, using K-means algorithms was one of the first avenues I investigated. As I discovered, however, the main issue with using K-means is that it does not cope well with clusters that vary greatly in size, and thus tends to overfit the results (Dabbura, 2018). This is an issue for this particular corpus as, to use an example, there is a much greater quantity of comments that contain the phrase “good song” versus “Alexa, play Despacito” (or variant thereof), and thus the respective cluster sizes are likely to be vastly different. Consequently, it was necessary that a different clustering algorithm was used in lieu of K-means and so I elected to once again make use of cosine similarity to identify links between comments.

The main challenge faced by using cosine similarity to compare individual comments is the raw size of the corpus. Unlike in chapter 4 wherein only the cosines between 89 vectorised texts were being calculated, on Despacito alone there were 1.7 million different comments to be compared, which overall resulted in $2.89x10^{12}$ (just under three trillion) operations to be performed. Initial tests using the cosine similarity function from scikit-learn that was previously employed in chapter 5 put the total run-time for 1.7 million comments at 4737 days – an unfeasible length of time for this project of this nature (and indeed any other) – but a number of optimisations were able to produce a speedup of 54x. First of all, instead of using native Python data structures, I stored all data in NumPy arrays – a data structure that forms the basis of the NumPy Python library. Python was not initially designed with scientific or heavily numeric computation in mind, however, as its popularity as the programming language used for artificial intelligence and machine learning has increased, so has the need for efficient numeric computation (Verma, 2019). This is an area in which NumPy excels for three reasons: firstly, the data within the arrays are much more densely packed in memory (RAM); secondly, NumPy is able to parallelise computations performed on its arrays; and finally, NumPy functions are executed in the programming language C – the language upon which Python is built. To use an analogy, if we imagine that we wish to order some goods (process some information) then using NumPy allows us to bypass the distributer (Python) and directly get these goods from the factory itself (C). Naturally, purchasing from a distributer is much simpler and in the vast majority of cases the ideal choice, however if we want to do a bulk order – in this case process huge numbers of comments – then for the sake of efficiency is preferable to access the manufacturer ourselves, hence using C rather than a package in Python. By making use of NumPy arrays for storing the data and results, I was able to reduce the runtime of calculating 1.7 million cosine similarities to around 200 seconds. As we still had to do so 1.7 million times, however, the overall runtime still remained around 10 years[18]. Thus, it was necessary to further optimise this script.

The Numba Python package is a scientific computing package designed to optimise Python code in much the same way that NumPy does: by translating Python into more efficient machine code (Numba, 2018). Whilst limited in the types of functions and algorithms it can optimise, (fortunately) cosine similarity using NumPy arrays is one such application that benefits greatly from optimisation from the Numba package. Chankedar provides a function for calculating cosine similarity between two NumPy arrays using Numba (2019), and by using this function instead of those found in the Scikit-learn library, the calculation of cosine similarity for 1.7 million comments was reduced to around 6 seconds. This is a vast improvement that is highly promising for future research and projects. This optimisation translated into 133 days of total run time - a length of time still out of the scope of this project but a helpful speedup, nonetheless. However, there still remained a large level of redundancy in the script. Table 5.1 shows the matrix of cosine similarities that is created by this script.

Table 5.1: Example matrix of cosine similarities:

	Comment 1	Comment 2	Comment 3	Comment 4
Comment 1	1	0.7	0.5	0.2
Comment 2	0.7	1	0.9	0.1
Comment 3	0.5	0.9		0.6
Comment 4	0.2	0.1	0.6	

[18] 200 seconds * 1.7 million = 340,000,000 seconds = 10.78 years

The coloured cells are those that are redundant calculations. As can be seen in the table, if calculating one column at a time, there is no need to compare a comment with itself (black) and there is also no need to compare a comment to those with which it has already been compared (red). By ignoring these redundant values, we are speeding our script up by just under 50%[19] as we only have to calculate n/2-n values, where n is equal to the total corpus size.

At this stage, there are two further methods to reduce the time it takes for this script to run: 1) parallelise the script into multiple processes, or 2) apply some criteria to the comments that reduces the quantity of comments that we need to compare using cosine similarity. In brief, the concept of parallelisation in this case refers to running an algorithm across multiple CPU cores as separate processes. The more CPU cores available, the more computation can be done at the same time and thus the faster the script can execute. I only had access to four CPUs during the application of this methodology, but at the time of writing there are more powerful options available through cloud computing. However, employing such drastic cloud computing measures is not in line with core digital Humanities goals. Throughout the course of this project, all methodological steps have been takes with the goal of easy reproducibility and open accessibility. Were we to design a methodology that requires the use of hundreds of cutting-edge units processing power only accessible at financial cost, we would have failed in these striving for these ideals. Naturally, if future researchers wish to extend this project further, it will simply be unavoidable that they will require more processing power. However, by employing the second method of reducing run-time outlines above – filtering out some (if not most) comments prior to processing – we will be able to produce a methodology that can process a maximal number of comments on hardware accessible to most, if not all, researchers.

[19] In reality the script must still iterate from one item in the list to the next and perform some negligible actions, the end result of which is that actual time savings are less than theoretically possible from this optimisation.

As established in section 1 of this chapter, memes are subject to a Darwinian process of propagation akin to natural selection (Marwick, 2013). On YouTube, the logical way this would be expressed is through the number of likes a given comment has received from members of the public. If contentions regarding the spreading of memes by authors such as Marwick, Stein et al., and Dawkins hold true for the YouTube comment space (1976; 2013; 2014), then we need only include those comments that have reached above a certain threshold of likes as markers of their success as a meme. We must choose this threshold carefully; too low and the quantity of comments with which must contend becomes unwieldy and increases computing time, too high and we run the risk of excluding valuable data. Furthermore, we will be examining music videos whose corpus of comments differs greatly in size. This means that, although we may be able to ignore the likes threshold on a video and include all the comments as there are less of them, we should not. The reason for this is that we may introduce statistical skew in our data if we do so. There is no reason to believe that, just because one video has a smaller corpus of comments, the 'Darwinian' process through which memes are selected would no longer function. Initially, I hypothesised that we ought to scale the threshold of likes to the proportion of comments on a video it counts, not a raw integer. For example, we need to ensure that we are capturing the top 40% most liked comments, not just all comments with more than 30 likes (as this number will change with the popularity of the video). However, thanks to the steps we have taken thus far to optimise our algorithm, we were afforded more flexibility. In the 'Gangnam Style' corpus, comments with at least 1 like account for 257,512 comments out of the 2,993,360 total. Predicted estimates for calculating the cosine similarities between 250,000 comments with all the optimisation steps outlined above put the total runtime at 12 hours for this step, which is an acceptable and manageable quantity of time considering we must also compare syntactic information as well. Consequently, we are able to reframe the likes threshold as simple Boolean, i.e., a true or false statement as to whether a comment has been liked. This renders the process of scaling the threshold for other videos much simpler as we can apply the above to all the comments instead of attempting to capture a necessarily continuous proportion of comments that have been divided into discreet categories.

2.4) Syntactic approach

As a field of study, syntax concerns the way in which sentences or phrases are structured out of their constituent words (Radford, 2004). As previously established, ‘mutating’ syntax in order to alter the semantic purpose of the meme occurs frequently in the propagation of text memes on YouTube and is thus a key part in identifying them and analysing their uses. In example five given in section 2.2, it was observed that a user had mutated the syntax from the original clause “Alexa play despacito” in order to create “Alexa forget about despacito and play taki taki”. By providing an additional verb phrase (“forget about despacito”), they repurposed the meme in order to provide a likely positive value judgement on DJ Snake’s ‘Taki Taki’. In this example, were we to perform a simple keyword extraction on the phrase, the output (once stopped and processed) would be the following list of words: “play”, “forget”, “Alexa”, “taki taki”, and “despacito”. This list is helpful as it is uncomplicated and generalises well as the number of strings that are processed also increases. It is for these reasons that this was the methodology employed when performing text processing in chapters 3 and 4, the eventual goal of which was token-frequency analyses. However, it also lacks any contextual information that would enable us to draw meaningful, closer analyses. For example, from the processed bag-of-words above is impossible to conclude that the object of the verb “forget” is “despacito”, and that the author is commanding “alexa” to instead “play taki taki”. Additionally, were we to utilise simple keyword collocation frequency tracking (such as the technique that was employed in chapter 4 when identifying common language attitude collocates), this use of syntax - and perhaps other, more obscure examples like it - would go unnoticed. This difficulty of moving from distant to close reading is a principal challenge in the Digital Humanities and has been discussed in depth thus far in the project. In spite of this I propose that we could make use of syntactic elements when computationally examining the comments. It would be ideal if were able to build a model for processing language that would be able to identify, for example, the general trend of “Alexa” followed by an imperative verb and a noun corresponding to a song title. We would then be able to incorporate this methodology alongside the cosine similarity calculations above in order to further increase our abilities to detect memes. This is the methodological aim of this section.

2.4 a) Theoretical framework of computational syntactic analysis

To date, much has been achieved in the domain of computational syntax processing. Roark claims that finite-state approaches are the most common of these (2007), that is to say programs that can only be in a finite number of states while performing computation. He also contends that these models are somewhat limited in their capacity to process some very common syntactic structures in natural language. Due to the constraints in this project however, we will be limited to the finite-state approaches as there is an abundance of work carried out by previous researchers/computer scientists on which we can draw (although it may prove a fruitful continuation of this research to examine the implications of non-finite state approaches to this methodology). Roark identifies four finite-state approaches: n-gram models, class-based language models, part-of-speech (POS) taggers and noun-phrase chunkers. It is the latter two of these approaches that we will employ in this methodology due to the fact that the tools to do so are readily available in the Natural Language Processing Toolkit (NLTK) (Bird et al., 2009). Roark contends that homonyms (words that have different meanings despite being morphologically identical) pose the greatest problem for POS tagging. As an example, he cites the phrase “the aged bottle flies fast” in which all but the definite article can have multiple syntactic functions (pg. 156). For example, it could be that elderly flies are foregoing food for a period or alternatively that pensioners are containing flies within bottles at speed. Roark provides a conceptual framework for overcoming this problem of homonyms: a Hidden Markov Model (HMM). This model consists of converting a sentence into a Markov chain – a list of probabilities (originating from a pre-tagged corpus) that a word corresponds to a given POS tag. These probabilities are calculated based off the Markov assumption which assumes in this application that the tag for a word is only dependant on the one directly preceding. For example, in the phrase “This is so sad” we need only calculate the probability that “so” is an adverb (as opposed to the various other grammatical functions that it can fulfil, e.g., a conjunction: “I don’t want to go, **so** I won’t”) based off the word “is”.

As discussed above, the reason for employing POS tagging in order to collect syntactical information is the availability of tools constructed to this end, namely the POS tagger in the NLTK module (Bird et al., 2009). NLTK has been noted to underperform with regards to efficiency when contrasted with other packages because, thanks to its doubling as a teaching tool, it contains a lot of extra computational baggage (Honnibal, 2013). This is not so much an issue for this project as, although millions of strings will be analysed, the overall processing time is not prohibitively long (roughly 10 minutes), which can be reduced even further if the task is multiprocessed. Additionally, the clarity and accessibility of NLTK will be advantageous for any future, non-expert researchers who wish to employ a similar methodology to that of this project. More generally, the advantage of using pre-existing tools that draw their information (in this case POS probabilities) from public resource is that it expedites the research process as we do not have to create any tools or training datasets. The drawback is that with a specialised corpus such as YouTube comments, there is a possibility that the algorithm will face data upon which it has not been trained. There are a number of corpora that have frequently been used as a resource for Natural Language Processing, in particular NLP that makes use of machine learning that requires a curated training dataset. One common example is the Wall Street Journal (WSJ) corpus – a corpus comprised of 30 million words from the Wall Street Journal that has been POS tagged and parsed (Linguistic Data Consortium, 2021). Whilst an immensely helpful tools for NLP practitioners, given that the excerpts date from 1987-1989 it is not going to contain references for inputs such as “Despacito” or recognise that, in the sentence “This song is fire”, the word fire is functioning as an adjective. However, given the time cost that would be associated with manually tagging a dataset of YouTube comments, it is unfeasible to make use of a specialised dataset in this way. Additionally, given the richness of the WSJ corpus (being 30 million words in length), the lack of specialisation will hopefully be counteracted by the general accuracy of using models trained on a corpus of this size.

NLTK also offers a broader array of POS taggers than the conceptual model provided by Roark (2007). Whilst Roark focuses on HMMs (and closely related methodologies), the NLTK documentation recommends making use of an averaged perceptron. Put simply, an averaged perceptron is a supervised machine learning model that attempts to find distinctions between classes of data. NLTK does also provide a POS tagger that makes use of an HMM in its implementation. As a result, it was trivial to devise a test that compared the two both in efficiency and accuracy. The first ten thousand comments of the Justin Bieber remix of 'Despacito' were POS tagged using each method and the start and end times were recorded. Using the recommended perceptron tagger, the algorithm required 11.6 seconds to tag ten thousand comments, whereas the HMM needed 27.5 seconds; the recommended tagger was over twice as efficient as the HMM. However, accuracy is also a consideration when tagging our dataset. As the YouTube comments corpus has not been manually POS tagged, it was not possible to perform and comprehensive statistical analyses of the results in order to assess the accuracy of the two techniques. I was able, however, to collect instances in which the two algorithms differed in their results. The HMM and perceptron only agreed on 527 comments out of 10,000, and manual sampling of the comments in which the results differed revealed that the HMM tagger was misclassifying the majority of tokens as plural proper nouns (NNP). It is difficult to identify the reason why this may be, however for the purposes of this case study it is clear that the perceptron is the superior algorithm to use. This does not mean that HMMs should be discounted for future, just that within the context of employing the NLTK package it is preferable to use their recommended perceptron tagger.

2.4 b) Practical application of the NLTK tagger

It is important to reiterate at this point that we are making use of POS tagging in an approach that mirrors that of Ferrara et al. (2013). Namely, we are incorporating extra-linguistic information into the comparative analysis of the comments in order to better assess whether we can classify the comments as belonging to a wider meme-cluster. In more practical terms, what this means is that as we are iterating through comments and collecting the non-stop words in them, we are also going to be noting the syntactic structure of the phrase. As an example of why this might help with the identification of memes, let us return to the meme of ‘Alexa, play Despacito’. Given that the name of the song is in the meme (and will likely occur in many thousands or other comments) it is logical to assume that solely a keyword approach is not going to capture the full scope of instances of this meme. Traditional text classification techniques often rely on the fact that lengthier documents and paragraphs reduce the ambiguity of the ideas expressed by the author through the context surrounding them in the text (Chen et al., 2019). This is not a luxury we are afforded when processing YouTube comments that, on average, number around 10 words in length. Thus, in the short phrase “Alexa, play Despacito”, a minor change of object from “Despacito” to “Taki Taki” would result in a 33% reduction in similarity for a BOW-based algorithm. To expand, it is clear to a human observer that “Alexa, play Despacito” and “Alexa, play Taki Taki” belong to the same ‘class’ of comments. However, if a cosine similarity measure was carried out between the two, it would report that there was a score of 0.66 between the two which would likely be below a threshold for clustering. However, the syntactical POS information for the two strings (proper noun, verb, proper noun) is a 100% match. Consequently, in order to capture this syntactic information, as we are iterating through the comments calculating cosine similarities, we will also collect information pertaining to the longest common sequence of POS tags between the two strings.

2.4 c) Optimisation of iterations

As discussed in section 2.3, a great concern of this methodology is efficiency. We have greatly increased the speed of iteration by reducing redundancies and multiprocessing the script. Furthermore, we were able to increase the how quickly we could calculate cosine efficiency by employing packages that streamline Python code into more efficient, lower-level code. Therefore, it is important that we also make the identification of the longest common sequence as efficient as possible so as to not counteract our previous optimisations. The principal way in which this was achieved was less through specific changes in how the algorithm was implemented (although it still remained an important aspect of the methodology), but rather through the broader conceptualisation of what we were actually aiming to achieve with this process. To expand, it was necessary that I calibrated the granularity of the data processing to that of the source data itself; the POS tags are an abstraction of the comments and therefore we need only return a lower resolution result, that is to say a Boolean true or false as to whether there was a common POS sequence between the two strings.

The term ‘longest common subsequence’ is somewhat of a misnomer. In computer science terms, what we are actually referring to is the ‘longest common substring’ problem which is the challenge of locating the longest consecutive sequence of characters (Arnold and Ohlebusch, 2011). This differs from ‘longest common subsequence’ problem which is in principal the same as the substring problem, however the characters do not have to be in consecutive order (Hirschberg, 1977). Thus far in this thesis I have used string to refer to textual information only, and so to maintain consistency I will continue to refer to what we are trying to find as the ‘longest common sequence’. Much of the methodologies in computer science, including those in pre-existing packages, are designed in order to cope with strings that are millions, or even billions, of characters long, and are incredibly efficient at doing so. However, what we wish to do is compute the longest common sequence between two very short sequences of POS tags millions of times. Consequently, making use of code that employs this higher level of computer science theory actually increased the estimated runtime between 300% and 500% as there is a large amount of overhead that is incredibly useful for enormous strings but not so much with shorter strings. Additionally, it is important to consider the level of granularity, or lack thereof, of using POS tags specifically for identifying memes. When we are processing these comments as their constituent parts-of-speech, what we are examining is the relationships between the elements of the comments regardless of any semantic or social context. Furthermore, we are doing so in order to identify repeated patterns that may correlate to memes. There is thus a necessarily large amount of generalisation happening in this process, and it is important to scale our collected data to this level generalisation. To expand, we need not process the comments in such a way that we are returned very high-resolution data such as the percentage of the part-of-speech sequences between two comments that match. This is because basing our analyses on generalised methodologies used to generate more specific data than they were designed to will provide ample opportunity for false positives to occur. An example may be a comment that simply reads "Alexa, play Despacito" being compared with a much longer string, one small part of which is the meme 'Alexa, play Despacito'. Were we to employ an algorithm that returns the percentage match between the two strings, we would likely discount these two comments as belonging to the same set of comments – those that contain the 'Alexa, play Despacito' meme. Naturally this is counterintuitive to the aims of this chapter, so we need to design the algorithm with a level of built-in generalisation. As stated earlier, I elected to do so with a simple Boolean true or false as to whether both strings contained a matching POS sequence.

Just because we are returning less-specific results for this section of the methodology means that we need not take steps to prevent over-generalisation. Within the practical implementation of the algorithm, I elected to exclude those comments that were less than three tokens in length and subsequently increase the minimum length of POS pattern to be matched from two tokens to three. There were two advantages to doing this: 1) We vastly reduced the potential for false positives, and 2) We decreased overall run time for the algorithm. The reason why we reduce the number of false positives is logical; from a probability standpoint there are more potential bigram matches between two strings than trigram. Furthermore, it is unlikely that we would collect any valuable data regarding the presence of a similar meme with pairs of POS tags as there are likely to be multiple instances of POS pairs such as "Noun-Verb" or "Adjective-Noun". This is not to say that there are not common structures present as trigrams (e.g., "Noun-Verb-Noun"), however by excluding a relatively small portion of the corpus, we are able to vastly reduce the number of potential false positives. The reason why we are afforded the second advantage stems from the basic algorithmic behaviour of this step in the process. A highly algorithmically efficient manner of iteratively comparing comment A and comment B is by creating a function in which a data structure that makes use of hash tables is created and to which we add all the trigrams in comment A as strings. The reason why a structure that uses a hash table is more efficient is because it allows us to search the structure for a particular key almost instantaneously[20] as opposed to physically traversing the structure in memory and checking each entry to see if it matches our input (Gorelick and Ozsvald, 2020). Consequently, when we loop through comment B as ‘rolling window’ of trigrams we can check each of these trigrams incredibly rapidly. With this in mind, it is clear why increasing the minimum POS sequence to three tags would decrease overall run time – there are simple less values to iterate through.

Thus far we have established the broad methodological framework of this chapter: for each comment in the corpus, we wish to compare it with every other comment for both cosine similarity and presence of a shared subsequence of part-of-speech tags. The algorithms for calculating the cosine similarity and identifying the sequences are themselves fairly rapid. Were they taken in isolation without any of the Pythonic mechanisms for accessing comments and passing them to these functions, we would be able to compare the entire 1.7 million comments on ‘Despacito’ in just a few hours. However, as a language, Python is particularly slow in its implementation of iterators such as those that we are employing here; making use of another programming language such as Go or C can lead to an increase in speedup of upwards of 500 times (Mamaev, 2018). Making use of Pythonic iterators to loop through an entire corpus once again puts the total run time of this script in the hundreds of days. The following section will describe how will a simple reconceptualisation of how we approach the problem will once again allow us to process cosine similarity and POS tags with Python and NumPy in a timely fashion.

2.5) Optimisation of Python: Efficient use of NumPy arrays and hash tables

The principal bottleneck that we are facing thus far is the fact that Python is prohibitively slow at iterating through sets of data in order to perform functions on them (Mamaev, 2018). This section will outline how we can make use of NumPy universal functions and Python hash maps in order to

[20] In computer science, the actual time it takes to look up a key in a hash table is described as O(1) time and is typically contrasted to O(n) time.

calculate our results, not on a comment-by-comment basis, but by operating on the data structure itself.

2.5 a) Using NumPy universal functions to calculate cosine similarity

We have established that storing and parsing our data with the use of NumPy arrays is more efficient for mathematical computation such as that required for cosine similarity. At present, we are iterating through the entire corpus of comments, and in each loop once again iterating from the first to last comment. This brings the total number of loops we have to complete to n^2, which for a corpus of just 100,000 would result in 10,000,000,000 (ten billion) separate iterations. By ignoring redundant calculations, we need only complete $(n^2/2)-n$ values, but computing 4,999, 900,000 values is still a prohibitively long. The solution lies with the vectorisation of NumPy arrays. Solomon classifies the process of vectorisation as the "ability within NumPy to express operations as occurring on entire arrays rather than their individual elements" (2018). To use an example, if we wished to double every element in a matrix with a pure Python implementation then it would involve manually indexing each element in a loop and then multiplying it by two. With the use of NumPy vectorisation, we can "delegate the looping internally to highly optimized C and Fortran functions", thereby performing the operation several orders of magnitude more quickly (*ibid.*). This extreme increase in speed does come at a cost, however: the fact that we are able to perform calculations on large arrays as a whole necessarily means the output takes up n^2 amount of memory to store. The GitHub user Denziloe provides a rapid solution to this memory problem by processing the word-frequency matrix in chunks (2018), which I modified to dump the results (in this case the indices relating to comments in a list of those values over a threshold of 0.5[21]) to file after every chunk. The fact that we were iterating over the chunks of data and also dumping to file contributed greatly to

[21] Note, this value was chosen semi-arbitrarily but roughly equates to a 50% match of tokens in strings of equal lengths. Naturally, this does not apply to string of greatly different lengths, however tests of randomly sampled comments were carried out and a score of 0.5 was roughly around one standard deviate above the mean cosine similarity of strings (zero-values excluded)

the run-time of the script, but this script was able to process 20,000 comments in this was in a matter of seconds – a vast increase over the previous method that employed Pythonic iterations.

In this particular case, one may wish to use cloud computing resources to process this script. Vectorised NumPy functions have multiprocessing built-in to the framework and given the speed of the script and that most resources are rented by the hour, it would not be financially prohibitive. It is self-evident that more processing power will increase the total time it takes to run the script, but it is especially true in this case. The more RAM one has available, the larger the chunks one can process (or not have to chunk the data at all), and the fewer loops and data dumps one must carry out. I only had 8GB of RAM and a CPU several years out of date. Consequently, I could process a maximum chunk-size of 100 comments at a time, and it was also necessary to increase the likes-threshold to only process 2.5% of the corpus. As discussed earlier, thresholding in this manner means that we must pick comments from discrete categories (the likes that they have received) up to a continuous threshold (2.5% of the total). In order to attain the requisite number of comments from each video, random sampling was employed if the quantity of liked comments was insufficient to reach 2.5% of the total comments. Naturally, this may impede on the reproducibility of the results and also potentially incorporate some statistical skew, but it is unlikely that a few hundred randomly selected comments will impact a corpus size numbering in the tens of thousands in any significant way. Additionally, all comments were sampled from a sub-corpus comprised of comments that had received at least one like. Consequently, we can be assured that all comments have at the least undergone some level of natural selection by the public. Nonetheless, this is a factor that will be carefully accounted for in the results-gathering stage of this case study.

2.5 b) Making use of hash tables for efficient processing of POS tags

In the original script, as we were iterating through the corpus, we would calculate the cosine similarity and then test whether there was a common subsequence between the two comments. If the comment pair passed both checks then it would be added to a database from which we could generate results. Now that we had disposed of nested Pythonic iterations for calculating cosine similarity, we were no longer able to concurrently test for common sequences. Instead, it was necessary to reconceptualise the way we approached locating common sequences and instead process the data as one discreet structure.

Instead of performing pairwise comparisons of the comments, instead I elected to simply iterate through the corpus once and add all the trigrams of POS tags to a Python data structure called a dictionary. In this context, a dictionary is a pair of hash mapped keys and values. For the purposes of this script the keys were the POS trigrams and the values of these trigrams was a set of all the comments in which this trigram occurred. As we were only performing a single loop through the corpus, this process was measured in the seconds; we still needed to parse both the POS tags data and the cosine similarities together in order to generate our results, and this is where the bulk of the processing time took place.

2.5 c) Collating and processing results

Given that we are performing this project with Python, it was inevitable that we would have to perform a process of lengthy iteration through our data in order to parse it. Thanks to the steps taken above, however, this was an incredibly streamlined process compared to what we were contending with beforehand. We had already created an array comprised solely of pairs of comments with a cosine similarity of above 0.5, and now it was simply a case of iterating through the cosine array and also the POS dictionary. As we were traversing the dictionary of POS tags, we would check the associated value (comprised of a list of all the comments containing that tag) for the presence of both elements of the cosine pair we were currently examining. This was not as prohibitively long a process as may first seem for three reasons: 1) POS tags are a generalisation of textual data; 2) common POS sequences were added to the dictionary first due the iterative nature it was filled; and 3) we stored the list of comments associated with the POS trigram key as a hash map. POS tags are an abstraction of textual data and consequently there are a great deal fewer POS tag combinations than real-word trigrams. Therefore, it is not necessary that we must loop through a prohibitively long list of POS trigrams – for the 42,418 comments collected from ‘Despacito’ the length of the dictionary of POS trigrams was only 9264 values long. Furthermore, as the comments were added (either to an existing entry or by creating a new one) to the POS keys in the dictionary as they were encountered in the corpus, it is logical that the most common sequences would be found nearer the beginning of the structure. This means that while iterating through the POS dictionary has the potential to take up O(n) time (i.e., an amount of time proportional to the length of the data structure), in reality, it was much quicker as we would typically encounter a POS tag with both the comments we were examining much earlier in the dictionary. Finally, we stored the list of comments for each POS tag as a hash map so we were able to look up the comment pairs in O(1) time. The overall run-time for this portion of the process was roughly four hours for the entire 77828 comments on ‘Gangnam Style’ – the largest corpus of comments. This is a vast improvement on previous Pythonic iterative methods and also will not scale as exponentially as when we were performing n^2 comparisons in the corpus.

The end result of this was a large array containing all the pairs of comments that both had above a cosine similarity score of 0.5 and also contained a common sequence of part-of-speech tags. From here it was trivial to parse this data and find clusters of comments by following the chains of links between them. With this, we hope to elucidate the extent to which memes play a role in the way the public receives foreign language music.

3) Discussion of results

Examining the role of the meme in the YouTube space became a two-step process, the first being the construction of our own methodological framework to this end, which we have detailed above. This section will be the next step in the process and will analyse the data yielded from our methodology

3.1) Preliminary analysis

3.1a) Statistical data

Table 5.2: Table showing the percentage of comments that were filtered from the raw data

Song name	Percentage of comments filtered
Gangnam Style	76.03%
Gentleman	52.46%
Despacito	80.36%
Mi Gente	48.56%

The output of our methodology was a list of comment pairs that both were above a cosine similarity threshold and contained a common part-of-speech sequence. By calculating a simple percentage of the comments that passed both of these checks, we will be able to draw some cursory conclusions that will provide a basis for the analyses to come. From the data in table 5.2, we can conclude that there is a clear disparity here between viral and non-viral songs. 'Gangnam Style' and 'Despacito' (viral) have a higher proportion of comments that passed the checks we outlined in section 2 than 'Gentleman' and 'Mi Gente' (non-viral). This implies that comments on their music videos tend to be more highly correlated with each other in terms of syntax and content and supports the hypothesis that repeated memes form a large part of the way that commenters interact with these cultural artifacts. As discussed in section 2, when the sub-corpus of comments above a given threshold of likes did not contain enough comments to reach the pre-decided proportion of 2.6%, random sampling was undertaken to pad out the data. The likelihood that this sampling is contributing to the disparity observed by diluting the pool of comments is relatively miniscule. A few hundred (or thousand depending on the corpus) comments are unlikely to influence the results greatly, however it is important to examine all the ways in which our data may be open to skew. In this case, it is more probable that the methodology proposed in this chapter has worked as intended and we can logically conclude instead that the corpora for 'Despacito' and 'Gangnam Style' present a great deal more interconnectedness. The question now remains as to whether this similarity between comments is as a result of the propagation of memes throughout the comments section, or whether there is another explanation entirely. This will be the focus of the following sections.

3.1b) Network graphs: Micro- and mezzo-level structures

One way to visualise the data produced by our methodology is as a network of comments linked by similar comments and syntactic structures. Figures 5.3 - 5.6 are network graphs visualising these relationships[22]. Whilst the data for the graphs were generated with the NetworkX Python package (Hagberg et al., 2008), as per recommendations by the authors of this package, the visualisations were generated with Gephi (Bastian et al., 2009). Gephi is an open-source tool that facilitates the visualisation and manipulation of network graphs such as our own. The reason why it is advisable to use an external graphing tool for network graphs is that those native to Python are maladapted for the complex task of drawing this kind of visualisation. Another advantage is that the use of open-source tools permits those researchers not proficient in Python to make use the .gexf files associated with the data we have generated. Simply by installing Gephi and accessing my repository, future scholars will be able to analyse the data to their own ends without the need to learn Python or any of the associated packages that I have employed throughout the project.

[22] It should be noted that the colourings of the community are arbitrary and are present purely to clarify the noisy data present in these graphs. However, we do employ the Leiden algorithm to generate these colourings which will be discussed in greater detail below.

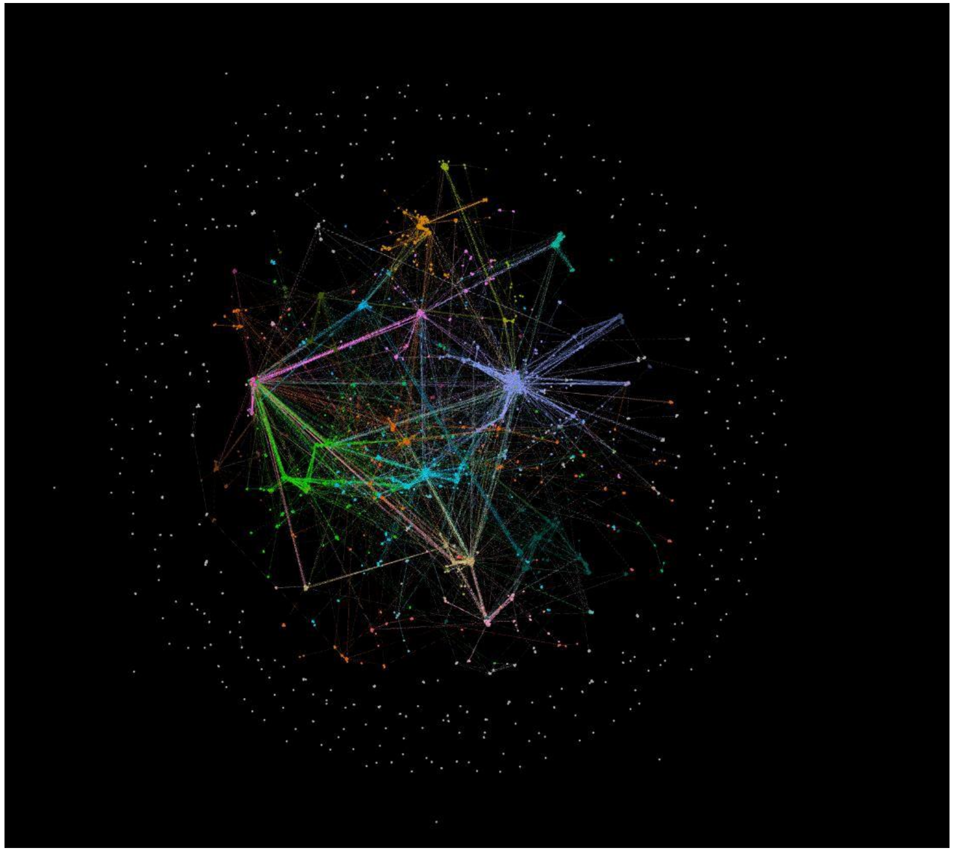


Figure 5.3: Network diagram of comments in ‘Gentleman'. Made with Gephi (Bastian et al., 2009).

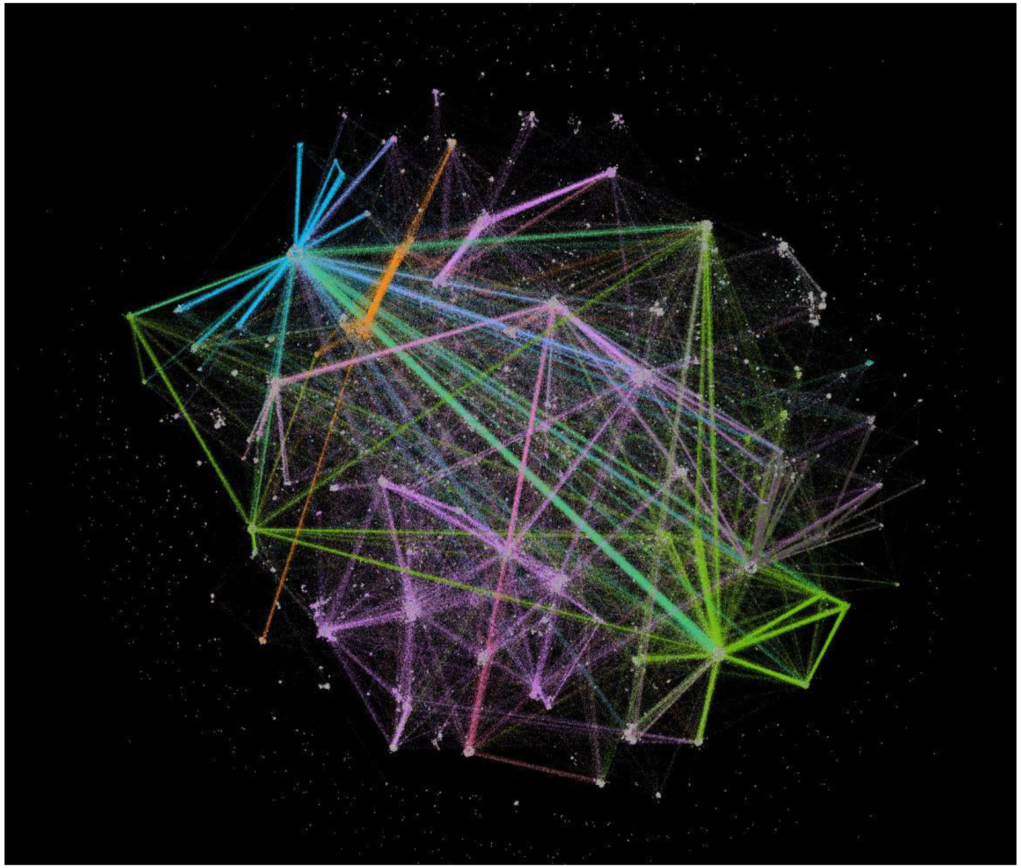


Figure 5.4: Network diagram of comments in 'Gangnam Style'. Made with Gephi (Bastian et al., 2009)

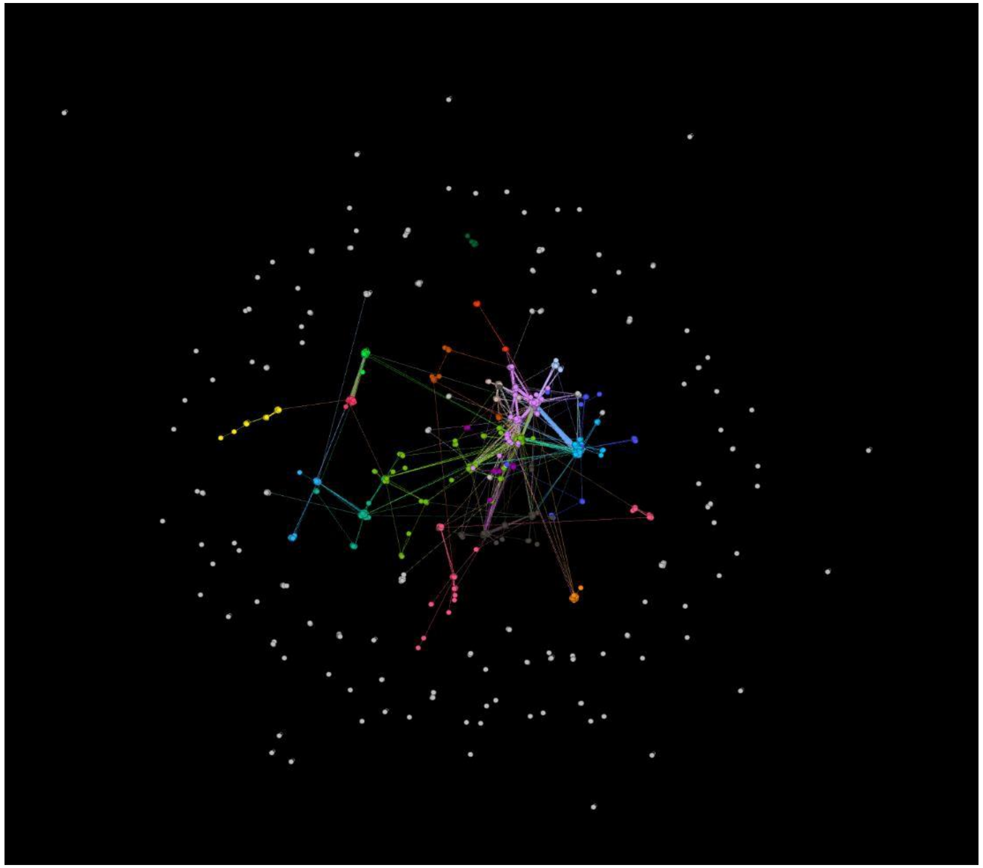


Figure 5.5: Network diagram of comments in 'Mi Gente'. Made with Gephi (Bastian et al., 2009).

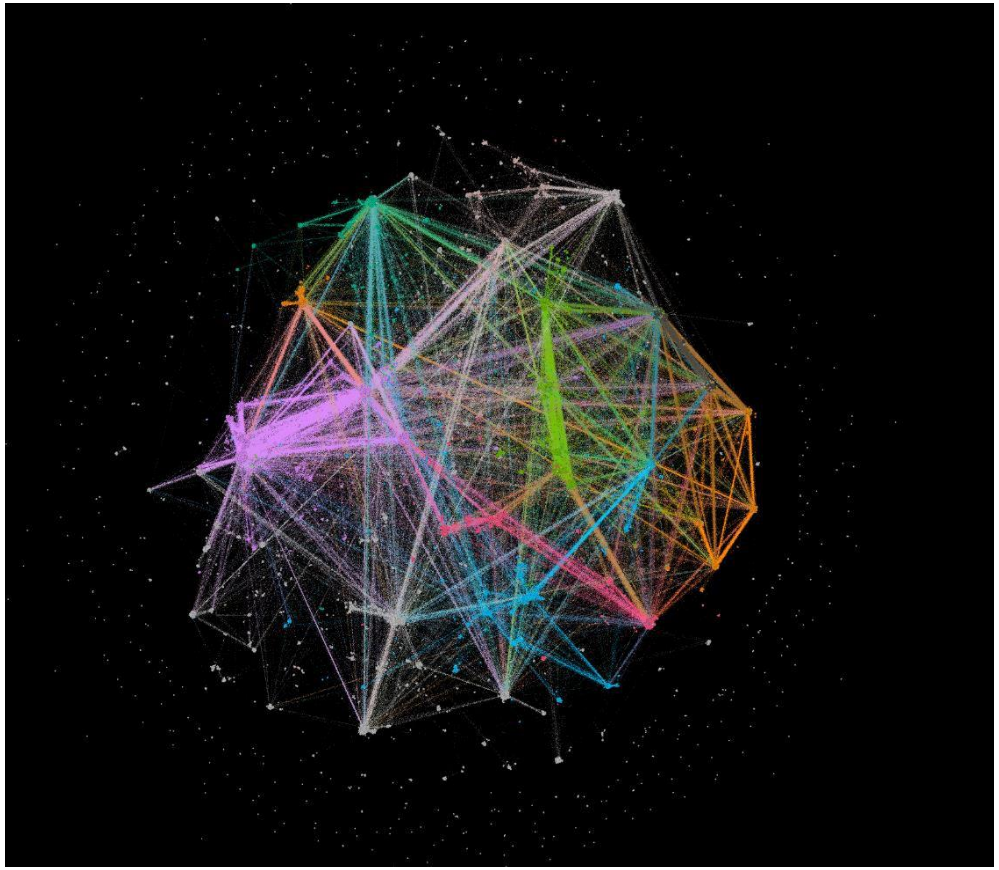


Figure 5.6: Network diagram of comments in 'Despacito'. Made with Gephi (Bastian et al., 2009).

Table 5.3: Table of average degrees and number of nodes in each song

Song	Average degree	Number of nodes
Gangnam Style	95.00	58,744
Gentleman	13.78	7156
Despacito	59.00	33,593
Mi Gente	11.37	1408

Table 5.3 displays the average degree for each of the networked comments corpora visualised in figures 5.3-5.6. In this context, the term 'degree' refers to the number of edges attached to each node. Consequently, table 5.3 is displaying the average number of edges per node across the entire graph. Cursory, visual analysis of the graphs fully supports the data in table 5.3.; 'Gangnam Style' is a densely populated graph with 58,744 well-connected, separate nodes (comments) whilst the network corresponding to 'Mi Gente' only contains 1408 nodes. The distribution of results follows a similar pattern for 'Despacito' and 'Gentleman'. This disparity is highly logical: if a network such as ours contains more comments, then it follows that there will be a greater number of connections between these nodes. However, this data only informs us of the micro-level structures in our network. The term micro-level and its comparative mezzo- and macro-levels are described by Soundarajan et al. as three levels of specificity one can operate at when examining networks (2014). A micro-level method will extract features from the network at the level of the node, the mezzo- from communities, and finally the macro- from the entire network. We wish to concern ourselves with the mezzo-level of analysis as we are most interested in communities of nodes, i.e., clusters of nodes within larger networks that correspond to comments our methodology has deemed similar. Figure 5.7 is an example of one such cluster in the 'Despacito' network. In this instance, it is the cluster that corresponds to the "Alexa, play Despacito" meme (or at the least those variations identified by our methodology). The following will be an analysis of the different communities such as this across the four songs that we have targeted in this chapter.

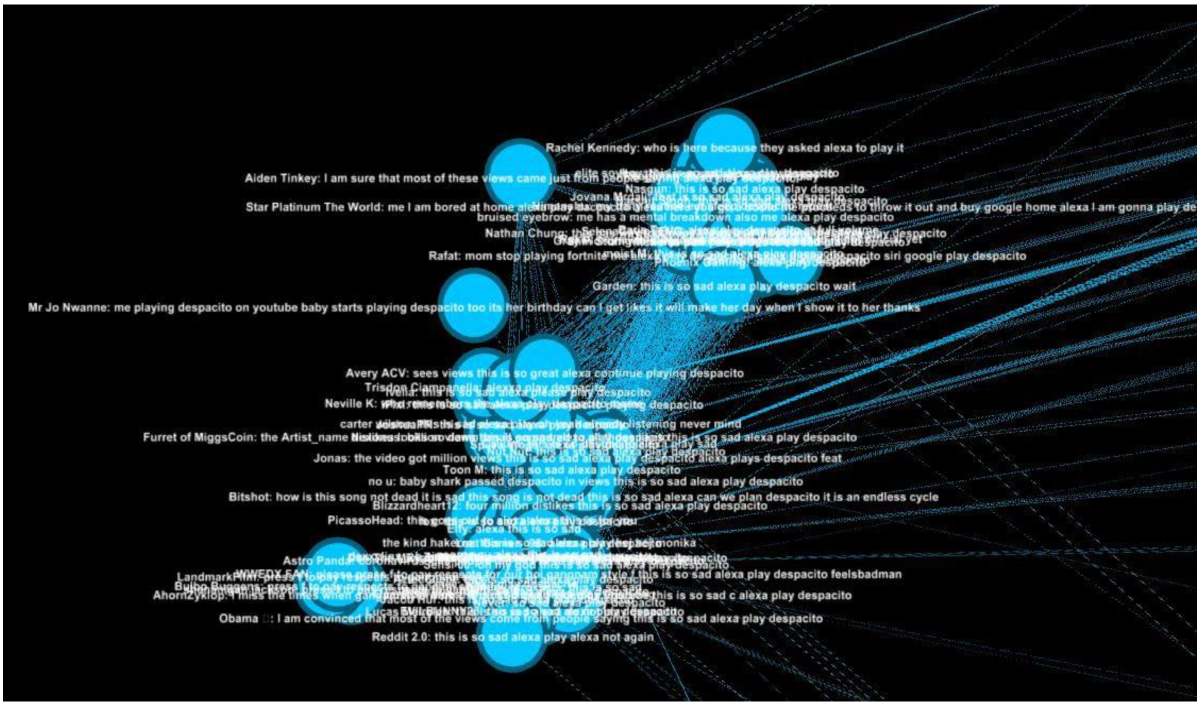

Rachel Kennedy: who is here because they asked alexa to play it
Aiden Tinkey: I am sure that most of these views came just from people saying alexa play despacito
Nasgun: this is so sad alexa play despacito
Star Platinum The World: me I am bored at home
to throw it out and buy google home alexa I am gonna play de
bruised eyebrow: me has a mental breakdown also me alexa play despacito
Nathan Chung:
play despacito
Rafat: mom stop playing fortnite
siri google play despacito
Phoenix Gaming: alexa play despacito
Garden: this is so sad alexa play despacito wait
Mr Jo Nwanne: me playing despacito on youtube baby starts playing despacito too its her birthday can I get likes it will make her day when I show it to her thanks
Avery ACV: sees views this is so great alexa continue playing despacito
Trisdon Ciampanella: alexa play despacito
Neville K:
carter
listening never mind
Furret of MiggsCoin: the Artist_name
this is so sad alexa play despacito
Jonas: the video got million views this is so sad alexa play despacito alexa plays despacito feat
Toon M: this is so sad alexa play despacito
no u: baby shark passed despacito in views this is so sad alexa play despacito
Bitshot: how is this song not dead it is sad this song is not dead this is so sad alexa can we plan despacito it is an endless cycle
Blizzardheart12: four million dislikes this is so sad alexa play despacito
PicassoHead:
Elfy: alexa this is so sad
the kind hake
monika
Astro Panda
Sensi00: oh my god this is so sad alexa play despacito
LandmarkPiml
WWEDX FAN
this is so sad alexa play despacito feelsbadman
Bulbo Buggans: press f to pay respects
AhornZyklop: I miss the times when
this is so sad c alexa play despacito
Lucas
Obama : I am convinced that most of the views come from people saying this is so sad alexa play despacito
Reddit 2.0: this is so sad alexa play alexa not again

Figure 5.7: Cluster of comments containing the "Alexa, play Despacito"

3.2) Identifying and analysing clusters

Rossetti et al. contend that discovering communities in networks is amongst the most researched problems in network analysis (2019). In essence, the goal of discovering communities is to decompose a complex network into meaningful clusters of nodes. In highly connected networks such as ours, the implementation of a robust algorithm is vital for accurately identifying such communities. Gephi provides algorithms to this end with the figures thus far having been generated with OpenOrd, an algorithm designed to elucidate clusters in very large graphs (Martin et al., 2011). As helpful as this implementation is for cursory, macro-level analysis, we require more finely tuned tools if we collate any meaningful, statistical data regarding the results we have generated. Rossetti et al. provide such a tool in the form of the Python package CDLib (Community Discovery Library). Built on NetworkX (the package used to generate the graph data passed to Gephi), this package allows us to identify communities from the raw data so we may perform statistical analyses thereon.

CDLib provides a number of different algorithms for the detection of communities in network graphs. Traag et al. contend that the Louvain algorithm is one of the highest rated in community detection, however it contains a defect that means that crucial connections can be overlooked (2019). In order to combat this defect, they propose the Leiden algorithm, which employs a similar methodology but without the same defect as the Louvain. Table 5.4 shows the number of communities identified with the Leiden algorithm, the number of nodes within these communities, the largest of these communities, and the average number of nodes across the communities in each of the four songs we are examining. For the latter, two values have been provided – the first is the average community size with all data included whereas the second is the average calculated only from communities that contain three or more nodes. By creating this threshold, we are providing ourselves with a more holistic picture of the data that excludes the potential influence of skew on the figures.

Table 5.4: Table showing the number of nodes, number of communities, average community, and largest community across the four songs examined in this case study.

Song	# of nodes	# of communities	Avg. community size (with/without thresholded data)	Largest community
Gangnam Style	58,744	767	76.59 / 286.80	8670
Gentleman	7156	384	18.65 / 52.74	853
Despacito	33,593	498	67.46 / 186.34	4963
Mi Gente	1408	132	10.67 / 20.95	165

Broadly speaking, the results in table 5.4 reflect what we would expect to see from the relative sizes of the corpora to which they pertain. From largest to smallest, the corpora are ordered as follows: 'Gangnam Style', 'Despacito', 'Gentleman', and 'Mi Gente'. The distributions of data across all metrics follows this pattern exactly, however not to the degree that we would expect. The 'Gangnam Style' corpus is roughly 8x larger than that of 'Gentleman' and it logically follows that we would expect 8x more communities to be present in the former. However, we only see approximately 2x more communities amongst in the corpus of 'Gangnam Style'. This disparity is even more pronounced in the comparison between 'Despacito' and 'Mi Gente' with the former containing 24x more nodes and yet only 4x more communities. The reason for this stark difference is hinted it by the average and largest community sizes.

Table 5.5: Table showing ratio between values in table 5.4 for song pairs

Song pair	Avg. community size ratio	Avg. community size ratio (stopped)	Largest community ratio
Gangnam Style and Gentleman	1:3	1:4	1:9
Despacito and Mi Gente	1:5	1:7	1:29

From table 5.5, we can draw two clear conclusions as to why 'Gangnam Style' and 'Despacito' present such different values for number of communities than what we would predict based off the corpus sizes alone. 1: The ratios between the sizes of the largest communities much more closely mirrors the difference in corpus sizes as would be predicted. 2: The fact that we only see a minor increase in the ratios when we remove clusters under 3 nodes in size implies that the size of the largest community for each song is not anomalous. In fact, this minor increase shows that there is a more uniform distribution at the top end of the spectrum. The combination of these two facts informs us that the corpora for 'Despacito' and 'Gangnam Style' contain several highly populous clusters of similar comments. Given the methodological steps we undertook to generate this data, there is a high probability that these clusters do in fact contain memes. However, in order to properly test this hypothesis, we will need to perform manual, qualitative analysis on the clusters.

3.3) Analysis of communities

We have computationally identified the presence of several large clusters of comments in 'Despacito' and 'Gangnam Style'. It now falls to me, the human researcher, to perform my own, manual analysis of these clusters to determine whether the content of the communities qualify as memes. Table 5.5 and 5.6 contain the most common keywords with example comments of the 25 largest communities in descending size-order.

Table 5.6: Table depicting the 25 largest communities in 'Gangnam Style' along with keyword frequency information and example comments.

#	Community size	5 most frequent tokens	Examples	Meme?
1	8670	['song', 'kpop', 'love', 'old', 'years']	"let us all agree that even the kpop haters love this song", "it is scary to think that this song is now NUMBER years old"	N
2	7039	['billion', 'views', 'video', 'people', 'world']	"how can this video have billion views if there are only billion people in the world"	N
3	5309	['like', 'comment', 'video', 'song', 'watch']	"how many likes can this comment get lets see like the comment its just for fun"	N
4	4391	['still', 'know', 'listening', 'song', 'years']	"if you are still listening to this song years later your a legend"	N
5	4271	['Artist_name', 'watch', 'video', 'views', 'like']	"better watch out Artist_name I am NUMBER views away from you", "the fact that most kpop haters like this song and other Artist_name songs when he is an idol and ceo of kpop agency makes me love this man even more"	N
6	3971	['style', 'gangnam', 'despacito', 'views', 'see']	"gangnam style see you again despacito billion in months the race to billion views is on"	N
7	3172	['views', 'see', 'came', 'many', 'video']	"who came here to see how many views it had in"	N
8	2585	['watching', 'else', 'like', 'video', 'today']	"who else is watching this today"	N
9	2245	['video', 'youtube', 'viewed', 'music', 'views']	"congrats you have the most viewed video on youtube with billion views"	N
10	2037	['comments', 'comment', 'find', 'million', 'legend']	"if you find this comment your a legend"	N
11	1991	['sub', 'channel', 'please', 'subscribe', 'back']	"please sub on my channel please"	N
12	1741	['likes', 'get', 'comment', 'reason', 'like']	"can I get likes for no reason"	N
13	1727	['check', 'views', 'came', 'view', 'else']	"who else came here in just to check the view count"	N
14	1452	['oh', 'god', 'guy', 'fortnite', 'girl']	"oh my god your in fortnite"	N
15	1445	['let', 'us', 'get', 'many', 'people']	"let us see how many people come here every day"	N
16	1021	['fuck', 'comments', 'top', 'shut', 'pen']	"the fuck is going on here with all the comments stahp"	N
17	851	['korean', 'korea', 'south', 'north', 'chinese']	"this is south korean not north korean north korea has laws against computers"	N
18	727	['thumbs', 'think', 'Artist_name', 'viewer', 'gay']	"thumbs up if Artist_name is gay", "thumbs up if you think I am gay"	Y
19	661	['randomly', 'came', 'anyone', 'mind', 'song']	"who came here because this randomly crossed your mind"	N
20	531	['bob', 'youtube', 'copy', 'paste', 'google']	"this is bob copy and paste him il all over youtube so he can take over youtube spread the word death to google bring the old youtube back"	Y
21	419	['views', 'many', 'checking', 'watching', 'comments']	"the only reason this has so many views is because of people checking for views"	N
22	374	['condom', 'style', 'open', 'op', 'store']	"it is who else is still listening to this and can still only hear open condom style"	N
23	230	['new', 'happy', 'year', 'internet', 'explorer']	"hi I am using internet explorer I hope this comment gets here on time happy new year"	Y
24	52	['step', 'go', 'search', 'key', 'enjoy']	"step go to google step search this key wixvi watch movies online step enjoy ee dd ee qq ff vv"	N
25	43	['lady', 'sexy', 'heyy', 'heey', 'ee']	"everyone: heyy sexy lady me NUMBER years oldn an intellectual: heyy pepsi lady"	Y

Table 5.7: Table depicting the 25 largest communities in 'Despacito' along with keyword frequency information and example comments.

#	Community size	5 most frequent tokens	Examples	Meme?
1	4963	['song', 'love', 'understand', 'like', 'spanish']	"I can not understand any spanish but I love this song it makes me feeling like amazing"	N
2	3217	['billion', 'views', 'despacito', 'song', 'people']	"Despacito: exists billion people: interesting", "youtube: do you want too see a song billion people: yes"	Y
3	3137	['despacito', 'views', 'see', 'vs', 'style']	"pewds vs t series despacito vs gangnam style"	N
4	2409	['like', 'listening', 'song', 'hit', 'despacito']	"who is listening this song in NUMBER if you are one of them hit like"	N
5	2119	['song', 'world', 'best', 'music', 'video']	"why tf has nearly the whole world watched this song like bruh", "most of the world has watched this music video"	N
6	1796	['Artist_name', 'like', 'despacito', 'better', 'si']	"who is better like Artist_name comment Justin Bieber"	N
7	1755	['watching', 'like', 'today', 'anyone', 'still']	"anyone else still watching this in December"	N
8	1722	['comment', 'like', 'people', 'likes', 'liked']	"the people who like this comment are here before the million likes"	N
9	1676	['views', 'check', 'came', '😂,', 'song']	"pov you just came to check the likes and views", "I am here to listen song not for check views"	Y
10	1559	['comments', 'find', 'comment', 'million', 'mine']	"about million comments if you find me then you are a legend"	N
11	1459	['views', 'people', 'see', 'come', 'song']	"this has so many views because of the people that come back so they can see how many views there are"	N
12	1120	['population', 'despacito', 'teacher', 'world', 'earth']	"teacher: what is population of earth Me: around one despacito xd"	Y
13	1060	['many', 'people', 'let', 'us', 'imagine']	"lets see how many people are listening to this masterpiece in February"	N
14	973	['video', 'youtube', 'viewed', 'despacito', 'views']	"anyone else randomly searched up most viewed video on youtube"	N
15	962	['get', 'likes', 'birthday', 'today', 'like']	"can I get a like it is my birthday"	N
16	535	['english', 'spanish', 'song', 'comments', 'puerto']	"Song: Spanish Me: indian Comment: English hotel trivago meme: overused comment: stolen"	Y
17	455	['first', 'fact', 'fun', 'watching', 'person']	"fun fact the first viewer must feel like the king of the world", fun fact if you put fun fact in a comment all people will read the whole thing	Y
18	368	['song', 'old', 'years', 'never', 'despacito']	"this song is literally years old but there are still comments with likes from week ago"	N
19	307	['one', 'nobody', 'literally', 'watch', 'comment']	"nobody: literally no one: me: let us just check despacito's views real quick"	Y
20	303	['spanish', 'learn', 'language', 'song', 'beautiful']	"I want to learn spanish now"	N
21	162	['happy', 'year', 'new', 'christmas', 'everyone']	"merry christmas and happy new year "	N
22	157	['live', 'may', 'years', 'parents', 'dear']	"to the person reading my comment in may your parents live years"	Y
23	88	['despacito', 'love', 'today', 'listen', 'confirmed']	"despacito confirmed by Bethesda", "despacito prequel trilogy confirmed by disney I can not wait"	Y
24	72	['video', 'water', 'music', 'income', 'whole']	"despacito is great but search on youtube for mc smook tanz mit mir the artist of this music video my boyfriend donates the whole income of all clicks to a solar water station in uganda do not be shy and search for the video click it like it comment it and spread it we can create a better world thank you"	N
25	70	['go', 'let', 'saad', 'lamjarred', 'coming']	"listen to saad lmjared let go the song very nice"	N

Based upon the data in these tables, I contend that the meme plays a much greater role in 'Despacito' than 'Gangnam' style. Only 3 out of 'Gangnam Style's' 25 communities contain a significant presence of what we can deem a meme. For 'Despacito', however, this number increases to 8, one of which being the second largest community. It would be useful at this point to provide some concrete examples as a way of explaining my qualitative rationale behind whether I have deemed a community to be comprised of a meme or not. The comments in Despacito's biggest community tend to express one of the language attitudes examined in chapter 5, To clarify, commenters tend to express that they enjoy the song in spite of a lack of comprehension through the following broad schema: "I do not understand the lyrics" – ADVERSATIVE CONJUNCTION – "I like this song". In contrast to this, the second largest community contained multiple instances of comments that follow the following broad format:

Entity a: *action*

Entity b: *reaction*

An example of this cited in table 5.6 is as follows:

"Despacito: exists

billion people: interesting"

In this case, the user is employing the format in order to remark on the enormous quantity of views that 'Despacito' had garnered with the use of humour. It is worth noting that this particular meme format is not unique to the second largest community. We can observe examples of this format in 4 of the 8 communities identified as meme-clusters. It is the presence of this template that principally informs my decision to classify these meme-clusters as such; as Marino states, a characteristic of memes in the current era, or more specifically that in which 'Despacito' was released, is their ability to be industrially produced (2015). With this 'cookie-cutter' template, memes are able to propagate throughout the comments with great ease. This is evidenced by the fact that we observe comments displaying this meme to be clustered chronologically in the community in which they are found. The implication this carries is that once this meme format was introduced, other users picked up and propagated it. We can additionally observe instances of another core feature of memes that Marino identifies: self-reflection (2015). In cluster 16, the following example was identified:

"Song: Spanish

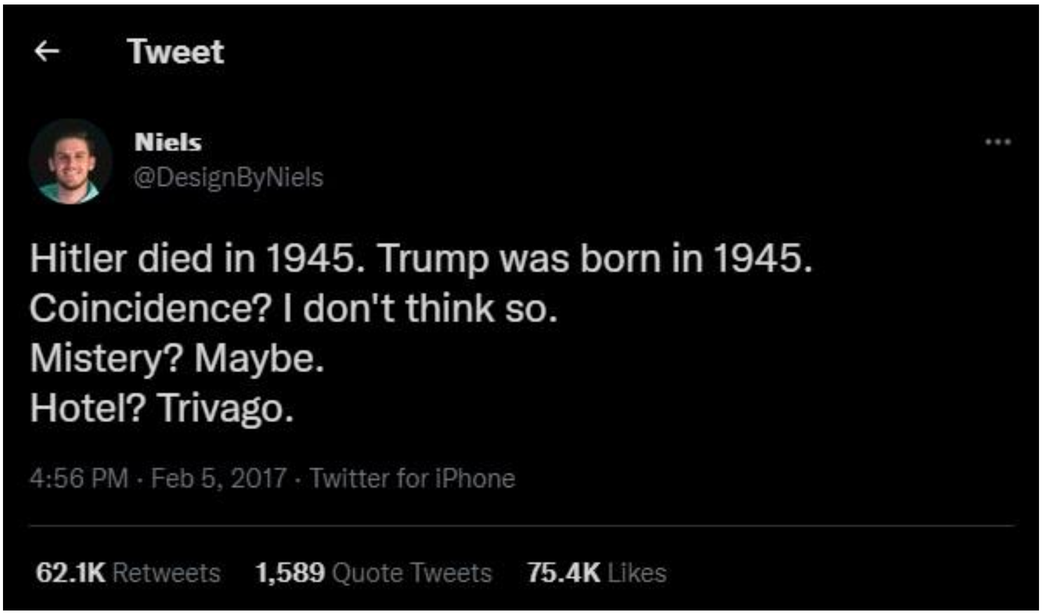
Me: indian

Comment: English

hotel trivago meme: overused

comment: stolen"

This comment meta-discursively examines the phenomenon of the 'Hotel Trivago' meme. According to 'Know Your Meme', a website dedicated to the archiving and explanation of internet memes, the 'Hotel Trivago' meme was formed as a parody of Hotel Trivago's slogan at the time (2021). The meme was principally popularised after the tweet displayed in figure 5.8 gained significant traction in February of 2017. Figure 5.9 highlights this popularisation by displaying the relative number of Google searches for the term "trivago meme" between 2015 and 2022 with early February presenting a spike in the frequency of searches for the term. The meme format we have identified with our clustering techniques presents a similar syntax to that of the tweet in figure 5.8. Given that 'Despacito' debuted at around the same time as this meme garnered popularity, we can logically presume that what we are observing here is a nascent meme evolving in the comments section of the video. It is worth noting that I elected to not collect the exact time that a given comment was posted when creating the corpus to reduce the noisiness of the data. However, I have re-downloaded a sample of songs with the inclusion of this metric and discovered that the majority of the comments are posted shortly after a video is published. This further supports the conclusion that instances of the meme template in question are an evolution of the ‘Hotel Trivago meme’.



Tweet

Niels
@DesignByNiels

Hitler died in 1945. Trump was born in 1945.
Coincidence? I don't think so.
Mistery? Maybe.
Hotel? Trivago.

4:56 PM · Feb 5, 2017 · Twitter for iPhone

62.1K Retweets 1,589 Quote Tweets 75.4K Likes

Figure 5.8: Image of the tweet that popularised the ‘Hotel Trivago’ meme.

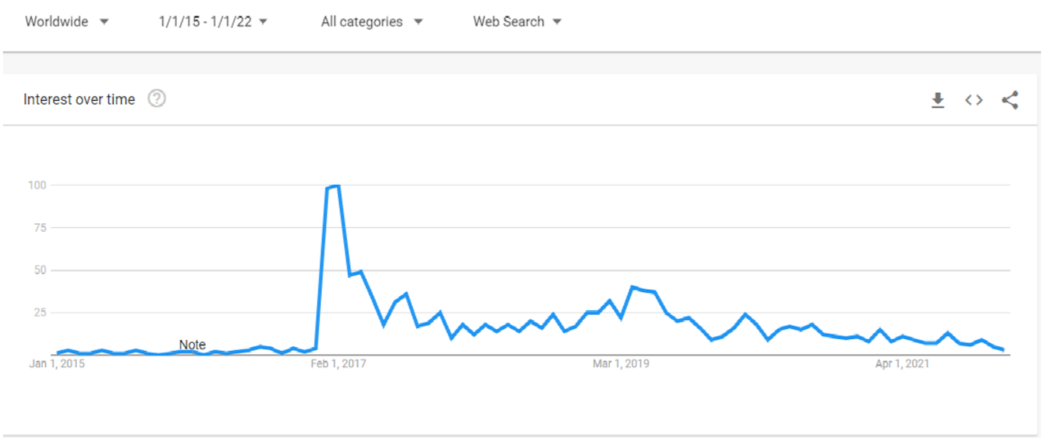


Figure 5.9: Graph displaying the Google searches for “trivago meme” since 2015 (Google, 2022)

This is an example of the type of analysis that was conducted to assess whether a given community could be classified as being comprised principally of memes. Although there would lie value in specifically outlining the ways in which each community does or does not fulfil the criteria of meme, the constraints of the project mean we are unable to do so. As evidenced by the example above, however, the main criteria as to whether we can classify these comments as memes is whether there is an element of reproducibility about them. The question that we now must consider is how what we have uncovered helps us fulfil the aim of the chapter, that is to say, to explore the role of the meme in the way in which the public receives these songs on YouTube.

4) Conclusion

4.1) The meme and viral foreign language songs

We have established that both 'Despacito' and 'Gangnam Style' present much more deeply connected corpora than those of 'Mi Gente' and 'Gentleman'. Taking this further, we identified that memes were more prevalent in the corpus of 'Despacito', although with the caveat that it seems to be a particular format of meme stemming from the Tweet displayed in figure 5.8. In answer to the question of the role that memes play in the way the English-speaking public receives these songs, I conclude that it is not particularly significant. Whilst there are some methodological decisions that may help to elucidate the matter further (see section 4.2), from the data we have collected meme-clusters appear to account for just over a fifth of the 25 most numerous in the comments of 'Despacito' and only an eighth of the 'Gangnam Style' corpus. Furthermore, there are perhaps broader socio-cultural factors that confound our ability to claim that the public chooses to engage with this video through the format of memes. As established in section 3.3, the majority of comments are posted within a short period of time of the video being published. With this in mind, it is possible that the (relatively) high prevalence of memes in the 'Despacito' corpus is due to the 'hotel trivago' meme appearing at around the same time. Had the two have been more temporally separated, it is entirely possible that we would not see as many memes in the corpus of 'Despacito'.

Whilst we have uncovered that memes are not a principal part of the way in which users receive these viral foreign language songs, we have been offered insight into what they are actually discussing in the comment sections of these songs. It appears that the majority of well-liked comments tend to preoccupy themselves with the popularity of the video. Whether it is remarking on the staggering number of views it has received ("why tf has nearly the whole world watched this song like bruh") or seemingly trying to profit from the popularity by asking for likes or subscribers ("can I get likes for no reason"), users will often post comments centred around this theme. Users also seemingly employ the comment sections of these viral videos as a form of social networking site. In chapter 1, we noted Bou-Franch et al.'s contention that YouTube presents a space for unique interpersonal interaction (2012). They describe the kind of communication that YouTube facilitates as polylogal; that is to say, it allows one-to-many interaction as opposed to the one-to-one form of communication posed by other sites. Perhaps due to the large quantity of visitors the page is bound to receive, commenters will often try to initial polylogal interaction with comments such as "who else is watching this today" or "the people who like this comment are here before the million likes". With this in mind, it is possible that, somewhat counterintuitively, the comment sections of very popular videos are not an ideal space for the examination of the interaction between memes and reception of songs and this brings us to the question of what future research may wish to accomplish in light of the work carried out in this chapter.

3.5) Future research

As digital humanists, we move from macro-level analyses to closer, oftentimes qualitative reading of texts (Honn, 2014). With the use of large datasets, we can employ complex methodologies to draw nuanced conclusions that previously would been out of reach. However, oftentimes these methodologies are the culmination of many minor decisions, and consequently future research may wish to alter the choices I made when implementing the methodology that I outlined in this chapter.

We can see that, whilst the POS tagging did play a role in the selection of grouped comments, it was mainly the content of the comments which informed the creation of the clusters. We know this because we can observe instances of the 'hotel trivago' meme in several of the communities in the corpus of 'Despacito'. Ideally, we would have grouped all instances of this meme-format into one group, and it is evident that a simple check as to whether there was a common subsequence of POS tags was not sufficient in this instance. Specifically for this meme, there is a distinct usage of punctuation that defines the format. Punctuation was not data that I elected to include when pre-processing the comments for the similarity-comparison methods as it can lead to noisy data, however, in this instance it may have been the key to identifying this particular thread of memes throughout the corpus. Broadly speaking, a more robust process for identifying and classifying POS tags would be beneficial for the general efficacy of the clusters. We established that for this project the principal limitation lay with the Python programming language, so other researchers who are proficient with other languages may be able to produce more specific results in this regard.

Another methodological decision that impacted the results was the choice of community identification algorithm we employed. We elected to make use of the Leiden algorithm, but, as Soundarajan et al. points out (2014), there are myriad different algorithms that may produce improved data. For example, although I was able to manually identify the "Alexa, play Despacito" meme in figure 5.6, there was no corresponding community in table 5.7. It is difficult to pinpoint which stage in the methodology exactly contributed to this phenomenon. It could be that our combined approach of cosine similarity and common POS tag sequence was not specific enough to separate out instances of this meme from other similar comments. More likely, it could be CDLib's implementation of the Leiden algorithm that resulted in the meme being obscured within another community (2019). Regardless, future researchers may wish to finely tune their network similarity method in order to either identify a flaw in previous methodological steps or acquire the correct resolution in order that these much smaller communities become salient.

Future research in Digital Cultures studies may also wish to further investigate the 'YouTube-as-social network' phenomenon that we observed within the corpus. Building on Bou-Franch et al. and Jones and Schieffelin's work, researchers may wish to further investigate the types of interactions present in the YouTube space (2012; 2009). A key component of this could be the examination of the ways in which commenters interact with each other on videos of differing popularity but are otherwise comparable in their socio-cultural characteristics, e.g., 'Despacito' and 'Mi Gente'.

Regardless of future researchers’ intentions, it is my belief that we have created in this chapter a flexible and powerful methodology that they may employ with ease to their own ends.

Conclusion

The principal aim of this project was to explore the reception of foreign language music in the English-speaking world. Through a range of digital techniques, we have employed three sub-corpora of songs as objects of investigation: K-pop, reggaeton, and viral songs. Within these categories, we have further narrowed our reading in order to elucidate different aspects of the way in which these foreign language songs are received. This general methodology has been an important foundational approach, which has allowed concrete insight into a range of aspects of the online reception of song in foreign languages, including the role that fandom plays within the differing communities of BLACKPINK and BTS. From this project, it is clear that there is much room for future research to be undertaken in the analysis of responses to digital cultures through YouTube comments and other digital, text-based communications. The methodology is easily adapted for analysing text in languages other than English; if, for example, one wished instead to examine how language attitudes are expressed within the space of French-language music on YouTube, it would be trivial to repurpose the methodology from chapter 4 to this end. With some additional development in collaboration with specialists in the relevant languages, it could also be applied to those languages with a non-Roman alphabet. One area that Digital Modern Linguists may struggle with is employing the POS-based clustering methodology in chapter 5 for languages that are more flexible with the order of their syntactic elements. For example, Levinsha et al. remark that, while English tends to be rigid with the its subject-object order, other languages are much more flexible, the maximal example being Lithuanian in which only roughly 60% of subject-object constructions adhere to the 'official' classification of the language (2021). Researchers in Lithuanian may thus wish to reconceptualise the way in which the tools created in chapter 5 approach syntactic elements. Future projects may also wish to take the various scripts and programs created throughout the project and transform them into more fully realised tools, that is to say a desktop program such as Wordsmith Tools 6 that was employed by Graedler during their investigation (2014). In doing so, we would open the techniques employed here up to an even wider range of scholars and researchers as there would be no requirement for even a minimum level of computer science knowledge.

The approach of this project lends itself well to application in other areas and objects of study within the sphere of digital cultures, beyond song – potentially encompassing film and television, online fora, social media, and other digital channels. With the exception of the like-count employed in chapter 5, the tools in this project are not YouTube-specific as the inputs for the various scripts are simple text strings in a Python data structure, allowing future researchers to pass data to the scripts that are not YouTube comments, for example Tweets or Tumblr posts. With this in mind, one potential future application of the methodology developed in this project might be in the study of the reception of re-imaginings of song in short-form video media such as TikTok. The adaptability of the methodology, both across languages and across different forms of digital media, and the future potential for this approach to serve as a model for other researchers justifies and supports the focus on documentation and the examination of the technical methodologies that we have employed. In tandem with this, we have taken steps to ensure that all the tools and datasets produced throughout this project are easily and openly available to the public in order that they can be reworked and reapplied, in line with the emphasis on accessibility within the culture of Digital Humanities studies. In sum, what we have created here is a body of work founded in interdisciplinary studies into song reception and modern languages that employs and creates various powerful methodologies of the Digital Humanities to the end of examining several facets of the phenomenon of popular foreign language music in the English-speaking world.

Bibliography

aboSamoor (2021) *PyCld2*. Available at: https://github.com/aboSamoor/pycld2 (Accessed: 6 June 2021).

Acevedo, N. (2019) *Reggaeton fuels Latin music boom despite lack of award recognition*. Available at: https://www.nbcnews.com/news/latino/reggaeton-fuels-latin-music-boom-despite-lack-award-recognition-n996161 (Accessed: 8 November 2021).

Achugar, M. and Pessoa, S. (2009) 'Power and place: Language attitudes towards Spanish in a bilingual academic community in Southwest Texas', *Spanish in Context*, 6(2), pp. 199–223.

Adeogun, A. O. (2018) 'A historical review of the evolution of music education in Nigeria until the end of the twentieth century', *Journal of the Musical Arts in Africa*, 15(1–2), pp. 1–18. doi: 10.2989/18121004.2018.1558954.

Aguila, J. (2014) *9. Becky G: 21 Under 21 (2014)*. Available at: https://www.billboard.com/music/music-news/becky-g-21-under-21-2014-6251181/ (Accessed: 4 May 2022).

Ailin, M. (2021) *Blackpink Discography*. Available at: https://kprofiles.com/blackpink-discography/ (Accessed: 31 July 2021).

Airoldi, M., Beraldo, D. and Gandini, A. (2016) 'Follow the algorithm: An exploratory investigation of music on YouTube', *Poetics*, 57, pp. 1–13. doi: https://doi.org/10.1016/j.poetic.2016.05.001.

Arbona-Ruiz, M. (2017) *The 'Despacito' effect: The year Latino music broke the charts*. Available at: https://www.nbcnews.com/news/latino/despacito-effect-year-latino-music-broke-charts-n830131 (Accessed: 7 September 2021).

Arnold, M. and Ohlebusch, E. (2011) 'Linear time algorithms for generalizations of the longest common substring problem', *Algorithmica*, 60(4), pp. 806–818.

Aslan, E. and Vásquez, C. (2018) '"Cash me ousside": A citizen sociolinguistic analysis of online metalinguistic commentary', *Journal of sociolinguistics*, 22(4), pp. 406–431. doi: 10.1111/josl.12303.

Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyżanowski, M., McEnery, T. and Wodak, R. (2008) 'A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press', *Discourse & society*, 19(3), pp. 273–306. doi: 10.1177/0957926508088962.

Barrus, T. (2021) *pyspellchecker*. Available at: https://pypi.org/project/pyspellchecker/ (Accessed: 23 April 2021).

Bastian, M., Heymann, S. and Jacomy, M. (2009) 'Gephi: an open source software for exploring and manipulating networks', in *Third international AAAI conference on weblogs and social media*.

Beal, V. (2021) *The Complete List Of 1500+ Common Text Abbreviations & Acronyms*. Available at: https://www.webopedia.com/reference/text-abbreviations/ (Accessed: 22 April 2020).

Big Hit (2021) *DISCOGRAPHY*. Available at: https://ibighit.com/bts/eng/discography/ (Accessed: 31 July 2021).

Billboard Staff (2007) *Billboard Hot 100 To Include Digital Streams*. Available at: https://www.billboard.com/music/music-news/billboard-hot-100-to-include-digital-streams-1050326/ (Accessed: 8 February 2022).

Bird, S., Klein, E. and Loper, E. (2009) *Natural language processing with Python: analyzing text with* *the natural language toolkit*. 'O'Reilly Media, Inc.'

Blanke, T. and Aradau, C. (2021) 'Computational genealogy: Continuities and discontinuities in the political rhetoric of US presidents', *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 54(1), pp. 29–43.

Boman, B. (2019) 'Achievement in the South Korean music industry', *International Journal of Music Business Research*, 8(2), pp. 6–26.

Born, G. and Hesmondhalgh, D. (2000) *Western music and its others : difference, representation, and appropriation in music / edited by Georgina Born and David Hesmondhalgh.* Berkeley : University of California Press, 2000.

Bou-Franch, P., Lorenzo-Dus, N. and Blitvich, P. G. (2012) 'Social Interaction in YouTube Text-Based Polylogues: A Study of Coherence', *Journal of computer-mediated communication*, 17(4), pp. 501–521. doi: 10.1111/j.1083-6101.2012.01579.x.

Brons, L. L. (2015) 'Othering, an analysis', *Transcience, a Journal of Global Studies*, 6(1).

Bruner, R. (2017) *The Top 10 Songs of 2017*. Available at: https://time.com/5034395/top-10-songs-2017/ (Accessed: 15 April 2021).

Burkholder, J. P. (James P. (2019) 'A history of Western music / J. Peter Burkholder, Donald Jay Grout, Claude V. Palisca.' Edited by D. J. Grout and C. V Palisca. New York : W.W. Norton and Company, 2019.

Cannam, C., Landone, C. and Sandler, M. (2010) 'Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files', in *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1467–1468.

Chankedar, P. (2019) *Speed up Cosine Similarity computations in Python using Numba*. Available at: https://medium.com/analytics-vidhya/speed-up-cosine-similarity-computations-in-python-using-numba-c04bc0741750 (Accessed: 30 December 2021).

Chen, J., Hu, Y., Liu, J., Xiao, Y. and Jiang, H. (2019) 'Deep short text classification with knowledge powered attention', in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 6252–6259.

Chun, E. W. (2013) 'Ironic blackness as masculine cool: Asian American language and authenticity on YouTube', *Applied linguistics*, 34(5), pp. 592–612.

Clark, E. and Araki, K. (2011) 'Text normalization in social media: progress, problems and applications for a pre-processing system of casual English', *Procedia-Social and Behavioral Sciences*, 27, pp. 2–11.

Coleman, E. G. (2010) 'Ethnographic Approaches to Digital Media', *Annual review of anthropology*, 39(1), pp. 487–505.

Crang, M. (2013) *Cultural geography*. Routledge.

Dabbura, I. (2018) *K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks*.

Damodaran, S. (2016) 'Protest and music', in *Oxford Research Encyclopedia of Politics*.

Davidov, D., Tsur, O. and Rappoport, A. (2010) 'Semi-supervised recognition of sarcasm in Twitter and Amazon', in *Proceedings of the fourteenth conference on computational natural language learning*, pp. 107–116.

Davidson, J. W. (1993) 'Visual perception of performance manner in the movements of solo musicians', *Psychology of music*, 21(2), pp. 103–113.

Dawkins, R. (1976) ‘11. Memes: the new replicators’, *The selfish gene*.

Denny, M. J. and Spirling, A. (2018) ‘Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It’, *Polit. Anal*, 26(2), pp. 168–189. doi: 10.1017/pan.2017.44.

Densmore, J. (2017) *Ethics in Web Scraping*. Available at: https://towardsdatascience.com/ethics-in-web-scraping-b96b18136f01 (Accessed: 15 April 2021).

Denziloe (2018) *Python: Cosine similarity between two large numpy arrays*. Available at: https://stackoverflow.com/questions/52030945/python-cosine-similarity-between-two-large-numpy-arrays (Accessed: 16 January 2022).

Dervin, F. (2012) ‘Cultural identity, representation and othering’, in *The Routledge handbook of language and intercultural communication*. Routledge, pp. 195–208.

Dobs, A. M. and Garcés-Conejos Blitvich, P. (2013) ‘Impoliteness in polylogal interaction: Accounting for face-threat witnesses’ responses’, *Journal of Pragmatics*, 53, pp. 112–130. doi: https://doi.org/10.1016/j.pragma.2013.05.002.

Dragojevic, M., Fasoli, F., Cramer, J. and Rakić, T. (2021) ‘Toward a Century of Language Attitudes Research: Looking Back and Moving Forward’, *Journal of language and social psychology*, 40(1), pp. 60–79. doi: 10.1177/0261927X20966714.

Drott, E. (2018) ‘Why the Next Song Matters: Streaming, Recommendation, Scarcity’, *Twentieth-Century Music*. 2018/11/29, 15(3), pp. 325–357. doi: DOI: 10.1017/S1478572218000245.

Elfving-Hwang, J. (2018) ‘K-pop idols, artificial beauty and affective fan relationships in South Korea’, in *Routledge handbook of celebrity studies*. Routledge, pp. 190–201.

Everett, Y. U. (2021) ‘From Exoticism to Interculturalism: Counterframing the East–West Binary’, *Music Theory Spectrum*, 43(2), pp. 330–338. doi: 10.1093/mts/mtab001.

Farrar, L. (2010) *‘Korean Wave’ of pop culture sweeps across Asia*. Available at: http://edition.cnn.com/2010/WORLD/asiapcf/12/31/korea.entertainment/index.html?iref=NS1 (Accessed: 20 August 2021).

Felderman, B. (2018) *SELECT ALL JUNE 29, 2018 ‘This Is So Sad Alexa Play Despacito,’ Explained(?)*. Available at: https://nymag.com/intelligencer/2018/06/what-is-the-this-is-so-sad-alexa-play-despacito-meme.html (Accessed: 26 November 2021).

Ferrara, E., JafariAsbagh, M., Varol, O., Qazvinian, V., Menczer, F. and Flammini, A. (2013) ‘Clustering memes in social media’, in *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*. IEEE, pp. 548–555.

Fitzsimmons-Doolan, S. (2014) ‘Using lexical variables to identify language ideologies in a policy corpus’, *Corpora*, 9(1), pp. 57–82.

Fowler, M. (1999) ‘Refactoring : improving the design of existing code / Martin Fowler’. Boston, Mass.: Addison-Wesley.

Garbade, M. J. (2018) *Understanding K-means Clustering in Machine Learning*. Available at: https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1 (Accessed: 15 February 2022).

Garbe, W. (2019) *symspellpy*. Available at: https://symspellpy.readthedocs.io/en/latest/index.html (Accessed: 23 April 2021).

Gawne, L., and McCulloch, G. (2019). Emoji as digital gestures. *Language@ Internet*, *17*(2).

Go FAIR (2021) *FAIR Principles*. Available at: https://www.go-fair.org/fair-principles/ (Accessed: 21 February 2021).

Goel, S. Anderson, A., Hofman, J. and Watts, D. J. (2016) 'The Structural Virality of Online Diffusion', *Management science*, 62(1), pp. 180–196. doi: 10.1287/mnsc.2015.2158.

Goldschmitt, K. E. and Seaver, N. (2019) 'Shaping the Stream: Techniques and Troubles of Algorithmic Recommendation', in *The Cambridge Companion to Music in Digital Culture*. Cambridge University Press, pp. 63–81.

González-Cruz, M. I. (2020) 'Othering and Language', *Love, Language, Place, and Identity in Popular Culture: Romancing the Other*, p. 453.

Gorelick, M. and Ozsvald, I. (2020) *High Performance Python: Practical Performant Programming for Humans*. O'Reilly Media.

Graedler, A.-L. (2014) 'Attitudes towards English in Norway: A corpus-based study of attitudinal expressions in newspaper discourse', *Multilingua-Journal of Cross-Cultural and Interlanguage Communication*, 33(3–4), pp. 291–312.

Griffiths, N. K. (2010) '"Posh music should equal posh dress": an investigation into the concert dress and physical appearance of female soloists', *Psychology of Music*, 38(2), pp. 159–177.

Hagberg, A., Schult, D. and Swart, P. (2008) 'Exploring network structure, dynamics, and function using NetworkX', in Varoquaux, G., Vaught, T., and Millman, J. (eds) *Proceedings of the 7th Python in Science Conference (SciPy2008)*. Pasadena, CA, pp. 11–15.

Hallman, J. (2021) *Efficient Featurization of Common N-grams via Dynamic Programming*. Available at: https://sisudata.com/blog/efficient-featurization-common-n-grams-via-dynamic-programming (Accessed: 8 September 2021).

Han, B. (2017) 'Korean Wave| K-pop in Latin America: transcultural fandom and digital mediation', *International Journal of Communication*, 11, p. 20.

Hirschberg, D. S. (1977) 'Algorithms for the longest common subsequence problem', *Journal of the ACM (JACM)*, 24(4), pp. 664–675.

hobogalaxy (2020) *youtube_multi_video_comment_downloader*. Available at: https://github.com/hobogalaxy/youtube_multi_video_comment_downloader/blob/master/yt-comment-scraper.py (Accessed: 20 November 2020).

Honn, J. (2014) *A Guide To Digital Humanities*. Available at: https://web.archive.org/web/20150919224700/http://sites.northwestern.edu/guidetodh/values-methods/ (Accessed: 5 October 2021).

Honnibal, M. (2013) *A Good Part-of-Speech Tagger in about 200 Lines of Python*. Available at: https://explosion.ai/blog/part-of-speech-pos-tagger-in-python (Accessed: 9 December 2021).

IFPI (2019) *Music listening in 2019*. Available at: https://www.ifpi.org/wp-content/uploads/2020/07/Music-Listening-2019-1.pdf.

Ilich, T. (2018) *Reggaeton Music Roots and Characteristics*. Available at: https://www.liveabout.com/reggaeton-puerto-rico-to-the-world-2141557 (Accessed: 7 October 2021)

Ingham, T. (2019) *ENGLISH-LANGUAGE MUSIC IS LOSING ITS STRANGLEHOLD ON GLOBAL POP*

CHARTS – AND YOUTUBE IS DRIVING THE CHANGE. Available at: https://www.musicbusinessworldwide.com/english-language-music-is-losing-its-stranglehold-on-global-pop-charts-and-youtube-proves-it/ (Accessed: 27 October 2020).

Iqbal, M. (2020) *Spotify Usage and Revenue Statistics (2020)*. Available at: https://www.businessofapps.com/data/spotify-statistics/ (Accessed: 12 February 2021).

Ivković, D. (2013) 'The Eurovision Song Contest on YouTube: A corpus-based analysis of language attitudes', *Language@Internet*, 10(1). Available at: http://nbn-resolving.de/urn:nbn:de:0009-7-35977.

Jafar, A. J. N. (2018) 'What is positionality and should it be expressed in quantitative studies?', *Emergency Medicine Journal*. pp. 323–324.

Jänicke, S., Franzini, G., Cheema, M. F. and Scheuermann, G. (2015) 'On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges.', in Borgo, R., Ganovelli, F. and Viola, I. (eds) *EuroVis (STARs)*, The Eurographics Association, pp. 83–103.

Jones, G. M. and Schieffelin, B. B. (2009) 'Talking Text and Talking Back: "My BFF Jill" from Boob Tube to YouTube', *Journal of computer-mediated communication*, 14(4), pp. 1050–1079. doi: 10.1111/j.1083-6101.2009.01481.x.

Joshi, T. (2020) *Will the UK ever love foreign-language pop?* Available at: https://www.theguardian.com/music/2020/aug/05/will-the-uk-ever-love-foreign-language-pop (Accessed: 7 April 2021).

Kerievsky, J. (2005) *Refactoring to patterns*. Pearson Deutschland GmbH.

Kharde, V. and Sonawane, P. (2016) 'Sentiment analysis of twitter data: a survey of techniques', *International Journal of Computer Applications,* 139(11) pp. 5-15. doi: 10.5120/ijca2016908625

Kim, J. O. (2021) 'BTS as method: a counter-hegemonic culture in the network society', *Media, Culture & Society*, 43(6), pp 1061–1077. doi: 10.1177/0163443720986029.

Kim, S., Hwang, S. and Kim, J. (2021) 'Factors influencing K-pop artists' success on V live online video platform', *Telecommunications Policy*, 45(3), p. 102090. doi: https://doi.org/10.1016/j.telpol.2020.102090.

Klebanov, B. B., Diermeier, D. and Beigman, E. (2008) 'Automatic annotation of semantic fields for political science research', *Journal of Information Technology & Politics*, 5(1), pp. 95–120.

Know Your Meme (2021) *Trivago Guy*. Available at: https://knowyourmeme.com/memes/trivago-guy (Accessed: 24 January 2022).

Latestly (2020) *K-Pop Bands, Fandom Names and Their Meanings: From BTS' ARMY to BLACKPINK's BLINK, Here's A Complete List of South Korean Music Groups' Fan Club Names*. Available at: https://www.latestly.com/social-viral/k-pop-bands-fandom-names-and-their-meanings-from-bts-army-to-blackpinks-blink-heres-a-complete-list-of-south-korean-music-groups-fan-club-names-2225132.html (Accessed: 27 September 2021).

Leeman, J. and Martínez, G. (2007) 'From identity to commodity: Ideologies of Spanish in heritage language textbooks', *Critical Inquiry in Language Studies*, 4(1), pp. 35–65.

Levshina, N., Namboodiripad., Allassonnière-Tang, M., Kramer, M. A., Talamo, L., Verkerk, A., Wilmoth, S., Rodriguez, G. G., Gupton, T. and Kidd, E. (2021) 'Why we need a gradient approach to word order'. To be published in *PsyArXiv* [preprint]. Available at: https://psyarxiv.com/yg9bf/

Lie, J. (2014) 'Why Didn't" Gangnam Style" Go Viral in Japan?: Gender Divide and Subcultural

Heterogeneity in Contemporary Japan', *Cross-Currents: East Asian History and Culture Review*, 3(1), pp. 6–31.

Linguistic Data Consortium (2021) *BLLIP 1987-89 WSJ Corpus Release 1*. doi: https://doi.org/10.35111/fwew-da58.

Lippi-Green, R. (2012) *English with an accent: Language, ideology, and discrimination in the United States*. Routledge.

Lipton, J. (2014) *Google's best and worst acquisitions*. Available at: https://www.cnbc.com/2014/08/19/googles-best-and-worst-acquisitions.html (Accessed: 27 November 2020).

Liu, Y. and Gao, X. (2020) 'Commodification of the Chinese language: investigating language ideology in the Irish media', *Current Issues in Language Planning*, 21(5), pp. 512–531.

Loria, S. (2018) 'textblob Documentation', *Release 0.15*, 2, p. 269.

Mamaev, M. (2018) *If you have slow loops in Python, you can fix it…until you can't*. Available at: https://www.freecodecamp.org/news/if-you-have-slow-loops-in-python-you-can-fix-it-until-you-cant-3a39e03b6f35/ (Accessed: 16 January 2022).

Mangold, W. G. and Faulds, D. J. (2009) 'Social media: The new hybrid element of the promotion mix', *Business horizons*, 52(4), pp. 357–365.

Marc, I. (2015) 'Travelling Songs: on Popular Music Transfer and Translation', *IASPM@Journal*, 5(2), pp. 3–21.

Marino, G. (2015) 'Semiotics of spreadability: A systematic approach to Internet memes and virality', *Punctum*, 1(1). pp. 43-66.

Marshall, W. (2006) *The rise of reggaeton*. Available at: https://thephoenix.com/Boston/Music/1595-rise-of-reggaeton/ (Accessed: 6 September 2021).

Martin, R. C. (2009) *Clean code: a handbook of agile software craftsmanship*. Pearson Education.

Martin, S., Brown, W. M., Klavans, R. and Boyack, K. W(2011) 'OpenOrd: an open-source toolbox for large graph layout', in *Visualization and Data Analysis 2011*. International Society for Optics and Photonics, p. 786806.

Marwick, A. (2013) 'Memes', *Contexts*, 12(4), pp. 12–13.

McCormack, A. (2020) *'A dumbing down of music': Have streaming services changed music releases forever?* Available at: https://www.abc.net.au/triplej/programs/hack/a-dumbing-down-of-music-have-streaming-services-changed-music-r/12916224 (Accessed: 12 February 2021).

McCourt, T. and Zuberi, N. (2016) 'Music and discovery', *Popular Communication*, 14(3), pp. 123–126. doi: 10.1080/15405702.2016.1199025.

Meindertsma, P. (2019) 'Changes in Lyrical and Hit Diversity of Popular U.S. Songs 1956-2016', *Digital Humanities Quarterly*, 13(4).

Meindertsma, P. (2020) *Lyric time series*. Available at: https://www.petermeindertsma.com/lyrics/ (Accessed: 18 February 2021).

Morris, J. W. and Powers, D. (2015) 'Control, curation and musical experience in streaming music services', *Creative industries journal*, 8(2), pp. 106–122.

Mueller, A. (2012) *A Wordcloud in Python*. Available at: https://peekaboovision.blogspot.com/2012/11/a-wordcloud-in-python.html (Accessed: 7 October 2021).

Nieva, R. (2016) *YouTube started as an online dating site*. Available at: https://www.cnet.com/news/youtube-started-as-an-online-dating-site/ (Accessed: 27 November 2020).

Numba (2018) *Numba*. Available at: https://numba.pydata.org/ (Accessed: 1 January 2022).

O'Connor, R. (2017) *Súbeme la radio: How audiences are adapting to non-English language music*. Available at: https://www.independent.co.uk/arts-entertainment/music/features/reggaeton-popular-international-hits-despacito-daddy-yankee-j-balvin-k-pop-bts-stotify-streams-language-sony-atlantic-a8110541.html (Accessed: 13 April 2021).

Official Charts (2021) *How the Official Charts are compiled*. Available at: https://www.officialcharts.com/getting-into-the-charts/how-the-charts-are-compiled/ (Accessed: 15 November 2021).

Otmazgin, N. and Lyan, I. (2014) 'Hallyu across the desert: K-pop fandom in Israel and Palestine', *Cross-Currents: East Asian History and Culture Review*, 3(1), pp. 32–55.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., (2011) 'Scikit-learn: Machine learning in Python', *the Journal of machine Learning research*, 12, pp. 2825–2830.

Radford, A. (2004) *English syntax : an introduction / Andrew Radford. [electronic resource]*. Cambridge : Cambridge University Press, 2004.

Raiford, T. (2021) *10 Things You Didn't Know about Nicky Jam*. Available at: https://tvovermind.com/nicky-jam/ (Accessed: 4 May 2022).

RapidAPI (2021) *Top 8 Best Websites to Find Song Lyrics (in 2021)*. Available at: https://rapidapi.com/blog/best-websites-song-lyrics/ (Accessed: 20 November 2020).

Rivera-Rideau, P. R. (2015) *Remixing reggaetón : the cultural politics of race in Puerto Rico / Petra R. Rivera-Rideau.* Durham : Duke University Press, 2015.

Roark, B. (2007) *Computational approaches to morphology and syntax / Brian Roark and Richard Sproat.* Edited by R. W. Sproat. Oxford .

Roettgers, J. (2011) *Most YouTube views come from non-English users*. Available at: https://gigaom.com/2011/11/03/youtube-global-language-stats/ (Accessed: 1 October 2020).

Romano, F. (2018) *Learn python programming : a beginner's guide to learning the fundamentals of python language to write efficient, high-quality code / Fabrizio Romano.* Second edi. Birmingham: Packt Publishing Ltd.

Rossetti, G., Milli, L. and Cazabet, R. (2019) 'CDLIB: a python library to extract, compare and evaluate communities from complex networks', *Applied Network Science*, 4(1), pp. 1–26.

Rousse-Marquet, J. (2012) *La K-pop, phénomène musical coréen*. Available at: https://larevuedesmedias.ina.fr/la-k-pop-phenomene-musical-coreen (Accessed: 20 August 2021).

Roxbergh, L. (2019) *Language Classification of Music Using Metadata*. M.Sc. Thesis, Upsalla Universitet. Available at: http://www.diva-portal.org/smash/get/diva2:1297032/FULLTEXT01.pdf (Accessed: 15 January 2021)

Russell, M. J. (2012) *The Gangnam Phenom*. Available at: https://foreignpolicy.com/2012/09/27/the-gangnam-phenom/ (Accessed: 20 August 2021).

Ryan, C., Wapnick, J., Lacaille, N., and Darrow, A. A. (2006). The effects of various physical characteristics of high-level performers on adjudicators' performance ratings. *Psychology of Music*, *34*(4), 559-572. doi: 10.1177/0305735606068106.

Satapathy, R., Guerreiro, C., Chaturvedi, I. and Cambria, E., (2017) 'Phonetic-based microtext normalization for twitter sentiment analysis', in *2017 IEEE international conference on data mining workshops (ICDMW)*. IEEE, pp. 407–413.

Schneebeli, C., 2017, November. The interplay of emoji, emoticons, and verbal modalities in CMC: a case study of YouTube comments. In *VINM 2017: Visualizing (in) the new media*. Available at: https://halshs.archives-ouvertes.fr/halshs-01632753/document (Accessed: 2 February 2021)

Solomon, B. (2018) *Look Ma, No For-Loops: Array Programming With NumPy*. Available at: https://realpython.com/numpy-array-programming/ (Accessed: 16 January 2022).

Soundarajan, S., Eliassi-Rad, T. and Gallagher, B. (2014) 'A guide to selecting a network similarity method', *SIAM international conference on data mining*. 14, Philadelphia: Sheraton Society Hill Hotel. Available at: https://epubs.siam.org/doi/abs/10.1137/1.9781611973440.118 (Accessed: 20 January 2022).

Srinivasa-Desikan, B. (2018) *Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*. Packt Publishing Ltd.

Stein, L., Jenkins, H., Ford, S., Green, J., Booth, P., Busse, K., Click, M., Li, X. and Ross, S., (2014) 'Spreadable media: creating value and meaning in a networked culture', *cinema Journal*, 53(3), pp. 152–177.

Stewart, J. (2012) 'What's Going On: Anti-War and Pro–War Hits on the Billboard Singles Charts during the Vietnam War Era (1965–1975) and the "War on Terror"(2001–2010)', in Gibson, S. and Mollan, S. (eds) *Representations of Peace and Conflict*. Basingstoke: Palgrave Macmillan, pp. 67–85.

Susam-Saraeva, Ş. (2019) 'Interlingual cover versions: how popular songs travel round the world', *The Translator*, 25(1), pp. 42–59. doi: 10.1080/13556509.2018.1549710.

Swan, A. L. (2018) 'Transnational identities and feeling in fandom: Place and embodiment in K-pop fan reaction videos', *Communication Culture & Critique*, 11(4), pp. 548–565.

Terraschke, A. (2007) 'Use of general extenders by German non-native speakers of English', *IRAL - International Review of Applied Linguistics in Language Teaching,* 45(2), pp 141-160

Traag, V. A., Waltman, L. and Van Eck, N. J. (2019) 'From Louvain to Leiden: guaranteeing well-connected communities', *Scientific reports*, 9(1), pp. 1–12.

Upton, T. A. and Ann Cohen, M. (2009) 'An approach to corpus-based discourse analysis: The move analysis as example', *Discourse studies*, 11(5), pp. 585–605.

Verma, S. (2019) *How Fast Numpy Really is and Why?* Available at: https://towardsdatascience.com/how-fast-numpy-really-is-e9111df44347 (Accessed: 31 December 2021).

Veselovsky, V., Waller, I. and Anderson, A. (2021) 'Imagine All the People: Characterizing Social Music Sharing on Reddit', in *Proceedings of the International AAAI Conference on Web and Social Media*, pp. 739–750.

Walters, K. and Chun, E. (2011) 'Orienting to Arab Orientalisms: Language, Race, and Humor in a YouTube Video', in *Digital Discourse*. Oxford University Press.

Wasser, L. and Joseph, M. (2017) *Lesson 4. Introduction to the JSON data structure*. Available at:

https://www.earthdatascience.org/courses/earth-analytics/get-data-using-apis/intro-to-JSON/ (Accessed: 18 February 2021)

Wayback Machine (2020) *YouTube.com*. Available at: https://web.archive.org/web/20120704144721/http://www.youtube.com/watch?v=JZwxgUODh8g&feature=g-logo-xit (Accessed: 30 November 2020).

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E. and Bouwman, J. (2016) 'The FAIR Guiding Principles for scientific data management and stewardship', *Scientific Data*, 3(1), p. 160018. doi: 10.1038/sdata.2016.18.

Williams, J. (2019) *What Does 'I Purple You' Mean? BTS Fans Celebrate 1,000 Days of Kim Taehyung's Phrase of Love*. Available at: https://www.newsweek.com/bts-kim-taehyung-purple-meaning-1453501 (Accessed: 20 August 2021).

Xu, W.W., Park, J.Y., Kim, J.Y. and Park, H.W. (2016) 'Networked cultural diffusion and creation on YouTube: An analysis of YouTube memes', *Journal of Broadcasting & Electronic Media*, 60(1), pp. 104–122.

Yoon, K., Min, W. and Jin, D. Y. (2020) 'Consuming the Contra-Flow of K-pop in Spain', *Journal of Intercultural Studies*, 41(2), pp. 132–147.

Zabus, C. (1990) 'Othering the Foreign Language in the West African Europhone Novel', *Canadian Review of Comparative Literature/Revue canadienne de littérature comparée*, pp. 348–366.

Appendices

Appendix 1: Full list of reggaeton songs examined in chapter 4

Song	Main Artist	Year released	Featuring artists
BEBE	6ix9ine	2018	Anuel AA
YAYA	6ix9ine	2020	
On My Way	Alan Walker	2019	Sabrina Carpenter, Farruko
Hasta Que Dios Diga	Anuel AA	2020	Bad Bunny
Keii	Anuel AA	2020	
Secreto	Anuel AA	2019	Karol G
Sigues Con Él	Arcangel	2019	Sech
Bichiyal	Bad Bunny	2020	Yaviah
Callaita	Bad Bunny	2019	
Dakiti	Bad Bunny	2020	Jhay Cortez
Está Cabrón Ser Yo	Bad Bunny	2020	Anuel AA
Ignorantes	Bad Bunny	2020	Sech
La Difícil	Bad Bunny	2020	
La Santa	Bad Bunny	2020	Daddy Yankee
Mia	Bad Bunny	2018	Drake
Pero Ya No	Bad Bunny	2020	
Safaera	Bad Bunny	2020	Jowell and Randy, Ñengo Flow
Si Veo a Tu Mamá	Bad Bunny	2020	
Solo De Mi	Bad Bunny	2019	
VETE	Bad Bunny	2019	
Yo Perreo Sola	Bad Bunny	2020	
Mayores	Becky G	2017	Bad Bunny
Sin Pijama	Becky G	2017	Natti Natasha
I Can't Get Enough	Benny Blanco	2019	Tainy, Selena Gomez, J. Balvin
MAMACITA	Black Eyed Peas	2019	Ozuna, J. Rey Soul
Ritmo	Black Eyed Peas	2019	J. Balvin
Reggaetón Lento (Bailemos)	CNCO	2017	Little Mix
Reggaetón Lento	CNCO	2017	
I Like It	Cardi B	2018	Bad Bunny, J. Balvin
La Bicicleta	Carlos Vives	2016	Shakira
Te Bote Remix	Casper	2016	Nio García, Darell, Nicky Jam, Bad Bunny, Ozuna
Armada Latina	Cypress Hill	2010	Pitbull, Marc Anthony

You Stay	DJ Khaled	2019	Meek Mill, J. Balvin, Lil Baby, Jeremih
Loco Contigo	DJ Snake	2019	J.Balvin, Tyga
Taki Taki	DJ Snake	2018	Selena Gomez, Ozuna, Cardi B
Con Calma	Daddy Yankee	2019	Snow
Dura	Daddy Yankee	2018	
Shaky Shaky	Daddy Yankee	2017	
Say My Name	David Guetta	2018	Bebe Rexha, J. Balvin
Dame Tu Cosita	El Chombo	2018	Cutty Ranks
Krippy Kush	Farruko	2017	Bad Bunny, Rvssian
La Tóxica	Farruko	2017	
Te Quiero	Flex	2008	
Ginza	J. Balvin	2015	
Mi Gente	J. Balvin	2017	Willy William
Un Dia	J. Balvin	2017	Dua Lipa, Bad Bunny, Tainy
Dinero	Jennifer Lopez	2018	DJ Khaled, Cardi B
No Me Conoce	Jhay Cortez	2019	J. Balvin, Bad Bunny
Tusa	Karol G	2019	Nicki Minaj
Whine Up	Kat DeLuna	2007	Elephant Man
Almost Like Praying	Lin-Manuel Miranda	2017	Artists For Puerto Rico
Despacito (Remix)	Luis Fonsi	2017	Daddy Yankee, Justin Bieber
Despacito	Luis Fonsi	2017	Daddy Yankee
Échame La Culpa	Luis Fonsi	2007	Demi Lovato
Soltera	Lunay	2019	Chris Jeday, Gaby Music
Corazón	Maluma	2018	Nego Do Borel
Felices Los 4	Maluma	2017	
HP	Maluma	2019	
Hawái	Maluma	2020	
Uptown Vibes	Meek Mill	2018	Fabolous, Anuel AA
Criminal	Natti Natasha	2011	Ozuna
El Amante	Nicky Jam	2017	
El Perdón	Nicky Jam	2015	Enrique Iglesias
Te Robaré	Nicky Jam	2015	Ozuna
X	Nicky Jam	2018	J. Balvin
La Jeepeta	Nio Garcia	2018	Brray, Juanka, Anuel AA, Myke Towers

Baila Baila Baila	Ozuna	2019	
Caramelo	Ozuna	2020	
El Farsante	Ozuna	2018	Romeo Santos
La Modelo	Ozuna	2018	CardiB
Vaina Loca	Ozuna	2018	Manuel Turizo
Bon Bon	Pitbull	2018	
I Know You Want Me (Calle Ocho)	Pitbull	2009	
The Anthem	Pitbull	2008	
Enjoy Yourself	Pop Smoke	2020	KarolG
TKN	ROSALÍA	2020	Travis Scott
Me Niego	Reik	2018	Ozuna, Wisin
Bella Y Sensual	Romeo Santos	2017	Daddy Yankee, Nicky Jam
Otro Trago	Sech	2019	Darell
Relación	Sech	2020	
Chantaje	Shakira	2016	Maluma
La La La	Shakira	2014	Carlinhos Brown
Me Enamoré	Shakira	2017	
Perro Fiel	Shakira	2017	Nicky Jam
Adicto	Tainy	2019	Anuel AA, Ozuna
Adrenalina	Wisin	2014	Jennifer Lopez, Ricky Martin
Arms Around You	XXXTentacion	2018	Lil Pump, Maluma, Swae Lee
I don’t even speak spanish lol	XXXTentacion	2018	
Go Loko	YG	2019	Tyga, Jon Z

Appendix 2: List of reggaeton songs in each cluster

Green group

Green parent-cluster members

Title	Artist
BEBE	6ix9ine
YAYA	6ix9ine

Green child-cluster members

Title	Artist
Keii	Anuel AA
La Difícil	Bad Bunny
Safaera	Bad Bunny
Si Veoa Tu Mamá	Bad Bunny
I Can't Get Enough	Benny Blanco
Dame Tu Cosita	El Chombo
Krippy Kush	Farruko
Te Quiero	Flex
Whine Up	Kat DeLuna
Uptown Vibes	Meek Mill
La Jeepeta	Nio Garcia
TKN	ROSALíA
I don’t even speak Spanish lol	XXXTentacion
Go Loko	YG

Blue Group

Blue parent-cluster members

Title	Artist
Secreto	Anuel AA
Te Bote Remix	Casper
Mayores	Becky G
Sin Pijama	Becky G
Con Calma	Daddy Yankee
Despacito	Luis Fonsi
Échame La Culpa	Luis Fonsi
Felices Los 4	Maluma
El Amante	Nicky Jam
El Perdón	Nicky Jam
X	Nicky Jam
Chantaje	Shakira
Me Enamoré	Shakira
Reggaetón Lento	CNCO
Despacito (Remix)	Luis Fonsi
Ginza	J. Balvin
Mi Gente	J. Balvin
Criminal	Natti Natasha
El Farsante	Ozuna
Loco Contigo	DJ Snake
Dura	Daddy Yankee
Corazón	Maluma
Hawái	Maluma
Perro Fiel	Shakira
Adrenalina	Wisin

Blue child-cluster members

Title	Artist
On My Way	Alan Walker
Hasta Que Dios Diga	Anuel AA
Sigues Con Él	Arcangel
Callaita	Bad Bunny
Dakiti	Bad Bunny
La Santa	Bad Bunny
Pero Ya No	Bad Bunny
Solo De Mi	Bad Bunny
VETE	Bad Bunny
MAMACITA	Black Eyed Peas
Ritmo	Black Eyed Peas
Reggaetón Lento (Bailemos)	CNCO
La Bicicleta	Carlos Vives
Armada Latina	Cypress Hill
Taki Taki	DJ Snake
Shaky Shaky	Daddy Yankee
Say My Name	David Guetta
La Tóxica	Farruko
No Me Conoce	Jhay Cortez
Almost Like Praying	Lin-Manuel Miranda
Soltera	Lunay
HP	Maluma
Te Robaré	Nicky Jam
Baila Baila Baila	Ozuna
Caramelo	Ozuna
Vaina Loca	Ozuna
I Know You Want Me (Calle Ocho)	Pitbull
Me Niego	Reik
Bella Y Sensual	Romeo Santos
Otro Trago	Sech
Relación	Sech
Adicto	Tainy

Red Group

Red parent-cluster members

Title	Artist
Un Dia	J. Balvin
I Like It	Cardi B
Tusa	Karol G
Mia	Bad Bunny
La Modelo	Ozuna

Red child-cluster members

Title	Artist
Bichiyal	Bad Bunny
Está Cabrón Ser Yo	Bad Bunny
Ignorantes	Bad Bunny
Yo Perreo Sola	Bad Bunny
You Stay	DJ Khaled
Dinero	Jennifer Lopez
Bon Bon	Pitbull
The Anthem	Pitbull
Enjoy Yourself	Pop Smoke
La La La	Shakira
Arms Around You	XXXTentacion