In Your Eyes: Identifying Clichés in Song Lyrics

Alex G. Smith, Christopher X. S. Zee and Alexandra L. Uitdenbogerd
School of Computer Science and Information Technology, RMIT University
GPO Box 2476V, Melbourne 3001, Australia

alex.geoffrey.smith@gmail.com xiashing@gmail.com sandrau@rmit.edu.au

Abstract

We investigated methods for the discovery of clichés from song lyrics. Trigrams and rhyme features were extracted from a collection of lyrics and ranked using term-weighting techniques such as tf-idf. These attributes were also examined over both time and genre. We present an application to produce a cliché score for lyrics based on these findings and show that number one hits are substantially more clichéd than the average published song.

1 Credits

2 Introduction

Song lyrics can be inspiring, moving, energetic and heart wrenching pieces of modern poetry. Other times, we find lyrics to be boring and uninspired, or *clichéd*. Some lyricists may aim to write truly original lyrics, while others are after a number one on the charts. The authors of *The Manual* (Drummond and Cauty, 1988), who have several hits to their credit, state that to succeed in achieving a number one hit one needs to "stick to the clichés" because "they deal with the emotional topics we all feel".

Despite dictionary definitions, it isn't easy to pinpoint what is cliché and what isn't. Dillon (2006) explains that linguists tend to prefer the term *idiom* or *fixed expression*. He also points out the subjective nature of the decision as to whether a phrase is a cliché, illustrating this with some frequently used phrases that are not considered clichéd, and other phrases such as 'armed to the teeth' that are, despite their relative infrequent appearance within corpora.

There is also a temporal component to whether something is cliché, since an expression would not be considered cliché on its first use, but only after widespread adoption. For song lyrics, clichés can arise due to the perceived need to make rhymes. Some words have limited possibilities for rhyme, and so using exact rhyme makes cliché more likely. Early songwriters believed that a good song must have perfect rhyme in its lyrics. However, recent thought is that alternatives, such as assonance and additive or subtractive rhymes, are valid alternatives in order to avoid clichéd writing (Pattison, 1991).

In this paper we use an information retrieval approach to defining what is clichéd in song lyrics, by using human judgements. We use statistical measures to build ranked lists of clichéd trigrams and rhymes, then combine these results to produce an overall cliché score for a song's lyrics. A simple count of the occurrences of terms in song lyrics, ranked according to frequency is likely to produce generic common phrases rather than lyric-specific terms. Therefore we investigated means of detecting typical rhymes and phrases in lyrics using a term-weighting technique. We examined trends in these attributes over musical genre and time. Using our results, we developed a cliché score for song lyrics.

The remainder of this paper is structured as follows: first, we discuss related work, then describe the data collection and preparation process. Next, our rhyme and collocation

Alex G. Smith, Christopher X. S. Zee and Alexandra L. Uitdenbogerd. 2012. In Your Eyes: Identifying Clichés in Song Lyrics. In *Proceedings of Australasian Language Technology Association Workshop*, pages 88−96.

techniques and results are shown. Finally, we present our application for cliché scoring.

3 Related Work

There are several areas of research that are relevant to our topic, such as other studies of lyrics, analysis of text, and work on rhyme. However, we have not found any work specifically on identifying clichés in either songs or other works.

Song lyrics have previously been studied for a variety of applications, including determining artist similarity (Logan et al., 2005), genre classification (Mayer et al., 2008) and topic detection (Kleedorfer et al., 2008). Whissell (1996) combined traditional stylometric measures with a technique for emotional description in order to successfully differentiate between lyrics written by Beatles Paul McCartney and John Lennon. In addition, several studies have recently appeared for retrieving songs based on misheard lyrics (Ring and Uitdenbogerd, 2009; Xu et al., 2009; Hirjee and Brown, 2010b).

Rhyme in poetry has been studied statistically for many years (for example a study of Shakespeare and Swinburne (Skinner, 1941)). More recently Hirjee and Brown (2009) (Hirjee, 2010) introduced a probabilistic scoring model to identify rhymes in song lyrics based on prior study of the complex rhyme strategies found in Hip-Hop. They define a normal or 'perfect' rhyme as 'two syllables that share the same nucleus (vowel) and coda (ending consonants). They found this method more accurate than rule-based methods and developed a *Rhyme Analyzer* application based upon their model (Hirjee and Brown, 2010a).

In our work, we considered the use of collocation extraction for finding words that frequently appear together in lyrics. Smadja (1993) described several techniques for collocation extraction and implemented these in the *Xtract* tool. The precision of this tool is tested on a corpus of stock market reports and found to be 80% accurate. Lin (1998) defined a method for retrieving two word collocations using a broad coverage parser and applied this to compute word similarties. Word n-grams have been used elsewhere as features for authorship attribution of text (Pillay and Solorio, 2011; Koppel et al., 2009), and source code (Burrows et al., 2009).

Genre	Proportion
Rock	24.70%
Hip-Hop	21.63%
Punk	13.09%
World	11.17%
Electronic	10.15%
Metal	7.00%
Pop	3.58%
Alternative	3.57%
Other	2.97%
Folk	2.12%

Table 1: Genre distribution of lyrics.

4 Experiments

Our aim was to detect and measure clichés in song lyrics. In normal text, clichés are likely to be stock phrases, such as "par for the course". Frequently used phrases can be found by looking at n-grams or collocated terms in text. The second source of cliché in song lyrics arises from rhyme pairs. Due to the typical subject matter of pop songs, and the tendency toward perfect rhyme use, particular rhyme pairs are likely to be common.

Our approach was to obtain a large collection of song lyrics, observe the effect of different formulations for ranking rhyme pairs and collocated terms, then create a cliché measure for songs based on the most promising ranking formulae. These were to be evaluated with human judgements.

4.1 Lyrics Collection

The collection was composed of a subset of lyrics gathered from online lyrics database LyricWiki[1] using a web crawler. Song title and artist meta-data were also retrieved. All lyrics were converted to lower case and punctuation removed. Digits between one and ten were replaced with their written equivalent. Duplicate lyrics were found to be a problem, for example the song 'A-OK' by 'Face to Face' was submitted three times under different names as 'A.O.K', 'A-Ok' and 'AOK'. Variations between lyrics included case, punc-

[1] http://lyrics.wikia.com

R	F(Lyrics)	F(Gutenberg)	tf-idf
1	be me	the a	baby me
2	go know	an man	real feel
3	away day	there very	be me
4	day way	than an	go know
5	way say	very their	tonight right
6	baby me	manner an	right night
7	you to	to you	apart heart
8	right night	pretty little	heart start
9	away say	cannot an	away day
10	real feel	then again	soul control
11	night light	any then	la da
12	away stay	their there	good hood
13	day say	anything any	night tight
14	town down	man than	away say
15	head dead	the of	away stay

Table 2: Top fifteen rhyme pairs ranked by frequency and tf-idf.

tuation and white space. Removing these distinctions allowed us to identify and discard many duplicates. This process resulted in a collection of 39,236 items by 1847 artists.

As our focus was on English lyrics, therefore the world music items were also discarded, removing the majority of foreign language lyrics. This reduced our collection to 34,855 items by 1590 artists.

4.2 Genre distribution

Using collected meta-data, we were able to query music statistics website last.fm[2] to determine the genre of each artist. We considered the top three genre tags for each and performed a broad general categorisation of each. This was done by checking if the artist was tagged as one of our pre-defined genres. If not, we checked the tags against a list of subgenres; for example 'thrash metal' was classified as 'metal' and 'house' fit into 'electronic' music. This resulted in the genre distribution shown in Table 1.

4.3 Lyric Attributes

In order to find typical rhymes and phrases, we applied the term-weighting scheme *tf-idf* (Salton and Buckley, 1988) to our collection. As a bag-of-words approach, tf-idf considers terms with no regard to the order in which they appear in a document. The objective of this scheme was to highlight those terms that occur more frequently in a given document, but less often in the remainder of the corpus.

The term frequency *tf* for a document is given by the number of times the term appears in the document. The number of documents in the corpus containing the term determines document frequency, *df*. With the corpus size denoted D, we calculate a term t's weight by $tf(t) \times \ln(D/df(t))$.

4.4 Rhyme Pairs

We modified Hirjee and Brown's 'Rhyme Analyzer' software in order to gather a set of all rhymes from our LyricWiki dataset. The pairs were then sorted by frequency, with reversed pairs, such as *to/do* and *do/to*, being combined. To lower the ranking of pairs that are likely to occur in ordinary text rather than in songs, we used tf-idf values calculated for rhyme pairs extracted from a corpus consisting of eighty-two Project Gutenberg[3] texts. The size of this corpus was approximately the same as the lyric collection. Note that most of the corpus was ordinary prose.

Table 2 shows that tf-idf has increased the rank of rhyme pairs such as 'right night' and introduced new ones like 'heart apart' and 'night light'. While not occurring as frequently in the lyrics collection, these pairs are given a greater weight due to their less frequent appearance in the Gutenberg texts. Note also, that the "rhyme pairs" found in the Gutenberg texts are not what one would normally think of as rhymes in poetry or songs, even though they have some similarity. This is due to the nature of the rhyme analyser in use, in that it identifies rhymes regardless of where in a line they occur, and also includes other commonly used rhyme-like devices, such as *partial rhymes* (for example, "pretty" and "little" (Pattison, 1991)). The benefit of using the Gutenberg text in this way is that spurious rhymes of high frequency in normal text can easily be filtered out. The technique may also make a rhyme analyser more robust, but that is not our purpose in this paper.

The results of grouping the lyrics by genre and performing tf-idf weighting are shown in Table 3.

[2]http://www.last.fm

[3]http://www.gutenberg.org

Rock	Hip-Hop	Electronic	Punk	Metal	Pop	Alternative	Folk
day away	way day	stay away	da na	night light	sha la	sha la	la da
way day	way say	day away	day away	la da	way say	insane brain	light night
say away	right night	control soul	say away	day away	feel real	alright tonight	day say
night light	good hood	say way	day way	real feel	say ok	little pretty	hand stand
way say	away day	getting better	way say	near fear	said head	write tonight	away day
stay away	dead head	way day	play day	say away	oh know	know oh	wave brave
night right	feel real	night right	day say	head dead	say day	way say	stride fried
way away	little bit	say away	way away	soul control	la da	la da	head dead
oh know	say play	heart start	head dead	high sky	right night	said head	sunday monday
da la	pretty little	light night	bed head	eyes lies	sha na	real feel	radio na

Table 3: Genre specific top ten rhyme pairs ranked by tf-idf.

4.5 Collocations

All possible trigrams were extracted from the LyricWiki collection and Gutenberg texts. Again, tf-idf was used to rank the trigrams, with a second list removing trigrams containing terms from the NLTK list of English stopwords (Bird et al., 2009) and repeated syllables such as 'la'. Table 4 provides a comparison of these techniques with raw frequency weighting. Similar attempts using techniques such as Pointwise Mutual Information, Student's t-test, the Chi-Squared Test and Likelihood Ratio (Manning and Schütze, 1999) did not yield promising ranked lists and are not included in this paper.

From Table 4, we can see that the difference between frequency and tf-idf in the top fifteen are both positional changes and the introduction of new terms. For example, 'i love you' is ranked fifth by frequency, but fifteenth using tf-idf. Also note how phrases such as 'its time to' and 'i just wanna' rank higher using tf-idf. This shows the influence of the Gutenberg texts - common English phrasing is penalised and lyric-centric terminology emphasised.

Interestingly, the filtered tf-idf scores 'll cool j' the highest. This is the name of a rapper, whose name appears in 136 songs within our collection. Hirjee's work involving hip-hop lyrics found that rappers have a tendency to 'name-drop in their lyrics, including their own names, nicknames, and record label and group names' (Hirjee, 2010). Examining these lyrics, we determined that many of these occurrences can be attributed to this practice, while others are annotations in the lyrics showing the parts performed by LL Cool J which we did not remove prior to experimentation.

Substituting document frequency for term frequency in the lyric collection, as in Table 6, decreases the weight of trigrams that occur repeatedly in fewer songs. As a result, this 'df-idf' should increase the quality of the ranked list. We see that the syllable repetition is largely absent from the top ranking terms and among other positional changes, the phrase 'rock n roll' drops from second place to thirteenth.

Decade	Collection
2000 - 2010	55.41%
1990 - 2000	33.49%
1980 - 1990	7.08%
1970 - 1980	2.88%
1960 - 1970	0.52%

Table 7: Time distribution of lyrics.

Several interesting trends are present in Table 5, which shows df-idf ranked trigrams by genre. Firstly, Hip-hop shows a tendency to use coarse language more frequently and genre-specific phrasing like 'in the club' and 'in the hood'. Repeated terms as in 'oh oh oh' and 'yeah yeah yeah' were more prevalent in pop and rock music. Such vocal hooks may be attempts to create catchy lyrics to sing along to. Love appears to be a common theme in pop music, with phrases like 'you and me', 'youre the one' and of course, 'i love you' ranking high. This terminology is shared by the the other genres to a lesser extent, except in the cases of hip-hop, punk and metal, where it seems largely absent. The term 'words music by' within the metal category is the result of author attribution annotations within the lyrics.

4.6 Time

There is a temporal component to clichés. There was probably a time when the lines "I'm begging you please, I'm down on my

Rank	Frequency	Frequency (filtered)	tf-idf	tf-idf (filtered)
1	la la la	ll cool j	la la la	ll cool j
2	i dont know	one two three	na na na	rock n roll
3	i want to	dont even know	yeah yeah yeah	cant get enough
4	na na na	rock n roll	i dont wanna	feel like im
5	i love you	cant get enough	oh oh oh	yeah oh yeah
6	i know you	theres nothing left	its time to	oh yeah oh
7	oh oh oh	feel like im	i wanna be	im gonna make
8	i got a	yeah oh yeah	i just wanna	theres nothing left
9	i dont want	cant live without	i just cant	dont wanna see
10	yeah yeah yeah	youll never know	give a f**k	cant live without
11	i dont wanna	two three four	dont give a	youll never know
12	up in the	oh yeah oh	du du du	let em know
13	i want you	im gonna make	i need to	im gonna get
14	i know that	never thought id	i need you	dont look back
15	you know i	dont wanna see	i love you	dont even know

Table 4: Top fifteen trigrams, ranked by term frequency and tf-idf.

Rock	Hip-Hop	Electronic	Punk	Metal	Pop	Alternative	Folk
its time to	*in the club*	its time to	its time to	its time to	i love you	and i know	and i know
i dont wanna	*give a f**k*	i dont wanna	i dont wanna	*time has come*	i dont wanna	i love you	*you are the*
i just cant	its time to	you and me	and i know	i can feel	and i know	its time to	i need to
i dont need	dont give a	i need you	*and i dont*	*in my mind*	you and me	*the way you*	close my eyes
i love you	*what the f**k*	cant you see	dont give a	its too late	oh oh oh	in your eyes	*i dont know*
yeah yeah yeah	*in the hood*	i need to	i wanna be	close my eyes	its time to	i try to	i love you
i need to	*i got a*	i love you	and i cant	*the time has*	i need you	*and you know*	*my heart is*
so hard to	*on the block*	what you want	i dont need	cant you see	yeah yeah yeah	i need you	*let it go*
i need you	*im in the*	*you feel the*	*i just dont*	*be the same*	in your eyes	*and i will*	i need you
in my eyes	i dont wanna	in your eyes	cant you see	*in the sky*	*to make you*	*you want me*	*i know youre*

Table 5: Top ten trigrams ranked by df-idf, grouped by genre. Terms that only occur in one genre's top 15 ranked list are *emphasised*.

1990-1995	1995-2000	2000-2005	2005-2010
da na	day away	away day	today away
way day	feel real	me be	town down
baby me	me be	know go	me be
go know	know go	wrong song	go know
be me	town down	day way	say away
soul control	day way	play day	right tonight
know show	oh know	right night	let better
tight night	stay away	say day	day away
down town	find mind	heart apart	alright light
day away	say away	say way	right night

Table 8: Top ten rhyme pairs ranked by tf-idf, five year period.

1990-1995	1995-2000	2000-2005	2005-2010
its time to	i got a	its time to	its time to
i got a	its time to	i got the	dont give a
i dont wanna	i dont wanna	dont give a	i got the
in the club	i need to	i got a	me and my
i need to	you need to	*i love you*	i got a
dont give a	*im in the*	and you know	i need to
and you know	and i know	*here we go*	check it out
and i know	in the back	check it out	in the back
in the back	i got the	in the back	i dont wanna
i got the	*i try to*	i dont wanna	you need to

Table 9: Top ten trigrams ranked by tf-idf, five year period. Terms that only occur in one genre's top 15 ranked list are *emphasised*.

knees", or the trigram "end of time" sounded fresh to the sophisticated audience. Fashions and habits in language also change over time. In this section we examine the rhyme pairs and trigrams across four time periods.

We queried the MusicBrainz[4] database using song title and artist in order to determine the first year of release for each song. This method was able to identify 22,419 songs, or 59% of our collection. Given the considerable size of MusicBrainz (10,761,729 tracks and 618,224 artists on 30th March 2011), this relatively low success rate can likely be attributed to incorrect or partial meta-data retrieved from LyricWiki rather than incompleteness of the database.

As shown in Table 7 our collection has a significant inclination towards music of the last twenty years, with over half in the last decade. It is suspected that this is again due to the nature of the source database — the users are likely to be younger and submitting lyrics they are more familiar with. Also, the distribution peak corresponds to the Web era, in which it has been easier to share lyrics electronically.

The lyrics were divided into 5 year periods

[4]http://musicbrainz.org

Rank	Frequency	Frequency (filtered)	df-idf	df-idf (filtered)
1	i dont know	dont even know	its time to	feel like im
2	i want to	one two three	i dont wanna	ll cool j
3	up in the	theres nothing left	give a f**k	dont wanna see
4	i know you	feel like im	dont give a	theres nothing left
5	i got a	ll cool j	i just cant	cant get enough
6	its time to	dont wanna see	yeah yeah yeah	dont even know
7	i know that	cant get enough	i need to	let em know
8	i love you	cant live without	what the f**k	im gonna make
9	and i dont	new york city	i wanna be	cant live without
10	you know i	let em know	in the club	im gonna get
11	you know what	im gonna make	im in the	dont look back
12	i dont want	two three four	check it out	new york city
13	i can see	never thought id	i just wanna	rock n roll
14	and if you	youll never know	i got a	never thought id
15	and i know	long time ago	i need you	im talkin bout

Table 6: Top fifteen trigrams, ranked by document frequency and df-idf.

from 1990 to 2010 and 2000 random songs selected from each. Rhyme pairs and trigrams were then found with the aforementioned methods, as shown in Tables 8 and 9.

4.7 Cliché Scores for Songs

We tested several cliché measures that combined the two components of our approach, being rhyme pairs and trigrams. We used precomputed tf-idf scores based on the Gutenberg collection for rhyme pairs and df-idf trigram weights. In this model, R and C are the sets of scored rhymes and trigrams respectively. The rhyme pairs and trigrams found in the given lyrics are represented by r and c. The length of the song lyrics in lines is denoted L, and $|R|$ denotes the number of rhyme pairs in the collection.

Our ground truth was based on human judgements. One coauthor prepared a list of ten songs, five of which were considered to be clichéd, and five less typical. The list was subjectively ranked by each coauthor from most to least clichéd. Spearman's rank-order correlation coefficient was used to compare each author's rankings of the songs. Two authors had a correlation of correlation of 0.79. The third author's rankings had correlations of -0.2 and -0.5 with each of the other authors, leading to a suspicion that the list was numbered in the opposite order. When reversed the correlations were very weak (0.09 and -0.1 respectively). We chose to work with an average of the two more highly correlated sets of judgements.

We report on results for the formulae shown below.

$$\frac{\sum R(r) + \sum C(c)}{L} \quad (1)$$

$$\frac{\frac{\sum (R(r))+1}{|R|+1} + \sum C(c)}{L} \quad (2)$$

$$(\frac{\sum (R(r)) + 1}{|R| + 1} + \sum C(c)) \times \ln(L+1) \quad (3)$$

$$(\frac{\sum (R(r)) + 1}{|R| + 1} \times \sum C(c)) \times \ln(L+1) \quad (4)$$

The lyrics were then ranked according to each equation.

An average rank list was prepared, and as the rankings of the third coauthor were an outlier, they were not included. Kendall's Tau and Spearman's rho were then used to compare this list to the equation rankings. These were chosen as they are useful for measuring correlation between two ordered datasets.

In order to test the accuracy of the application, we randomly selected ten songs from our collection and again subjectively ranked them from most to least cliché.

4.8 Results

Table 10 shows that the third equation produces the ranked list that best correlates with the coauthor-generated rankings. Table 11 shows the rankings obtained applying the

	Eq. 1		Eq. 2		Eq. 3		Eq. 4	
		p		p		p		p
τ	0.4222	0.0892	0.5555	0.0253	**0.7333**	0.0032	0.6888	0.0056
ρ	0.5640	0.0897	0.6969	0.0251	**0.8787**	0.0008	0.8666	0.0012

Table 10: Correlation measure results for training list using Kendall's tau (τ) and Spearman's rho (ρ).

	Ex. 1		Ex. 2		Ex. 3	
		p		p		p
τ	0.333	0.180	0.333	0.180	0.244	0.325
ρ	0.466	0.174	0.479	0.162	0.345	0.328

Table 12: Correlation measure results for random list, using Kendall's tau (τ) and Spearman's rho (ρ).

same formula to the test data. Table 12 shows the correlations obtained when compared to each author's ranked list. The results show a drop to about 50% of the values obtained using the training set.

4.9 Discussion

The third equation showed the greatest correlation with human ranking of songs by cliché. It suggests that log normalisation according to song length applied to the trigram score component in isolation, with the rhyme score normalised by the number of rhymes. Dividing the summed score by the length of the song (Equation 1) performed relatively poorly. It is possible that introducing a scaling factor into Equation 3 to modify the relative weights of the rhyme and trigram components may yield better results. Oddly, the somewhat less principled formulation, Equation 2, with its doubled normalisation of the rhyming component outperformed Equation 1. Perhaps this suggests that trigrams should dominate the formula.

The different expectations of what clichéd lyrics are resulted in three distinct lists. However, there are some common rankings, for example it was unanimous that *Walkin' on the Sidewalks* by Queens of the Stone Age was the least clichéd song. In this case, the application ranks did not correlate as well with the experimental lists as the training set. Our judgements about how clichéd a song is are generally based on what we have heard before. The application has a similar limitation in that it ranks according to the scores from our lyric collection. The discrepancy between the ranked lists may be due to this differerence in lyrical exposure, or more simply, a suboptimal scoring equation.

The list of songs was also more difficult to rank, as the songs in it probably didn't differ greatly in clichédness compared to the hand-selected set. For example, using Equation 3, the range of scores for the training set was 12.2 for *Carry*, and 5084 for *Just a Dream*, whereas, the test set had a range from 26.76 to 932.

Another difficulty when making the human judgements was the risk of judging on quality rather than lyric clichédness. While a poor quality lyric may be clichéd, the two attributes do not necessarily always go together.

Our results suggest that there are limitations in how closely human judges agree on how clichéd songs are relative to each other, which may mean that only a fairly coarse cliché measure is possible. Perhaps the use of expert judges, such as professional lyricists or songwriting educators, may result in greater convergence of opinion.

5 How Clichéd Are Number One Hits?

Building on this result, we compared the scores of popular music from 1990-2010 with our collection. A set of 286 number one hits as determined by the Billboard Hot 100[5] from this time period were retrieved and scored using the aforementioned method. We compared the distribution of scores with those from the LyricWiki collection over the same era. The score distribution is shown in Figure 1, and suggests that number one hits are more clichéd than other songs on average. There are several possible explanations for this result: it may be that number one

[5]http://www.billboard.com

Ex. 1	Ex. 2	Ex. 3	Score	Song Title - Artist
2	1	6	931.99	Fools Get Wise - B.B. King
8	8	1	876.10	Strange - R.E.M
1	7	7	837.14	Lonely Days - M.E.S.T.
3	2	4	625.93	Thief Of Always - Jaci Velasquez
9	4	9	372.41	Almost Independence Day - Van Morrison
7	6	3	343.87	Impossible - UB40
5	5	2	299.51	Try Me - Val Emmich
4	3	8	134.05	One Too Many - Baby Animals
6	9	5	131.38	Aries - Galahad
10	10	10	26.76	Walkin' On The Sidewalks - Queens of the Stone Age

Table 11: Expected and application rankings for ten randomly selected songs.

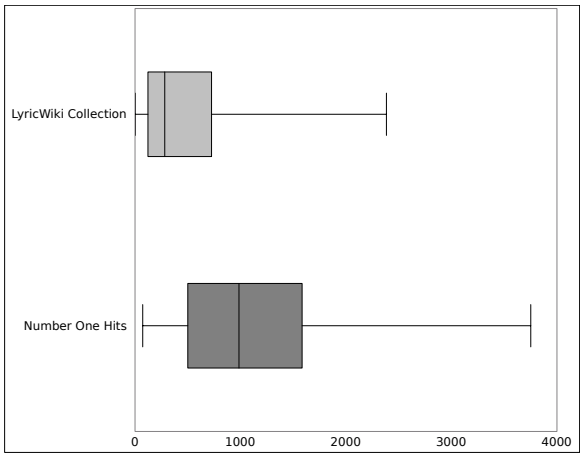

Figure 1: Boxplots showing the quartiles of lyric scores for the lyric collection from the 1990-2010 era, and the corresponding set of number one hits from the era.

hits are indeed more typical than other songs, or perhaps that a song that reaches number one influences other lyricists who then create works in a similar style. Earlier attempts to compare number one hits with the full collection of lyrics revealed an increase in cliché score over time for hit songs. We believe that this was not so much due to an increase in cliché in pop over time but that the language in the lyrics of popular music changes over time, as happens with all natural language.

6 Conclusion

We have explored the use of tf-idf weighting to find typical phrases and rhyme pairs in song lyrics. These attributes have been extracted with varying degrees of success, dependent on sample size. The use of a background model of text worked well in removing ordinary language from the results, and the technique may go towards improving rhyme detection software.

An application was developed that estimates how clichéd given song lyrics are. However, while results were reasonable for distinguishing very clichéd songs from songs that are fairly free from cliché, it was less successful with songs that are not at the extremes.

Our method of obtaining human judgements was not ideal, consisting of two rankings of ten songs by the research team involved in the project. For our future work we hope to obtain independent judgements, possibly of smaller snippets of songs to make the task easier. As it is unclear how consistent people are in judging the clichédness of songs, we expect to collect a larger set of judgements per lyric.

There were several instances where annotations in the lyrics influenced our results. Future work would benefit from a larger, more accurately transcribed collection. This could be achieved using Multiple Sequence Alignment as in Knees et al. (2005). Extending the model beyond trigrams may also yield interesting results.

A comparison of number one hits with a larger collection of lyrics from the same time period revealed that the typical number one hit is more clichéd, on average. While we have examined the relationship between our cliché score and song popularity, it is important to note that there is not necessarily a connection between these factors and writing quality, but this may also be an interesting area to explore.

References

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

S. Burrows, A. L. Uitdenbogerd, and A. Turpin. 2009. Application of information retrieval techniques for source code authorship attribution. In Ramamohanarao Kotagiri Xiaofang Zhou, Haruo Yokota and Xuemin Lin, editors, *International Conference on Database Systems for Advanced Applications*, volume 14, pages 699–713, Brisbane, Australia, April.

G. L. Dillon. 2006. Corpus, creativity, cliché: Where statistics meet aesthetics. *Journal of Literary Semantics*, 35(2):97–103.

B. Drummond and J. Cauty. 1988. *The Manual (How to Have a Number One the Easy Way)*. KLF Publications, UK.

H. Hirjee and D.G. Brown. 2009. Automatic detection of internal and imperfect rhymes in rap lyrics. In *Proceedings of the 10th International Society for Music Information Retrieval Conference*.

H. Hirjee and D.G. Brown. 2010a. Rhyme Analyzer: An Analysis Tool for Rap Lyrics. In *Proceedings of the 11th International Society for Music Information Retrieval Conference*.

H. Hirjee and D.G. Brown. 2010b. Solving misheard lyric search queries using a probabilistic model of speech sounds. In *Proc. 11th International Society ofor Music Information Retrieval Conference*, pages 147–152, Utrecht, Netherlands, August. ISMIR.

Hussein Hirjee. 2010. Rhyme, rhythm, and rhubarb: Using probabilistic methods to analyze hip hop, poetry, and misheard lyrics. Master's thesis, University of Waterloo.

Florian Kleedorfer, Peter Knees, and Tim Pohle. 2008. Oh Oh Oh Whoah! Towards Automatic Topic Detection In Song Lyrics. In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR 2008)*, pages 287–292, September.

P. Knees, M. Schedl, and G. Widmer. 2005. Multiple lyrics alignment: Automatic retrieval of song lyrics. In *Proceedings of 6th international conference on music information retrieval (ismir05)*, pages 564–569.

M. Koppel, J. Schler, and S. Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, 60(1):9–26.

Dekang Lin. 1998. Extracting collocations from text corpora. In *First Workshop on Computational Terminology*, pages 57–63.

B. Logan, A. Kositsky, and P. Moreno. 2005. Semantic analysis of song lyrics. In *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on*, volume 2, pages 827–830. IEEE.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, USA.

R. Mayer, R. Neumayer, and A. Rauber. 2008. Rhyme and style features for musical genre classification by song lyrics. In *ISMIR 2008: proceedings of the 9th International Conference of Music Information Retrieval*, page 337.

P. Pattison. 1991. *Songwriting : essential guide to rhyming : a step-by-step guide to better rhyming and lyrics*. Berklee Press, Boston.

S.R. Pillay and T. Solorio. 2011. Authorship attribution of web forum posts. In *eCrime Researchers Summit (eCrime), 2010*, pages 1–7. IEEE.

N. Ring and A. L. Uitdenbogerd. 2009. Finding 'Lucy in Disguise': the misheard lyric matching problem. In *The Fifth Asia Information Retrieval Symposium*, Sapporo, Japan, October.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24:513–523, August.

B. F. Skinner. 1941. A quantitative estimate of certain types of sound-patterning in poetry. *American Journal of Psychology*, 54(1):64–79, January.

F. Smadja. 1993. Retrieving collocations from text: Xtract. *Computational linguistics*, 19(1):143–177.

C. Whissell. 1996. Traditional and emotional stylometric analysis of the songs of Beatles Paul McCartney and John Lennon. *Computers and the Humanities*, 30(3):257–265.

X. Xu, M. Naito, T. Kato, and K. Kawai. 2009. Robust and fast lyric search based on phonetic confusion matrix. In K. Hirata and G. Tzanetakis, editors, *Proc. 10th International Society ofor Music Information Retrieval Conference*, pages 417–422, Kobe, Japan, October. ISMIR.