Universität Leipzig

Fakultät für Mathematik und Informatik

Institut für Informatik

Einführung in die Digital Humanities

WS 2017/18

Dr. Francesco Mambrini

The Lyrical Characteristics of Pop Music: Studying Differences in Song Lyrics between Genres

Authors:

Jachnow, Kristofer	Schuler, Paul
BA Sociology, 7th Semester	BA Sociology, 5th Semester
3706201	3748049
kj49zike@studserv.uni-leipzig.de	ps66kuqa@studserv.uni-leipzig.de

Submitted: 19th March 2018

Number of character (incl. blank character): 29 562

List of contents

List of figures

Abstract

The basic question of this report is what differences exist between Pop music lyrics and that of other genres. Especially the characteristics of Pop lyrics are examined for the purpose of answering the question whether the simplicity and generality of this type of music causes its prevalence. With the aid of scraping lyrics of music charts it is found that a lower lexical density has a slight positive effect on the popularity of songs. Pop lyrics show a relatively low text density but extensive vocabulary, while no special thesaurus is used. The results provide points for development and a basis for further research.

1 Introduction and Motivation

The analysis of lyrics is a relatively young but rapidly growing target for research. For instance, Chi et al. (2009) studied the significance of lyrics in predicting song emotion. The broadening of digital audio formats and portable music players is nowadays accompanied by an increasing demand for music recommendation systems. Since these developments allow people to have access to a seemingly unlimited amount of music, it is impossible to organise that quantity manually. For this reason, mainstream music mood estimation technologies try to predict song emotion by audio tracks while lyrics are mostly ignored. On the contrary, Chi et al. showed on a basis of a 600 Pop song database that lyrics are not only a valid measure for estimating a song's mood rating but also provide supplementary information that can help improve audio-only recommendation systems. Furthermore, predictions based on lyrics text are better estimators of the overall song mood for songs with conflicting predictions. Thus, Chi et al. claim based on their findings, that lyrics can be treated as the key design factor in future music mood estimation systems. Another study was conducted by Gao et al. (2016). The authors of this study examined whether the complexity of song lyrics influence the likelihood to be successful. The stated goal of this investigation is to provide record labels and similiar organisations with a model to predict whether a song will become famous or not, before it is published - therefore, if an investment is appropriate or not.

Besides that, another interesting observation is that over the years the variety of music genres increased tremendously, with new styles such as "Comedy Rock", "Outlaw Country" and "Eurodance" occuring. This development even lead to a point in which it is suggested to discard the entire concept of music genres[1]. Nevertheless, some genres seem to be timeless, since songs have been labeled by them for many years. Probably the most prominent representative for this is the so-called Pop genre. Although genres in genral are exceedingly difficult to define, this is particularly true of Pop.

Looking at both depicted facts, it would be worthwhile taking a closer look at the lyrics of Pop songs. For this reason, the focus of this report is to examine the differences of Pop lyrics to other genres and especially to find out, which characteristics constitute lyrics of Pop songs. Furthermore, the connection between the popularity and the lyrics of a song shall be examined. These issues shall be analysed with the aid of text evaluation and computational analysis.

In the first section the theoretical approach of this work and in particular the problem of defining Pop music will be targeted. The selection and acquisition of the data used will be described in the second part. The third section will address the analysis of the data. Finally, the results of this work will be summarised and discussed.

[1] http://www.econotimes.com/Musical-genres-are-out-of-date-%E2%80%93-but-this-new-system-explains-why-you-might-like-both-jazz-and-hip-hop-244941

2 Preliminary Considerations

2.1 Research Approach

Pop music, short for "popular music" has from the 1950's onwards dominated the music market and radio playlists. Many songs, originally forbidden for propriety reasons were most successful only a few years later. The question is, why some artists and their music are so popular to huge crowds of people across different continents and cultures. The unique features and progressiveness might be one important reason. Another fundamental factor could be the intersection of musical preferences of generations across nations. This requires simplicity and slenderness of the music plus lyrics that are not too unfamiliar to the listeners. Thus, Pop lyrics are expected to rather consist of every day language which is used among all cultures. This phenomenon is noticeable in the broadcasting habits of radio stations, which exclude songs of certain lengths and with explicit words.

Obviously a song attributed as a Pop song must have specific characteristics. Pop music is not equivalent to a list of the most popular or most successful songs. The importance of language and lyrics for music are in the main focus of this piece of work. It is assumed that the need for simple language and little lexical complexity will be observed best in Pop music for the mentioned reasons, whereas other further distanced genres do not require such restrictions and therefore should be more diverse in the use of words. It is further expected that the field of word usage will show a much more limited dimension for Pop music than for other types, even though every type of music usually has its own vocabulary. To clarify this statement, it is expected that there is a rather general word pool for Pop music and a more type associated pool for the others.

From these deliberations two central hypotheses are derived, which shall be examined in the following:

1. The more successful a song in Pop music, the less complex the vocabulary; and
2. compared to other genres Pop music is based on a much more general word field

2.2 Operationalisation

One difficulty, which definitely has to be targeted, is the distinction between certain genres. Since this is the central argument of this report, it has to be decided what characteristics define the parts of the spectrum such as "Pop", "Rock" "R&B/Hip-Hop", et cetera. As the central independent variable, the operationalisation must be clear and determined before the first calculations and data analysis are undertaken, that there is no risk to alter the initially chosen methods due to the characteristics of certain songs.

By the Oxford Music definition, genre means a shared tradition or a set of conventions for pieces of music[2] The fact that Pop music comes from "popular music" makes a simple definition more difficult, since all popular songs from different genres could be included in the list, even

[2] http://www.oxfordmusiconline.com/grovemusic/view/10.1093/gmo/9781561592630.001.0001/omo-9781561592630-e-0000040599?rskey=MqTH9Wß&result=1 (last accessed 13th March 2018)

though the "classic definition" might have rather been "Country Music", "Punk" or "Grunge". Therefore it is clear that Pop music cannot be a classification of liked songs measured by sales or times played on the radio but must be on the same level as other genres such as rock music. Another matter to be mentioned is that certain songs or music can be ascribed to more than one genre.

As one can see, the question for the most clear definition of classes of music would take up enough space and effort to write a musicological monography. To faciliate the decision for the independent variable the help of others is required, without necessarily diminishing the quality of the analysis, namely of the "Music Chart" concept. The benefit of using it is the fundamental coherence in the classifications, plus rankings which describe the success of records. Underlying the charts are algorithms, combining airplay, streaming and sale figures to different ratios. To be clear, a higher success of a song means here a lower chart ranking.

One emerging problem of the mere observation of these factors[3] is the restrictiveness. This could constitute a distortion of the true most popular or respectively most typical songs of each genre. Since the ranks are calculated on the basis of limited measurements, it is possible that certain songs fail to be noticed and are not represented on the chart as they are in the music scene. One example are Madonna's singles released in 2003. Boycotted by republican radio stations, the ranks in the "Hot 100" list, which includes airplay figures, were significantly lower (in terms of 'less successful') than the "Billboard Sales Charts". As the focus is on genres in general rather than the most popular and successful representatives, this fact seems to be negligible for the research question. Naturally, a study on the success of certain songs and their features might be of interest too, but shall not be undertaken here.

For an easier comparison the data is merged to a corpus. To examine the vocabulary and application of words, quantitative statistical methods are used, which will be further described in section "Data Analysis". In order to verify the second hypothesis, the vocabulary of the music types is compared through the examination of the most common words.

[3] example Billboard charts: ratio for calculations changes every week, with the average of sales (35-45%), airplay (30-40%) and streaming (20-30%) (https://www.billboard.com/articles/columns/ask-billboard/5740625/ask-billboard-how-does-the-hot-100-work (last accessed 13th March 2018))

3 Data

3.1 Acquisition

To acquire the data needed to answer questions on the differences between genres, a lyric dataset is created, containing the printed lyrics of songs representing the concerned genres. The easiest and most simple way to find representative songs for each type of music appears to be the use of music charts, as mentioned above. For this *Billboard Charts* are used, an old and renowned institution to list the most successful songs and albums.

Billboard uses a system adding a "fingerprint" to every song and collecting data from shops, radio stations and streaming services[4]. Using R, data is scraped from the online Billboard archive (www.billboard.com, 27th february 2018) for two dates, 1st January 2010 and 1st January 2015. Six different genres were chosen to examine differences. These are "Alternative", "Country", "Gospel", "R&B/Hip-Hop", "Pop", "Rock". The dates were chosen randomly since the research question is not specifically about certain points in time, but the differences between types of music. However, the choice was limited because of the archive dating back to the 70's for some genres and only to 2009 for others.

With the information dowloaded from the internet, a data frame was created, containing the artist's name, the song name, the date of the chart and the rank at that time. Using the song's and artist's name it was possible to fetch the corresponding lyrics from originally three different websites, namely "www.metrolyrics.com", "www.songlyrics.com" and "www.lyricsmode.com", which, after complications with "metrolyrics", were eventually reduced to the two latter ones, with a 93.4% success rate, which means that 31 lyrics couldn't be successfully downloaded or were incomplete.

Finally, a dataset containing 488 valid Songs (510-31 missing) of six different genres was compiled. One has to mention the differences in numbers between the genres reaching from 100 lyrics (Rock) to 50 lyrics (Gospel). Keeping that in mind and including the different sample sizes in the analysis, the validity of the results can not be compromised.

3.2 Conditioning

The first step of the analysis is cleaning and conditioning the dataset. Given that the analysis is for reasons of comparability limited to lyrics in English, it is necessary to fix the most common contractions by expanding them to their long form, for example, "won't" to "will not", "can't" to "can not". After that, every character that does not carry any meaning, like "," or "*", is removed. Finally, to have a consistent dataset and to not make incorrect differentiations between some values, every element of the dataset is converted to lowercase. Since song lyrics are a special kind of text, they often contain phrases like "Bridge", "Chorus" or "Repeat". In terms of the pursued analysis these phrases do not contain relevant information, which is why

[4] for further information: https://www.billboard.com/articles/business/8006673/billboard-charts-adjust-streaming-weighting-2018; https://de.wikipedia.org/wiki/Billboard_(Magazin) (last accessed on 19th March 2018)

they are also removed. Same is done to known error messages of the websites used. At this point, the dataset and particularly the lyrics are free from obvious noise.

The examination of the lyrics with the lowest number of words reveals that some songs have incorrect lyrics. As far as a verification was possible, songs with erroneous lyrics were removed from the dataset. The conditioned dataset contains 477 observations, which includes 76 Alternative songs, 99 Country songs, 39 Gospel songs, 90 R&B/Hip-Hop songs, 77 Pop songs and 96 Rock songs.

4 Data Analysis

The fundamental question at the beginning of this investigation is, given the present dataset, which quantitative measurements are possible and meaningful to examine the issues raised. Following Debbie Liske in her tutorial "Lyric Analysis with NLP & Machine Learning with R"[5], the following three measurements will be used: First of all word frequency, which means the number of words in a song or even in the entire genre. Second the lexical diversity, which represents the used vocabulary and is defined as the number of unique words in a text or genre. Finally, the third measurement will be lexical density, which is defined as the number of unique words divided by the total number of words. This measurement aims to take the length of the song and repetitions into account, as far as the possibilities exist. A detailed explanation of the measurements will be given further below.

To begin the exploratory analysis, Figure 1 shows the different distributions of the word counts for every genre in the data. What can be seen here, is the typical range of the number of words used in a song specific for every genre. Since the typical ranges differ a lot between the genres, it is crucial to analyse the number of words always in regard to the genre. The distribution of word counts in R&B/Hip-Hop songs is much higher located than the other genres. In this dataset R&B/Hip-Hop songs show the highest number of words, while Gospel and Rock songs have the lowest distributions of word counts. An important observation is that Pop songs have overall lower word counts than R&B/Hip-Hop songs, but are nevertheless higher than the remaining genres.

In order to be able to assess how much the different genres use repetitions as a stylistic element in the lyrics, it is necessary to know the vocabulary used in a genre. Simply counting how much different words are used in a genre shows the following results. Again, R&B/Hip-Hop is at the top of the genres with 4272 different words used. On the other side, Gospel has the lowest vocabulary with 1180 different words used. The other genres are relatively close to each other, Alternative has 2117 different words, Rock 2289, Country 2371 and finally Pop 2465. Due to the fact that the dataset contains a different number of songs for every genre, it is important to check, if this is the cause for the variety in the number of different words. The vocabulary of a genre divided by the number of songs is therefore a measurement which shows the vocabulary in relation to the number of words. Once again, R&B/Hip-Hop is first with an average of 47.5 unique words used per song and Pop is second with 32. Surprisingly Gospel is this time in front of the remaining genres with 30.3 unique words per song. This could be an indicator, that the litte amount of Gospel songs in the data is causing the low number of different words used in the genre in general, but this needs to be further explored. In conclusion, Alternative shows a value of 27.9, Country 23.9 and Rock 23.8.

[5] https://www.datacamp.com/community/tutorials/R-nlp-machine-learning (last accessed on 19th March 2018)

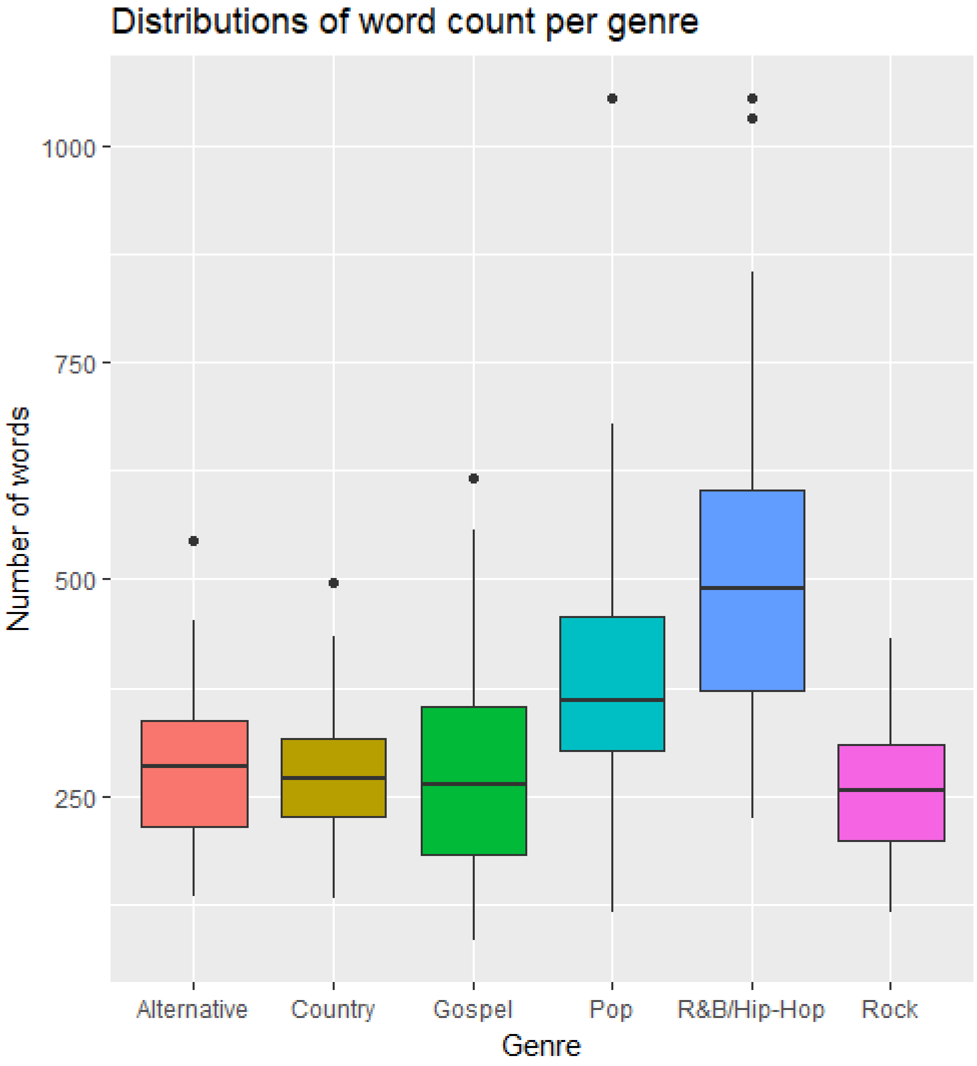

Figure 1: Distribution of word counts per genre

It is vital to stress that the used measures are only different approaches to view and examine the data. One has to be very careful about the conclusions to be drawn from that, especially because the vocabulary in general is somehow a limited value, thus the number of songs greatly influences the measurements used. In fact, the number of unique words per song tends for every genre towards zero if the sample of songs is large enough. For this reason, the measure is just illustrative. However, these values help to evaluate the goodness of the dataset and in particular the sample size. The results to be kept in mind at this point are that in comparison to the other genres the number of words per song is relatively high for Pop songs as well as the amount of vocabulary used.

The next step is to investigate the differences of the vocabularies itself. Since text analysis is a well known and long-standing method in social research, there has been much discussion about how it should be done and what should be considered. For this work, it is important to note that the sheer quantiative analysis of texts is heavily criticised since the 1950s (Kracauer 1952; Glaser and Strauss 1967). The main argument is that characters and symbols carry meanings, which highly depend on the context. For this reason, the semantic content is insufficiently or not at all covered by quantitative methods and this is why these methods are at least inappropiate to analyse texts in their wholeness. Although it is not possible to reach the whole underlying meanings, examining the frequencies of the words used could offer some insight.

To not only see which are the most frequent words per genre, but also to obtain some understanding of the content, some assumptions had to be made. First, a large part of the words used, are more or less without any substantial meaning. These words are called "stop words" and their function is to connect the elements of the sentence or they are used for grammatical reasons, but they do not have an independent meaning. For this reason, these words do not contribute to an understanding, which is why they are excluded. There are many lists to choose from, but for this analysis the lexicon called *stop_words* from the *tidytext* package is used to remove those words from the data. Second, because the goal is to analyse the meaningful words, every word with less than four characters is removed.

After that, Figure 2 shows the eight most frequent words per genre and the number of songs, in which they appear. Althought these words are out of context, there are some interesting observations to be made. First of all, these words are suprisingly similar for every genre. "Love" is always under the top three words and "time" is represented in every column. Furthermore there are many words that appear in most of the columns, like "girl", "night" or "life". The most specific words and the greatest differences in comparison to the other genres are in Gospel and R&B/Hip-Hop. The most frequent word in Gospel is "lord", the words "heaven" or "hell" are only in this genre among the top words. R&B/Hip-Hop shows very crude words like "shit", "fuck" or "bitch". Besides that, the remaining genres are overall very similar. For this reason and regarding the goal of this work, there are no considerable idiosyncrasies of Pop lyrics to be found. As said before, this result has to be seen in relation to the method.

One of the main goals of this report is to examine the lyrical complexity of different genres. Lyrical complexity means different things in different contexts. An important stylistic element of song writers - as it was mentioned already above - is the use of repetitions. If a song has many repetitions, it probably has also a high number of words. Thus, a song can have lots of words without holding much content. For this reason, counting just the number of words to assess the lyrical complexity probably misses the point. The assumption here is, that hoding much semantical content requires the use of many different words. Therefore the lexical diversity should be a more appropriate measure for the lyrical complexity. This approach was already adopted further above for the entire genres. As the dataset includes the chart position for every song, a worthwhile investigation is to study the connection between chart position and lexical diversity.

Indeed, there is a positive but weak correlation between chart ranking and lexical diversity in general. This finding is not significant, but with every position that a song is closer to the top, the number of different words used in the song decreases by an average of 0.2549 words. This means in effect, that the better the chart position, the lower the lexical diversity. It has to be clearly mentioned that this connections only hold true for the entire datset.

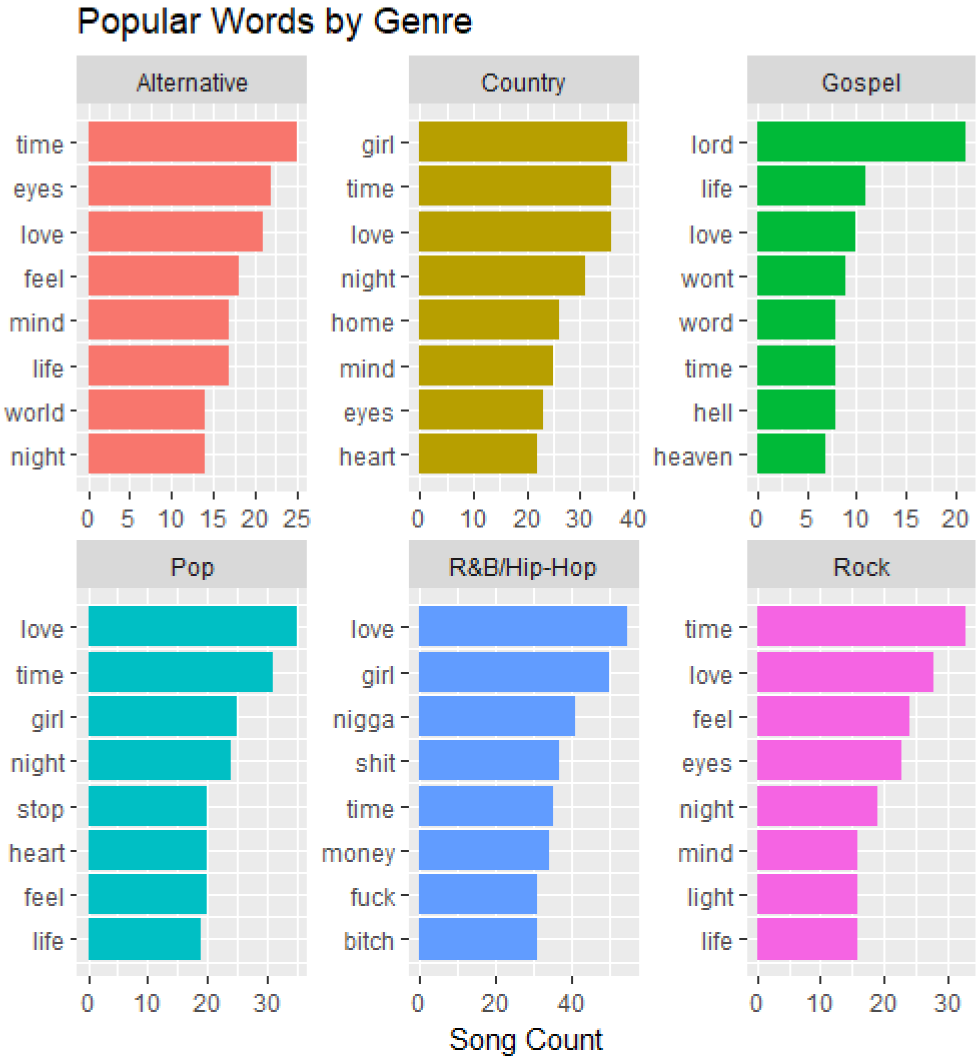

Figure 2: Most frequent words per genre

Calculated seperately for every genre, this effect is reversed for Alternative, Country and Rock. The results for R&B/Hip-Hop are similar to the overall model, for Pop songs the effect is even enforced with an average of 0.32 less words with every position closer to the top. Yet, all these results are not resilient, because none of these are significant except for Gospel, where the effect is significant with a correlation of $r = 0.308$ and a with an average decrease of 1.236 unique words used per rank closer to the top.

A substantial factor determining the amount of words used is probably the length of a song. For a profound analysis it is therefore essential to consider this factor. Unfortunately, the dataset does not hold information about the song length, hence it is necessary to find a measurement similar to that. To realise this intention, the assumption has to be made that the amount of words in general used in a song is highly positive correlated with the length of a song. This would mean that long songs tend to have a high number of words, while in short songs the number of words tends to be low. Again, because of stylistic idiosyncrasies this measure has to be seen in relation to the genre. Nevertheless, the assumption is not free of criticism, but given the circumstances, this should not be too far from reality.

After that, it is possible to create a measurement which considers on the one hand the use of repetitions, as it was done already further above, and on the other hand the length of a song with the aid of the number of words used. This measure is called lexical density and is defined as the number of unique words divided by the total number of words. Considering the goal of this report, this is the most suitable measure to quantify the lyrical complexity of a song.

A simple linear regression model of lexical density on the chart position shows a highly significant correlation coefficient of $r = 0.214$ for the whole dataset. With every rank closer to the top, the lexical density of a song decreases with an average of 0.00176. This means, that the lexical density and thereby the lyrical complexity is lowest among the ranks at the top and increases the higher the chart position is. Again, if this analysis is calculated seperately for every genre, the results differ, but they are only significant on the level of the whole song selection and never for the seperate genres.

A final analysis is the comparison of the distributions of lexical density per genre. Therefore the distributions of the lexical density are plotted for each genre in Figure 3. What can be seen here is the typical range of the lexical density in a song specific for every genre. According to whether lexical density is accepted and appropriate to measure lyrical complexity, this plot depicts the typical lyrical complexity of a song specific for the genre. First of all, the highest located distribution is that of Country songs. Country songs seem to have the highest lyrical complexity, followed by Gospel and R&B/Hip-Hop. The lowest located distributions are in descending order Pop, Rock and Alternative songs. Regarding the goal of this report it should be stressed that the lexical density of Pop songs is relatively low in comparison to other genres.

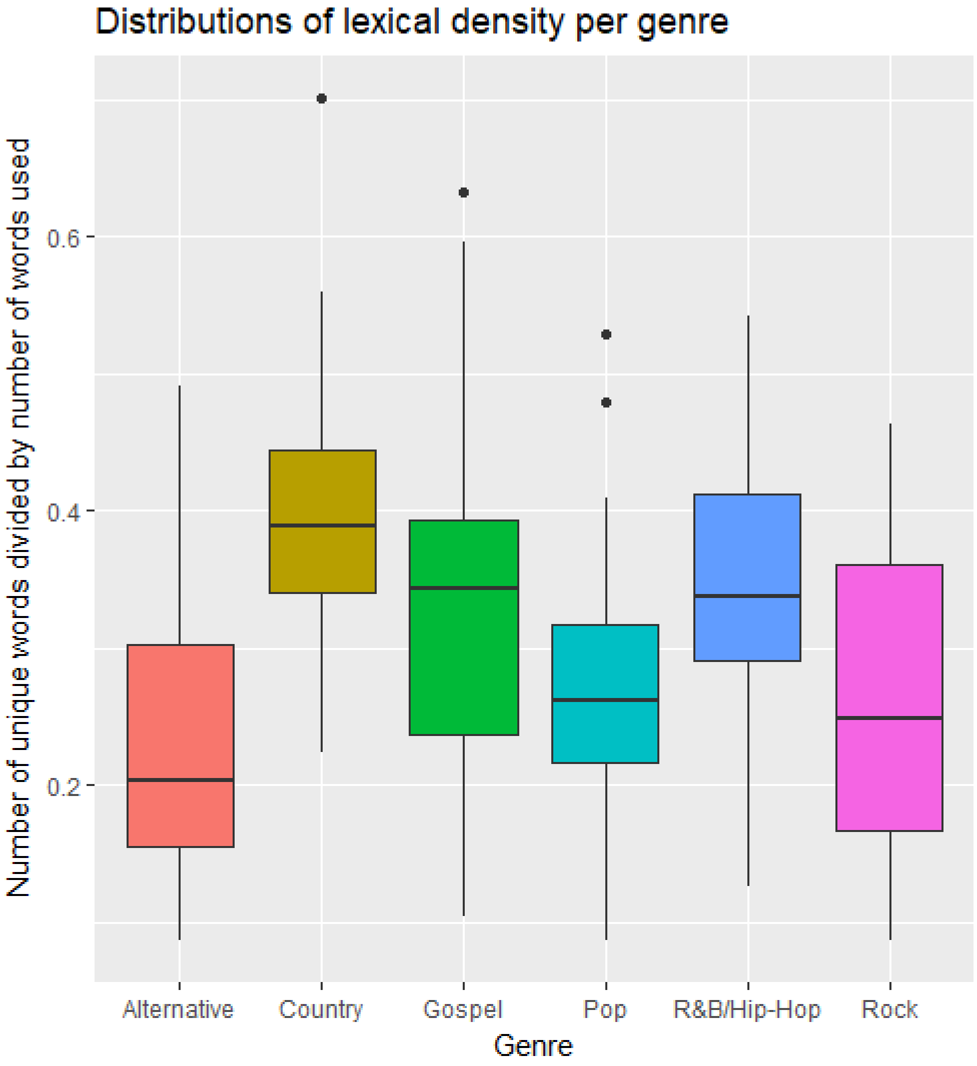

Figure 3: Lexical density per genre

5 Conclusion

To summarise, it can be said that Pop songs have in general a high number of words per song in relation to the other genres, direct after R&B/Hip-Hop songs. Also the extent of the used vocabulary is ahead of other genres except for R&B/Hip-Hop. Conversely, there were no great differences found in the most frequent words used, with only Gospel and R&B/Hip-Hop each having a very characteristical vocabulary. Studying the connection between lexical diversity and ranking, a weak but positive correlation was found. This finding reinforces when lexical density is used instead of lexical diversity. Finally, the lexical complexity of Pop songs is much below Country and R&B/Hip-Hop songs, just like Alternative and Rock.

In the interests of completeness it must be mentioned, that the limits of lyric analysis are not reached yet. The methods used in this report are relatively simple and there are more sophisticated approaches, like topic modelling or natural language processing. Moreover, the results could be much more precise with the use of additional information, like song length or the differentiation between instrumental and vocal parts. The musical aspects of a song are not considered at all, which is another point to mention.

Nevertheless, as it was shown above some expectations were met and others discarded. Low text density seems to have a positive influence on the popularity in general, not only for Pop songs, which means that successful songs tend to be less "dense". For this reason, the data are in line with the first hypothesis. On the other hand, the vocabulary, when comparing Pop against Rock, Country and Alternative, does not differ essentially. Thus, the second hypothesis was not confirmed statistically.

It must be mentioned that these results only apply to the used dataset but do not necessarily have to be true for genres at large. The results found are only explorative and therefore one should be careful with generalisations. On the other hand it becomes obvious which possibilites lie in the use of computernalisation of sciences such as the humanities. It is possible to aggregate a large amount of data which for the most part already exist and has not to be collected at high costs. Even though the hypotheses could not be fully approved, some very interesting insights were produced. Hence, the results found provide a basis for further research. As mentioned at the beginning, the analysis of lyrics is a modern branch of research. For this reason, the methods and possibilities are still in development and thus remain to be seen.

Bibliography

Chi, C., Wu, Y., Chu, W., Wu, D C., Hsu, J Y., & Tsai, R T. (2009). *The power of words: Enhancing music mood estimation with textual input of lyrics.* 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 1-6.

Gao, Y., Harden, J., Hrdinka, V. & Linn, C. (2016). *Lyric Complexity and Song Popularity: Analysis of Lyric Composition and Relation among Billboard Top 100 Songs.* 2016 23rd Static Analysis Symposium.

Kracauer, S. (1952). *The Challenge of Qualitative Content Analysis.* In: Public opinion quarterly, S. 631–642.

Glaser, B. & Strauss, A. (1967). *The Discovery of Grounded Theory: Strategies for qualitative Research.* New York: Aldine.

Declaration of Authorship

We hereby declare that the thesis submitted is our own unaided work. All direct or indirect sources used are acknowledged as references. We are aware that the thesis in digital form can be examined for the use of unauthorized aid and in order to determine whether the thesis as a whole or parts incorporated in it may be deemed as plagiarism. For the comparison of our work with existing sources we agree that it shall be entered in a database where it shall also remain after examination, to enable comparison with future theses submitted. Further rights of reproduction and usage, however, are not granted here. This report was not previously presented to another examination board and has not been published.

Leipzig, 19th March 2018

Leipzig, 19th March 2018