Emotion Painting: Lyric, Affect, and Musical Relationships in a Large Lead-Sheet Corpus

Emotion Painting: Lyric, Affect, and Musical Relationships in a Large Lead-Sheet Corpus

SOPHIA H. SUN
Wellesley College

MICHAEL SCOTT CUTHBERT [1]
Massachusetts Institute of Technology

ABSTRACT: How are lyrical emotions expressed in music? This paper explores the correlation between affect-carrying lyrics and musical features such as beat strength, duration, pitch height, consonance, and mode. Using computer-aided musicology software music21 and the NRC emotion lexicon, we conduct a corpus study on 1,895 folk and popular song lead-sheets encoded as MusicXML. The study reveals that metrical strength and note lengths are highly correlated with affects, while correlations of pitch height, consonance, and mode are in general less significant, at times contradicting previous research. Measurements of minor vs. major chordal context and tonal certainty, however, reveal certain previously unknown differences among emotional states. The paper uses a larger dataset of observations and gives greater values of significance than has appeared in symbolic corpus analysis of emotions in the past, and includes general discussions and directions for future work.

Submitted 2017 June 5; accepted 2017 November 21.

KEYWORDS: *emotion, affect, sentiment, lead sheet, lyric, corpus studies, NRC EmoLex, Wikifonia, music21*

WORDS and musical moments are generally both felt to have intrinsic emotions, but while the emotional content of text is determined at least in part from the semantics (meaning) of individual words and sentences, the emotional content of individual musical parameters is far less well defined. This lack of definition is especially acute in written musical scores which are separated from the aural and visual elements of particular performances or recordings. A major question for music perception and the psychology of music is, do particular musical features such as consonance and dissonance, beat placement, and pitch height—separated from their performative elements—carry particular emotional meanings such as exhilaration or sadness?

It might not be possible to answer such a question directly, but it may be possible to approach the answer indirectly by asking if there are musical features which are used in conjunction with texts that are associated with particular sentiments or emotions with such regularity that we might believe that the musical features are used to strengthen the effect of these affects or even carry the affect themselves.

The problem with making such connections with traditional music analytical means is that there is no perfect correlation between any musical feature and any emotional state. One might think that (at least in the Western, common practice tradition) words connoting joy would tend to be higher in pitch than the notes preceding it. It is easy to find examples to confirm this theory: in Schubert's "Gretchen am Spinnrade" (Gretchen at the Spinning Wheel), every use of "Herz" ("heart"), "Lächeln" ("smile"), "Zauber" ("magic") and "Küss" ("kiss") is at the same pitch level or higher than the proceeding note and in the upper register for the singer. But then a bit of Elvis enters the head and one hears "love" and "Darling" below the previous notes, and even an expression as madrigalistic as "falling" is set with a rising motion (see Figure 1). Individual pieces can have clear relationships among text, emotion, and musical features, but the aggregated experience of listening to common practice, folk, and popular music is far more muddled.

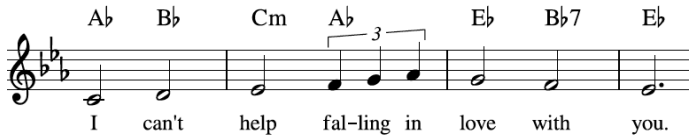

Fig. 1. Excerpt from George Weiss et al. "Can't Help Falling in Love" (major performer: Elvis Presley)

Clarity may come in the form of corpus studies, especially those involving computational analysis of large corpus datasets, whose sheer volume makes it infeasible for human researchers to analyze manually. With computational methods for extracting connections among music, text, and emotion, statistically significant trends in music-affect relationships can stand out from a large corpus even if the magnitude of the trend is quite small. For instance, the difference in average pitch for notes setting words connected to "anger" compared to those set to emotional-neutral words might be less than a semitone, but if such difference is consistent throughout the thousands of examples, it could still be highly significant.

This paper attempts such an examination of text, music, and emotion relationships by applying tools for musical corpus studies, specifically the generalized framework of music21, on a corpus of 1,895 leadsheet scores of popular and folk music in English. We find that some connections that might be considered intuitive, such as joyous texts being higher in pitch than those with negative connotations, are correct, but other intuitions, such as joy being more consonant or anger being more dissonant, are either not statistically significant or the opposite is demonstrated.

Related Works

The literature on music and emotion and text (in a non-musical context) is immense. Thus for the most part the related works cited below will only contain those for which the affective content of text in a musical context is the focus. A useful exception is Gabrielsson and Lindström (2001) which reviews the current state of research on the perceived emotional content of individual musical elements, to which we will refer often below as (G&L).

Most existing works on relating lyrics, affects, and musical properties computationally concern the goal of music mood classification. He et al. (2008) found that using the bag-of-words representation of lyrical data is effective for classifying songs into different mood clusters. Hu et al. (2010) improved the clustering accuracy by extracting lyrical features based on a word-affect lexicon translated from the Affective Norms for English Words (ANEW). Hu and Downie's system (2010) combined lyrical data with audio analysis and was able to achieve more accurate and nuanced mood classification results. McVicar et al. (2011) found strong correlation between musical features extracted from audio data and the clustering results of lyrical text, suggesting that semantic meaning expressed in the lyrics are often closely manifested in the music. These findings are mostly about the piece as a whole; our study concerns aggregating results on the local level of individual words. A recent dissertation on emotional content of musical lyrics that includes some audio features is Malheiro (2017) employing a substantially different basis for determining text affective quality than our own.

Corpus studies has shown that semantic meanings of the text can be reflected in musical features such as pitch height. A valuable contribution was Strykowski's study of text painting in the 201 madrigals composed by Luca Marenzio (2016), published last year in this journal. In Strykowski's study, he compiled a lexicon of height related words and conducted analysis specific to the repertories of madrigal. In this paper, we use similar approach but a more generalized labeling and analysis method. In his response to Strykowski's study, Sapp (2016) encourages the use of a larger corpus for similar studies in the future, advice we have taken to heart. There has not been, to our knowledge, a large scale corpus study of musical scores to corroborate the hypothesis that lyrical affects are manifested in music.

A complementary paper to ours, using the same NRC Lexicon, is Davis and Mohammad (2014) who sought to use received correlations between music and emotion (mainly from the literature review in G&L) to compose music with similar emotional states to words found in particular novels such as *Peter Pan* and *A Clockwork Orange.*

TOOLS AND CORPUS

NRC Emotion Lexicon

To automatically assess the sentiment and emotion connoted by lyrics, we used the NRC Word-Emotion Association Lexicon, or EmoLex, created by the National Research Council Canada.[2] (Mohammad and Turney: 2010 and 2013) The human-generated database contains annotations of eight emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, and trust) and two sentiments (negative and positive), for 14,182 English words (with Canadian spellings). The emotions and sentiments are collectively called "affects." Examples of certain affect-word associations are given in Table 1 below.

Table 1. Examples of word-affect associations in the NRC dataset.

	Words with this affect	**Words with this and *only* this affect**
anger	teasing, havoc, bad, repay, remand, tumultuous, criminality, consternation	sizzle, confront, rapping, harassing, rocket, tension, slash, kicking
anticipation	skewed, recreational, zeal, holiness, intimately, harbinger, birthday, plan	clock, linger, inaugural, contingent, arrive, bugle, prerequisite, destined
disgust	blatant, pillage, monstrosity, fungus, bastard, sneeze, scoundrel, disgraced	slime, dung, barf, diarrhoea, gratuitous, heartworm, excretion, catechism
fear	miscarriage, infestation, shell, mysterious, intense, smite, cyst, pray	dragon, posse, mortgagor, flood, gunpowder, eel, escaped, dart
joy	heroic, star, bride, orchestra, notable, perfect, grow, intense	whimsical, doll, beach *(the only three examples in the database)*
negative	plea, distasteful, indigent, mace, stocks, endangered, leaky, lemon	overwhelm, horribly, exhaust, trance, flagging, skip, juvenile, hypocrisy
positive	valuable, technology, enforce, concentric, aspiring, reconstruction, housekeeping, objective	clairvoyant, avatar, foreman, core, twin, acknowledgment, relaxation, muscular
sadness	exhaustion, relics, sarcasm, descent, couch, retirement, comatose, pine	widower, hospice, epitaph, pity, stricken, terminate, wrinkled, weeping
surprise	bewildered, synchronize, favorable, wonderful, warn, resignation, generosity, unintentional	overestimate, sally, zany, outrageous, trip, rapid, variable, mouth
trust	cohesive, chaplain, sadness, authorize, proctor, improve, inclusion, happy	tribe, establish, fact, inform, institute, protected, jury, consult

Examples of words in the NRC vocabulary but which are neutral include, "economics, senator, logarithmic, splice, galloping, hectic, mirror, ignite." While some affects can largely be seen to be antonyms of each other, a handful of words have, for instance, both positive and negative sentiments (81, including, "celebrity, prodigal, retirement, income, and liberal") or both joyful and sad emotions (38, including, "weight, closure, retirement, opera, and heartfelt").

music21

The analytical tools for connecting the lead-sheet corpus to the NRC Emotion Lexicon come from the music21 toolkit. The toolkit is a collection of libraries coded in Python for computer-aided musicology and music theory (Cuthbert and Ariza, 2010). The toolkit included built-in methods for analyzing pitch height, beat strength, consonance, mode, and tonal certainty (details of these attributes will be discussed further below), along with parallelizing iterators and context-searching for chord symbols which were crucial to the process.

Wikifonia Corpus

Wikifonia was a licensed collection of lead sheet scores, related to the MuseScore project (Bonte et al. 2006–13). It consisted primarily of folk and popular songs input by users of MuseScore. After the expiration of its license for distributing digital music in 2013, the Wikifonia corpus is no longer freely available. The collection of the pieces in the corpus was obtained before 2013.

This paper examines a subset of the corpus that removes non-English-language or textless pieces, songs without chord annotations, and works without at least one correctly encoded syllable division. Example titles of pieces in the corpus include “Down By the River,” “Mr. Bojangles,” and “Unforgettable.” The corpus subset contains 1,895 songs, 237,428 notes, 69,498 chord symbols and 205,724 words.

The number of observations for each sentiment differed slightly from test to test (or in the case of the major-minor correlator, significantly), however for a rough estimate of the number of observations the Wikifonia corpus allowed see Table 2.

Table 2. Number of observations for each sentiment for the Beat Strength test.

anger	2853	negative	6134
anticipation	6885	positive	12387
disgust	2163	sadness	4701
fear	3403	surprise	3470
joy	9175	trust	6010
NRC Neutral	25620	**Stopword/NRC neutral**	123464
Comprehensive Neutral	183391		

DEFINITIONS

The **vocabulary** of a test will be defined as the set of words that will be examined in the test. There are three vocabularies used in this paper. The (1) **NRC vocabulary** comprises only the 14,182 words in the NRC dataset. The (2) **Stopword/NRC vocabulary** comprises those 14,182 words plus the stopword list comprising 142 of the most common words not in the NRC dataset after removing words which might be seen to be emotive (mainly negative words such as “can’t”).[3] The (3) **comprehensive vocabulary** comprises any word found (in songs detected to be in English).

In any vocabulary, an **affect-carrying** word is one with a true value for at least one affect in the NRC dataset. All other words in the vocabulary are considered **neutral** words (for instance “abacus” in the NRC vocabulary, “the” in the Stopword/NRC vocabulary, or “nifty” in the comprehensive vocabulary); for the comprehensive vocabulary, this is a powerful and potentially incorrect assumption, however the results of the comprehensive vocabulary and the stopword vocabulary are generally quite similar. For any particular affect, **matching** and **non-matching** words are words in the vocabulary that do and do not carry this affect, respectively. As an example, for the trust emotion, “abbot” is a matching word while “sugar” is non-matching.

METHODOLOGY

Overview

In this paper, ten *correlators*, or ten musical features and their relation to lyric affects, are studied. The main procedure for examining each correlator is similar: We first label all lyrics in our corpus with the affect they connote according to the NRC emotion lexicon, and then collect the musical features of the notes or excepts set against these lyrics (called *observations*), grouping them by affect labels. These groups are then compared to the three vocabularies defied above and to other groups; we measure the magnitude, direction, and statistical significance of their differences.

We will elaborate on the system developed to carry out such analysis. Each correlator is coded as a SentimentCorrelator[4] object. For each SentimentCorrelator type and vocabulary, we parsed in parallel each of the 1,895 songs with music21 (v.4 beta) and created a BaseOneWork object associated with the affect. For instance, the PitchCorrelator creates 1,895 PitchOneWork objects.

The OneWork object then iterates over the notes of the melody of the piece. For each moment in musical time that is associated with a word, the word is examined to see if it lies in the vocabulary of the test. If it does then the object runs getObservation which returns either a numeric value or None if no value can be extracted (for instance, the first note of the piece cannot get its pitch relative to the preceding note). The range and meaning of numeric values differs by Correlator type. The observations for each moment, excepting those that returned None, are then classified and stored according to the presence of a given affect for the word. For instance, there is a list of observations for joy=Found and a list of observations for joy=notFound. There is a twenty-first list which contains all the observations for neutral words. For debugging, we occasionally generated images adding observations and abbreviated affect names for all words in the vocabulary. An example of the RelativePitchCorrelator using the NRC vocabulary running on the opening of the traditional English song “Greensleeves” is given as Figure 2:

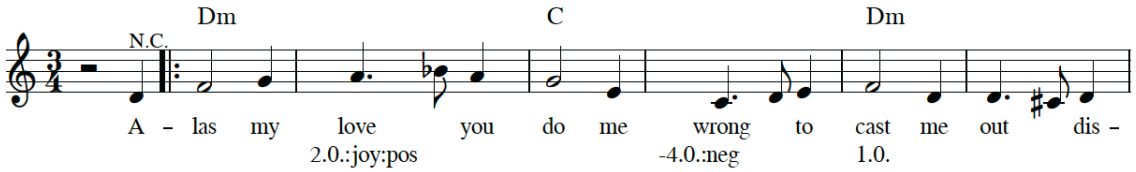

Fig. 2. Opening of Greensleeves with NRC vocabulary

Three words in the first six measures were found in the vocabulary. “Love” is marked as “joy” and “positive” while “wrong” is marked as “negative.” The 2.0 under “love” indicates that it is two semitones above the previous pitch. “Cast” is labeled with an observation but no affects, indicating that it is in the NRC vocabulary but is a neutral word.

Running the same piece with the same RelativePitchCorrelator but using the Stopword/NRC vocabulary gives a label to every word except “alas” (see Figure 3).

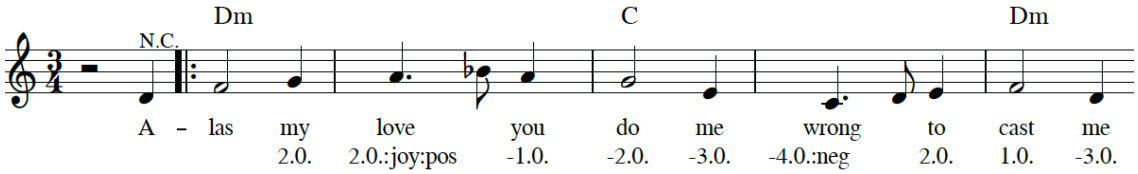

Fig. 3. Opening of Greensleeves with Stopword/NRC vocabulary

At the end of the parsing, the observations are stored on disk to prevent the necessity of rerunning the same parsing method on unchanged data. The observations are then used to compute differences among the medians of the data values and the p-value of the differences using all observations of a given type (using Welch’s t-test which is more reliable when populations have different sizes). Each affect was compared to the observations for neutral words in the vocabulary, for all non-matching words in the

vocabulary, and for each of the other nine affects. For each test, a text results file and a visualization were generated.

Details

MULTISYLLABLE AND MELISMA

All parts of this study aim to discover correlations between emotion connoted by lyrics and musical properties of a note that the particular lyric accompanies. In the case of a melisma where one lyric or syllable corresponds to multiple notes, we use the musical properties of the first note; ideally, the note attached to the accented syllable, or an average of the properties of all notes contained within the word would be used. However, since our dataset comprises mainly syllabic texts and the proportion of multi-syllable words is small compared to the total corpus (206k words and 251k syllables), we believe that the assumption that the first note will carry affect-music value is acceptable in light of the difficulties that detecting metrical accent in text would entail.

STEMMING

In early tests, we included two flavors of each vocabulary, normal and stemmed versions where words were reduced to their base words (e.g., “swimming” -> “swim”) in order to capture more words. We used the suffix-removal algorithm developed by M. F. Porter (1980), implemented in Python by Vivake Gupta. However, in observing the results of tests on a small number of scores, few new words were added and too often the stemming process merged words with opposing meanings, such as “happy” and “unhappy” or “careful” and “carelessness.” Thus the stemmed results are not discussed, but they are available for most correlators in the electronic appendix.

SIGNIFICANCE AND VISUALIZATION

For each comparison, we label p-values on a scale of zero to five stars, as follows in Table 3.

Table 3. Translation of p-values to significance stars and colors

+	–	***p-value***	***as 10^E value***	**stars**	
		>= 0.01	E-01, E-02	0	
		< 0.01	E-03	1	*
		< 0.001	E-04	2	**
		< 0.0001	E-05	3	***
		< 0.00001	E-06	4	****
		< 0.000001	E-07 or less	5	*****

The brightness of color corresponds to the number of stars for each block, or how low the p-value is. The darker the color, the more statistically significant the result.[5] A sign (+, –) and the hue of each block indicates whether the row affect’s statistic is *above* (+, blue) or *below* (–, brown) that of the column affect in the current metric.

Because each experiment involves 85 comparisons,[6] there were bound to be some values that would be statistically significant at the $p < 0.05$ threshold in a single comparison but due to the number of comparisons could easily be the result of chance. (This is the XKCD “Green-Jellybean” problem.[7]) For instance, in our “example” correlator, which plots pitch-class usage by affect (presumably meaningless in the tonal contexts of the vast majority of corpus works), there is still one comparison (trust vs. surprise) that would receive a starred p-value. In general, in this paper we will only discuss values of two-stars or above ($p < 0.001$) unless a prior hypothesis or received wisdom makes discussing a one-star or null result interesting.

RESULTS

In the following subsections, we will explain, present, and analyze the result from each of our ten correlators. The correlators are ordered roughly by the statistical significance of findings—beat strength, note length, pitch height, consonance, chord type, tonal certainty, and mode. We will also present two complementary example correlators, pitch class and word length, to help discuss our methodology and results.

Beat Strength

Beat strength as defined in music21 is a value between zero and one reflecting the metrical accent of this object in the most recently positioned Measure. For example, the four quarter notes in a 4/4 measure has beat strength 1.0, 0.25, 0.5, and 0.25 respectively.

A caveat of using these numbers must be given: The beat strength measurement in music21 is quite rough and not based directly on empirical evidence of perceived measurements of metrical accent. In particular, pieces using many tuplets or in meters besides 4/4, 3/4, etc. and common compound meters, such as 6/8, may give inconsistent or wrong results. However, these problems are mitigated in this dataset since the vast majority of songs are in the Western popular or folk traditions and 1456 of the 1895 pieces begin in 4/4.[8]

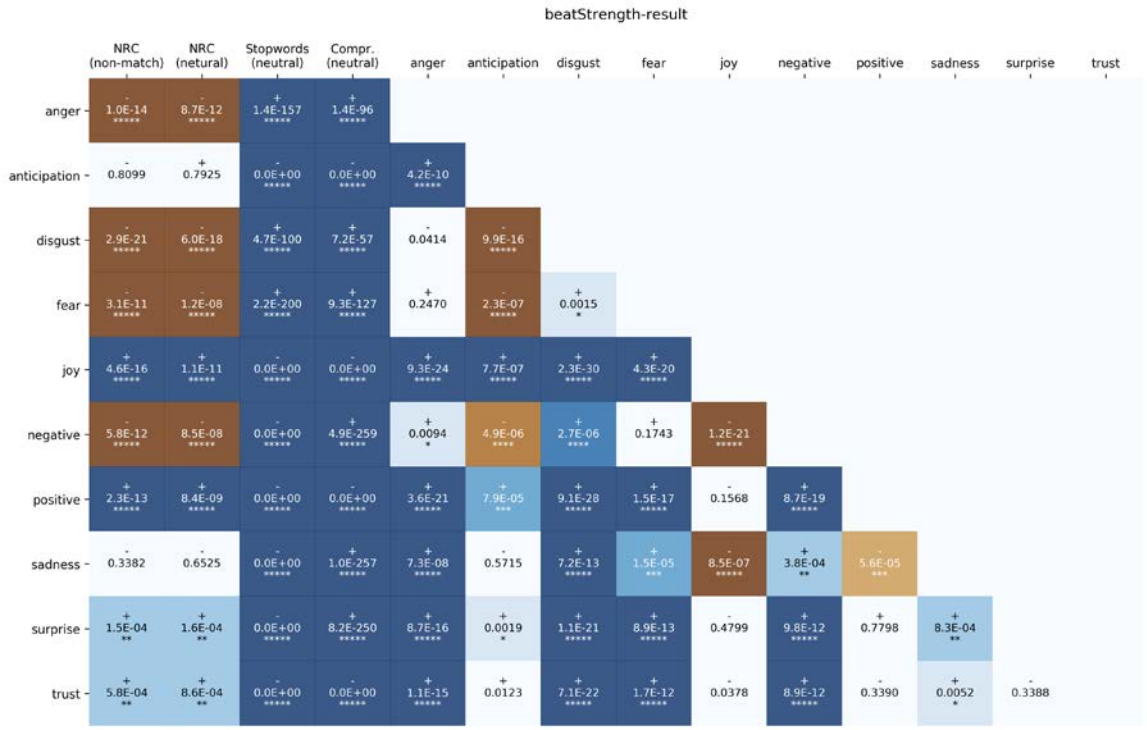

Fig. 4. Beat strength significance matrix

Each block in Figure 4 represents a comparison between the affects of the rows and the observations of the columns. Each affect is first compared to the three general vocabulary definitions defined earlier in the paper, with the comparison against the NRC vocabulary divided into a comparison against all non-matching observations and against neutral observations alone. In the remaining nine columns, the affect is compared against each of the other affects. Since the correlations between affects are symmetric, we only show a correlation between two affects once for clarity, and leave half of the matrix blank.

The values in the blocks are the p-values for the comparison. Values below the threshold for floating point representation (around 10^{-308}) are given as 0.0E+00.

We will show an example for how to interpret our results. Take row 6: the *negative* sentiment. The brown color of the *negative/Non-Match (NRC)* and *negative/Neutral (NRC)* blocks indicates that notes carrying texts with negative sentiment are consistently on weaker beats than words with other affects or words in the NRC vocabulary with neutral affect. The darkness of the shading indicates that the relationships in this column are highly significant (*****). The deep blue of the following two columns (Stopword/NRC vocabulary and Comprehensive vocabulary) indicates that negative sentiment words are, on average, found on stronger beats than stopwords, this relation is also highly significant (*****).

Moving over to the right half of the matrix, negative sentimental words are on weaker beats compared to anticipation (with weaker significance ***) and joy (*****). They are, on the other hand, indistinguishable as to beat strength with respect to anger (*) and fear (no stars). Note that the degree of *how much* one is higher or lower is not manifested in the visualization; the brightness of the color instead reflects statistical significance.

To see whether the negative sentiment differs significantly in its beat strength from any of the remaining four affects (positive, sadness, surprise, and trust), it is necessary to read down the negative *column* and to "flip" the color. So it is doubtless that negative sentiment words appear more commonly on weak beats than positive words and those connoting surprise or trust (*****) and the same relationship is also likely for words conveying sadness (**).

Looking across affect, the first two columns suggest that works carrying negative sentiments and those connoting emotion such as anger or disgust are more likely to be sung on weaker beats, while positive and joyful words are more often sung on strong beats. The third and fourth columns, on the other hand, show that words with any sentiment fall on stronger beats than stopwords, presumably because stopwords such as "the", "and", and "is" are more likely to fall on passing tones.

Mean beat strength for observations for each affect and for the observations for neutral words in the NRC and Stopword/NRC vocabulary are given in Table 4.

Table 4. Beat strength results

anger	.60	negative	.62
anticipation	.65	positive	.67
disgust	.58	sadness	.64
fear	.61	surprise	.67
joy	.67	trust	.66
NRC	.65	**Stopwords**	.41

Before moving on to other tests, it is worth noting that consistently in our results, the differences between the observations in the third and fourth columns is nearly always slight to negligible. A likely explanation is that Stopword/NRC vocabulary covers most occurrences of lyrics in our dataset; this was certainly the case for the *Greensleeves* example above. For instance, in the Stopword vocabulary, the average value was .41 while for the comprehensive vocabulary it was .45, both of which are far below the values found for any affect. As the differences between observations based on these two vocabularies are small, all further discussions of the Stopword/NRC vocabulary can be assumed to hold for the comprehensive vocabulary as well. Detailed information can be found in the data files in the electronic appendix.

Note Length

The quarterLength property in music21 was adapted as the note length feature. This property measures the length of a note by how many quarter notes it lasts. A whole note is worth 4, while an eighth note is worth .5. As with beat strength, many of the comparisons among affects and with neutral words are highly significant.

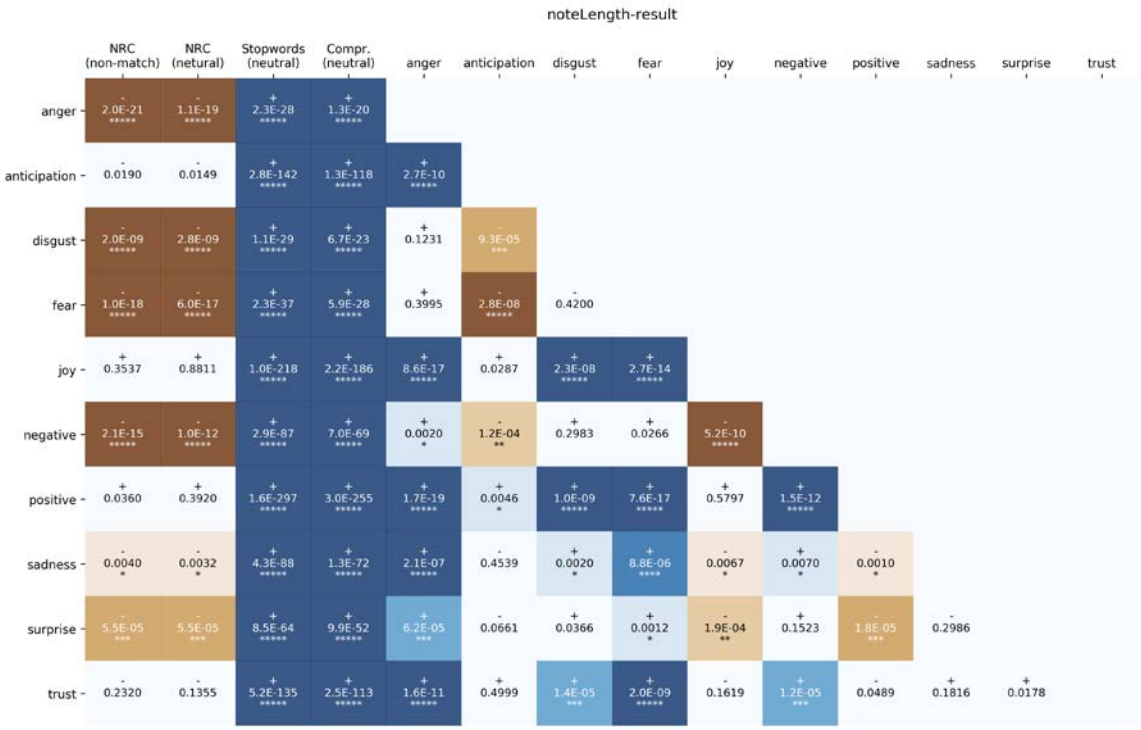

Fig. 5. Note length significance matrix

All affects use longer note values than the Stopword/NRC vocabulary (*****) (Figure 5). This makes intuitive sense since there is little reason for composers to hold out words such as "the" compared to words such as "sun." However, while none of the affects used significantly longer note values than non-matching or neutral words in the NRC vocabulary, five used significantly shorter note values, including the emotions of anger, disgust, fear, and the negative sentiment (all *****) along with surprise (***). These affects were often able to be significantly distinguished from other sentiments. On the other hand, affects such as anticipation, joy, and trust, and positive are on average granted longer note length than other sentiments (*** to *****).

Table 5 reports the mean note lengths for each affect and the first two vocabularies.

Table 5. Note length results

anger	1.13	negative	1.18
anticipation	1.25	positive	1.29
disgust	1.16	sadness	1.23
fear	1.14	surprise	1.21
joy	1.28	trust	1.26
NRC	1.28	**Stopwords**	.95

Pitch Height

Pitch height of individual notes was studied in the context of the range and register of the melody of the piece to which it belongs. For each note, we calculated the number of standard deviations of pitch height away from the average pitch height of the piece. This normalization allowed comparisons between, say, the tiny range of many Gregorian Chants and the huge ranges of the choral music of Luigi Nono (neither of which appear in the corpus). The observational value, therefore, reflects not only how high the note is but also the rarity of such a pitch height.

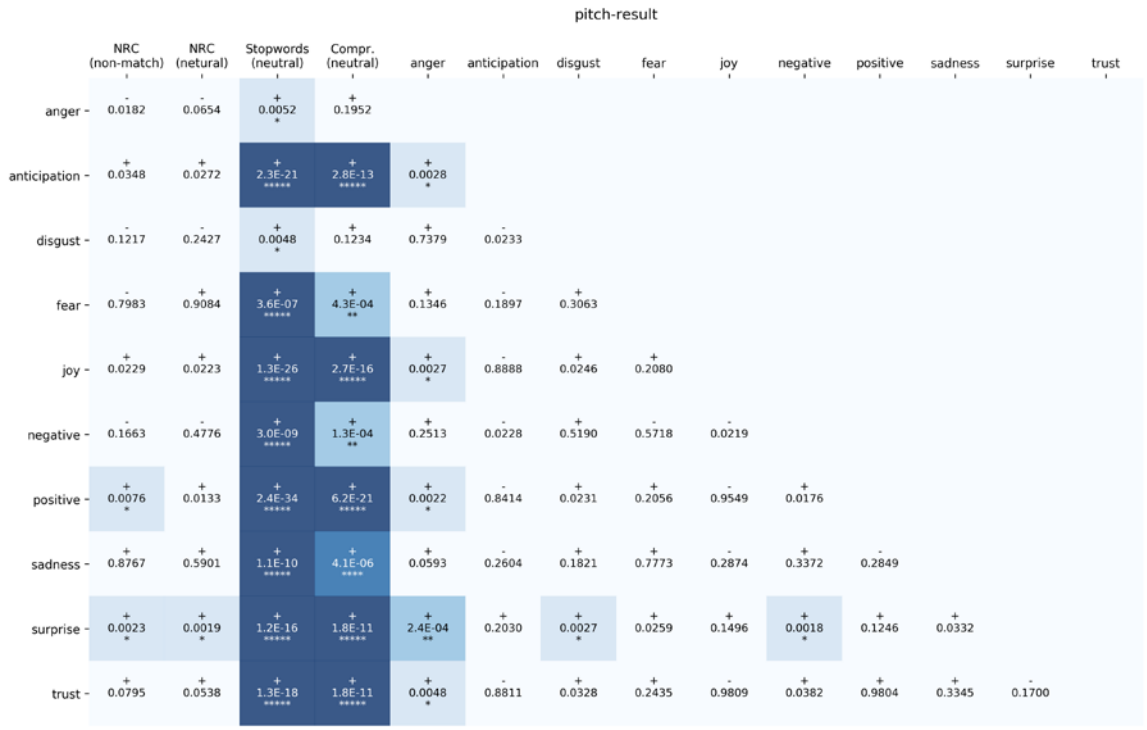

Fig. 6: Pitch height significance matrix (relative to the melody as a whole)

Most affects in Figure 6 (with the exception of anger (*) and disgust (*)) produced pitch levels higher than neutral words in the Stopword/NRC (all *****) and Comprehensive (** to *****) vocabularies, but none produced significantly higher pitch levels above the NRC vocabularies except the loosest of significance of positive and surprise (*). Weakly significant differences (*) did occur between anger and anticipation, joy, positive, and trust, as with surprise and disgust and negative words. Only between anger and surprise was there a truly significant (**) inter-affect result, with words connoting surprise being on average eight-tenths of a semitone higher. The high p-values demonstrates the low significance of pitch height in our dataset – no affects are consistently higher or lower in pitch compared to other affects. Because all of the values on the verge of significance (*) coincide in direction with those that received wisdom has expected to find (angry, disgusting, fearful, negative, and sad words being lower than positive, anticipatory, joyous, trusting, and surprising words), gathering a larger corpus to see if significance improves as the number of observations increases may be of value.

Observed values for pitch height follow in Table 6.

Table 6. Pitch height (relative to piece) observations

anger	**.**007	negative	**.**032
anticipation	.072	positive	.067
disgust	.016	sadness	.050
fear	.044	surprise	.098
joy	.070	trust	.069
NRC	.042	**Stopwords**	–.045

Values are in standard deviations away from the mean pitch of the piece. The median standard deviation for a work was 8.84 semitones, so for instance, in an average piece, the surprise value of .098 means that the expected value of a word connoting surprise would be to be sung 86/100ths of a semitone above the mean pitch of the piece. Standard deviations ranged from 3.30 to 13.83 semitones, with 25th and 75th percentile values of 8.01 and 9.63 respectively.

We also studied whether lyrical affect is correlated to pitch change on the local level, such as leaps and stepwise motion. The relative pitch measures the interval (number of semitones) between a note with lyrical affect and its preceding note. In this case, we did not normalize for range of the piece as a whole.

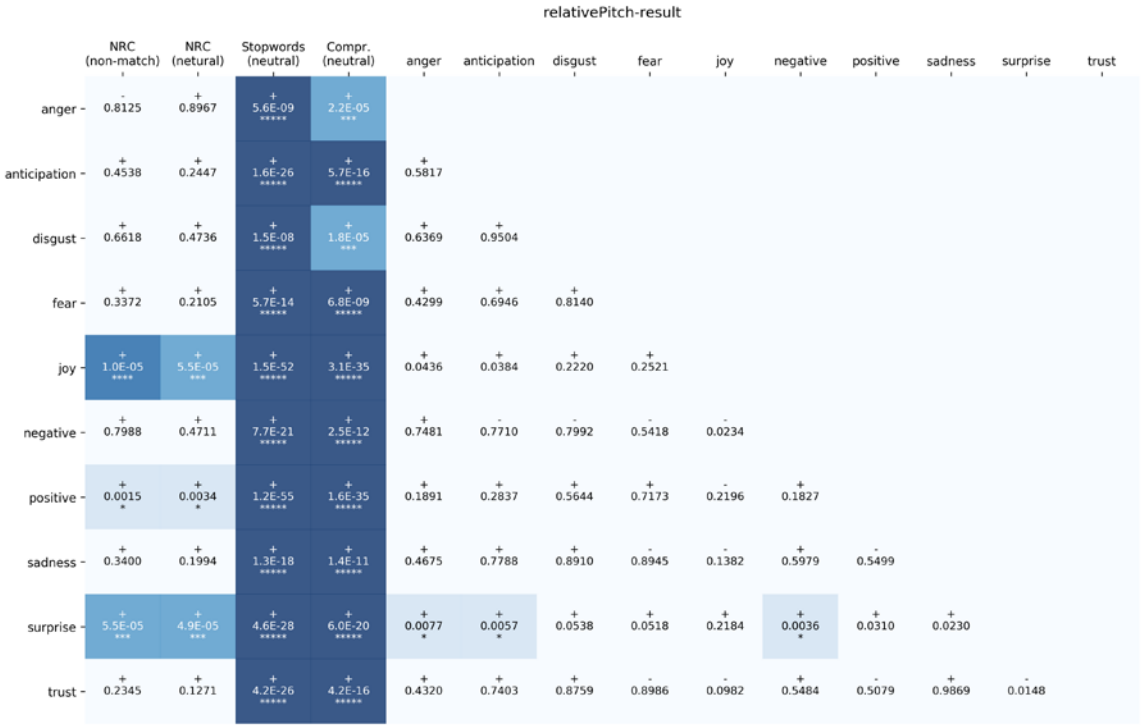

Fig. 7. Pitch height significance matrix (relative to the previous note)

While the results in Figure 7 in general are similar to the previous experiment, two new values of significance compared to non-matches and to neutral words in the NRC vocabulary stand out. Words connoting surprise are significantly higher than either set of words (***) while joyous words appear above other words in the NRC dataset and neutral words to significant extents (****, *** respectively). These observations concord with observations in the perception literature (G&L, and esp. Scherer and Oshinsky 1977), but the lack of downward directions on sad and negative affective words is a break with received wisdom.

Table 7 gives measurements, in semitones above the preceding pitch, for each affect:

Table 7. Pitch height results (relative to previous note)

anger	.34	negative	.36
anticipation	.38	positive	.43
disgust	.38	sadness	.40
fear	.41	surprise	.56
joy	.49	trust	.40
NRC	.32	**Stopwords**	–.05

Consonance

Our consonance tests take advantage of the fact that all the pieces in the corpus define at least one chord symbol representing the chord active at the time that the lyric is being sung. Two definitions of consonance are used in this study. The first definition, traditional consonance, add the melodic note to the currently active chord symbol and checks if resulting chord is consonant (that is, entirely made up of major or minor thirds or sixths, perfect fifths, or unisons/octaves above the bass). This method will, for instance, categorize all dominant-seventh chords as dissonant—consistent with traditional tonal harmony but inconsistent with much jazz and popular music harmony. A drawback of using the traditional music21 consonance methods is that enharmonic spelling matters, so G, B-flat, D will be a consonant minor triad, but G, A-sharp, D will not. A handful of pieces seem to have been input without regard to enharmonic spelling.

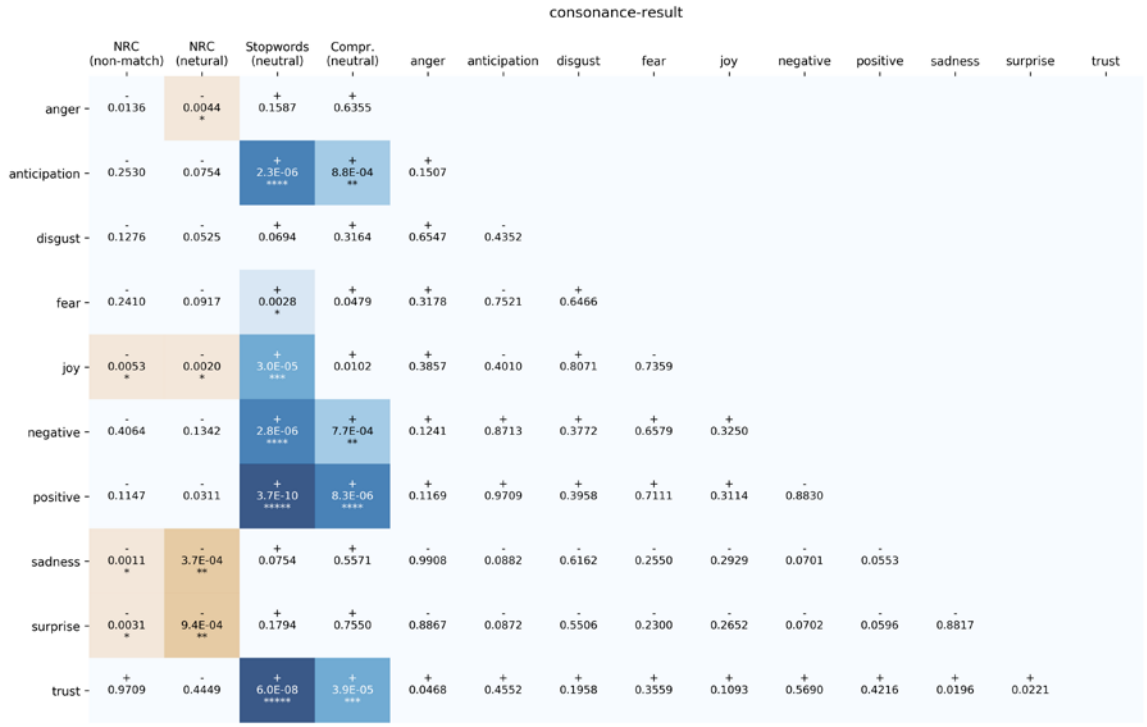

Fig. 8. Traditional consonance significance matrix

Only two emotions had statistically significant differences between their consonance values and the NRC dataset (Figure 8). Both sad and surprising words were found to be more dissonant than neutral affect words (**). The observation for sadness conforms with many received studies (G&L at 236, 241) but the lack of statistically significant distinction among lyrics with sharply different affects is unusual.

In the actual observations, a zero was given to all observations for non-consonant observations and a one for consonant observations, and no observation was recorded for moments which had no active chord symbol. Results are the mean of observations multiplied by 100, and thus represent percentage consonant. All affects were more dissonant than the NRC vocabulary and all were less dissonant than the Stopword/NRC vocabulary (probably because of the propensity to place stopwords in passing and other non-harmonic tones), but as noted above, not all of these relationships are statistically significant (Table 8).

Table 8. Traditional consonance results

anger	43.5	negative	45.3
anticipation	45.1	positive	45.2
disgust	44.2	sadness	43.5
fear	44.8	surprise	43.4
joy	44.5	trust	45.8
NRC	46.3	**Stopwords**	42.2

The second definition, which we called “loose consonance,” regards a note being consonant if the note is contained in the chord played in its context. Under this definition, a flat-5 played in a diminished chord will be considered consonant. We felt that this definition might more properly capture a larger sense of dissonance for the many pieces from the jazz and popular music traditions that use many seventh and ninth chords as stable harmonies.

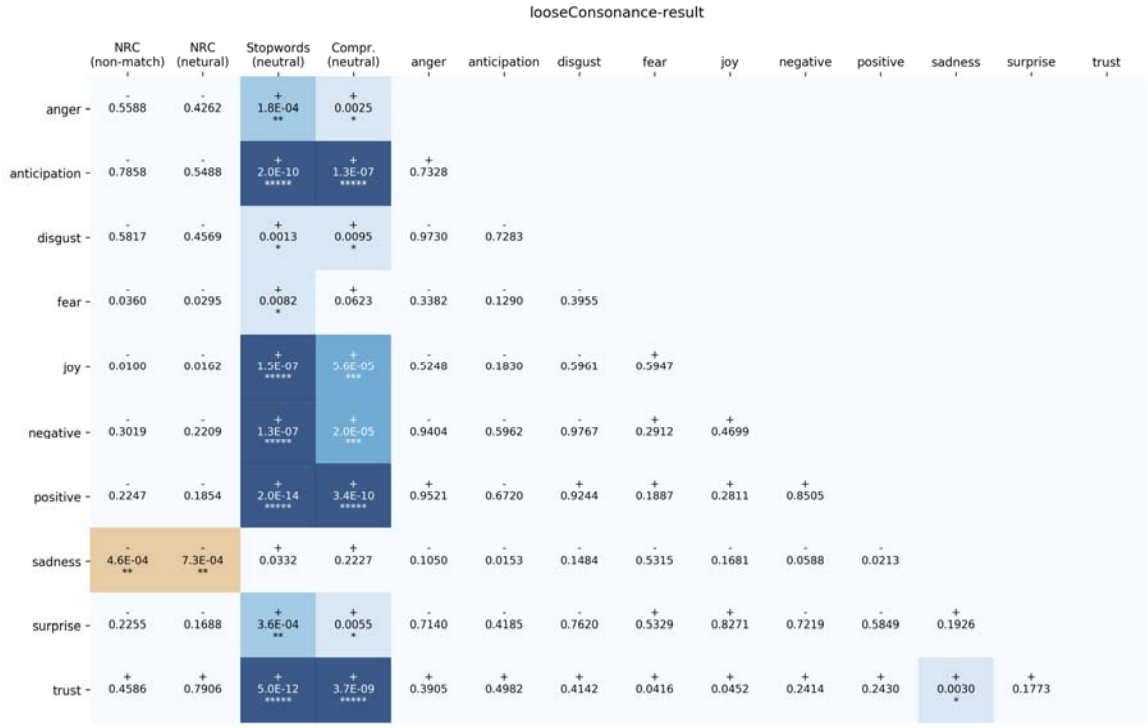

Fig. 9. Loose consonance significance matrix

The results in Figure 9 and Table 9 were similar to the traditional consonance metric, though the difference between sad lyrics and all other lyrics in the NRC vocabulary became more significant (**).

Table 9. Loose consonance results

anger	74.4	negative	74.3
anticipation	74.7	positive	74.4
disgust	74.3	sadness	72.7
fear	73.3	surprise	74.0
joy	73.8	trust	75.2
NRC	75.0	**Stopwords**	71.2

Major-minor chord context

Continuing to use the chord symbol as context, we examined whether lyrics with happy and sad emotions and positive and negative affects differ in their presence in the chordal context with respect to whether the chord is a major or a minor chord. (All other chords, including sevenths, were ignored). Prior work had suggested that happier lyrics should be associated with major chords and sad lyrics with minor chords (G&L at 228–9 under “mode”).

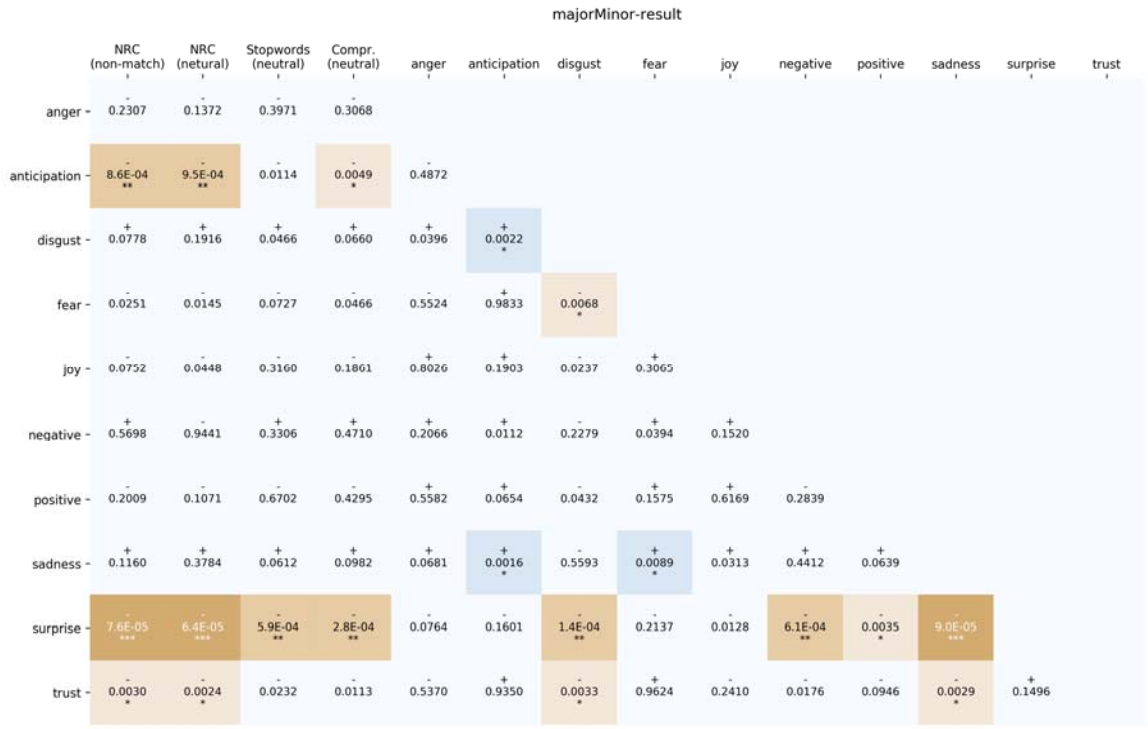

Fig. 10. Major-minor chord significance matrix

The test found significantly lower use of major chords for words associated with anticipation (**) and surprise (***) than for words in the NRC dataset (Figure 10). More interesting than the significant results are the null hypotheses that cannot be rejected. Positive lyrics were not statistically different from negative lyrics (and even the p = .28 result showed positive lyrics are more often accompanied by minor chords than negative lyrics). Among the significant results, lyrics associated with surprise were more often accompanied by minor chords than lyrics associated with sadness (***).

The results table (Table 10) presents the total percentage of observations that are major chords.

Table 10. Major-minor chord results

anger	79.7	negative	81.2
anticipation	78.9	positive	80.3
disgust	82.7	sadness	82.0
fear	78.9	surprise	77.4
joy	80.0	trust	79.0
NRC	81.3	**Stopwords**	80.6

Tonal Certainty

The next three tests use music21's key analysis methods. They use a variation of Carol Krumhansl and Mark A. Schmuckler's probe-tone key-finding algorithm (Krumhansl 1990), which statistically correlates pitch occurrences in a piece with the profile of a major key or a minor key.

Total certainty is a metric reflecting the confidence of key matching, based on the correlation coefficient of the piece with its most likely key and the difference between that coefficient and that of the second-most likely key; it ranges between 0 and 3, with 3 being the most tonally certain result. The method is implemented as a feature in the music21 feature code (Cuthbert, Ariza, Friedland 2011). A single tonal certainty score was generated for each piece, but the same piece will be weighed more heavily if a sentiment appears multiple times in the same piece.

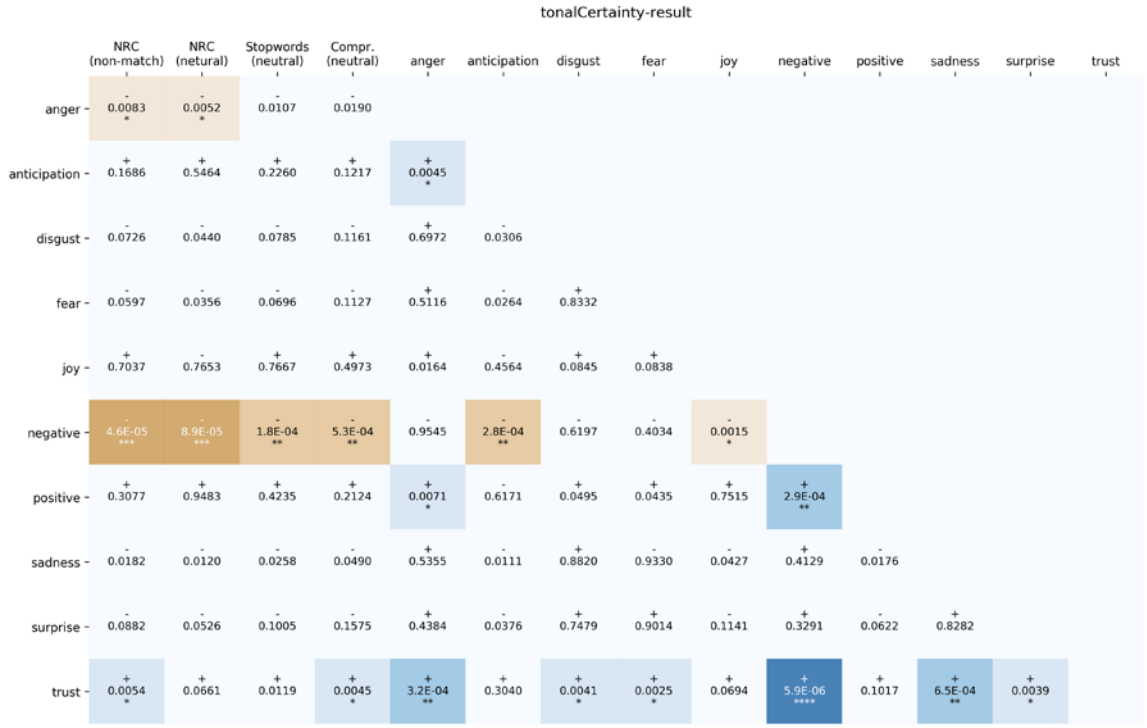

Fig. 11. Tonal certainty significance matrix

Ignoring values from Figure 11 at the lowest level of significance (*), only the negative sentiment differs significantly from other words in the NRC vocabulary (***) and the Stopword/NRC vocabulary (**), with a difference in the metric of .013 and .011 respectively. (Anger and fear had the same difference, but owing to fewer observations, were not significantly different). Between certain affects, there were differences in the tonal certainty (each will be given with the affect more clearly in a key listed first): anticipation vs. negative (**), positive vs. negative (**), trust vs. anger (**), trust vs. negative (****), and trust vs. sadness (**). In each of these cases, the direction of the difference conforms to prior hypotheses, that positive and happy emotional states would be more connected to a secure sense of key than those of negative states. Sadness, fear, and surprise were nearly indistinguishable from each other, as Table 11 makes clear.

Table 11. Tonal certainty results

anger	1.043	negative	1.043
anticipation	1.058	positive	1.056
disgust	1.046	sadness	1.047
fear	1.047	surprise	1.048
joy	1.055	trust	1.062
NRC	1.056	**Stopwords**	1.054

Mode

Each of the two mode correlator tests the correlation between the mode of a piece (major or minor), as determined by the key-finding algorithm described above and the affect of its lyrics. We assigned 0 to pieces in minor and 1 to pieces in major and took the mode average of melodies corresponding to lyrics of certain affect, weighted by the number of occurrences of this affect.

The global mode test analyzes the entire song and reports the mode of the work.

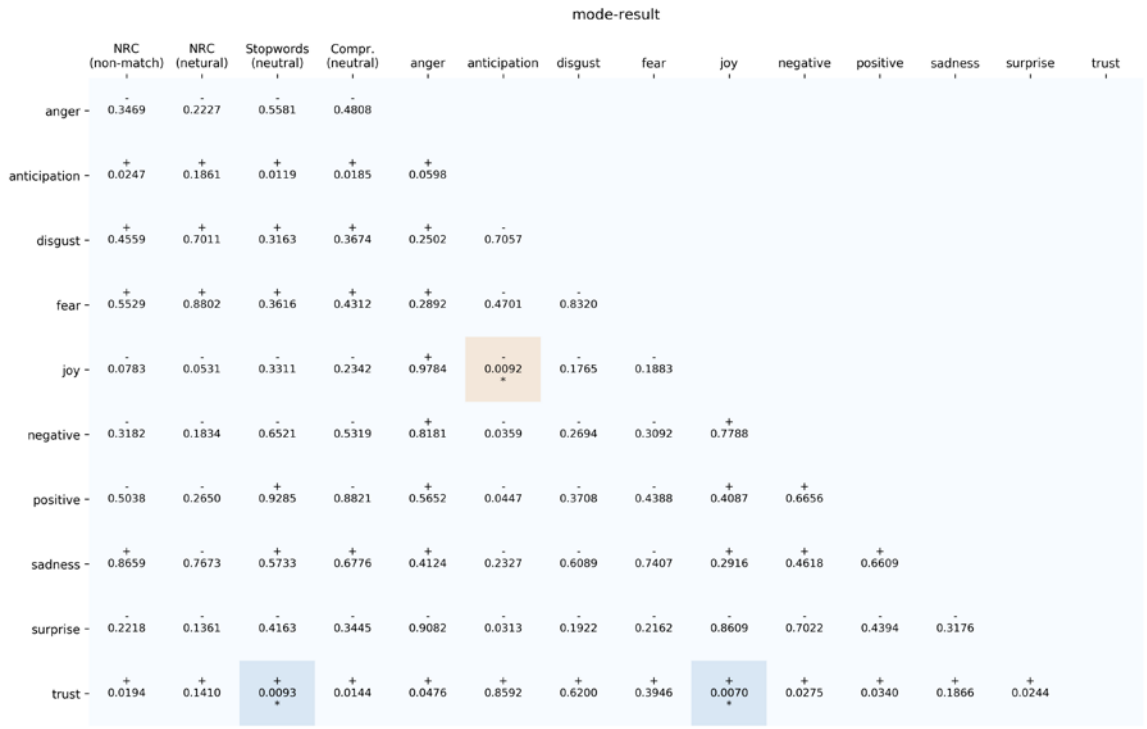

Fig. 12. Overall mode significance matrix

No significantly different results except at the lowest (*) level were detected (Figure 12). Thus, the hypothesis that minor mode pieces would use more negative and sad words than major mode pieces cannot be sustained. In fact in the corpus there was a higher percentage of joyous words paired with minor mode pieces than sad words, however, the result was not statistically significant (Table 12).

Table 12. Overall mode results

anger	.698	negative	.700
anticipation	.717	positive	.703
disgust	.713	sadness	.707
fear	.710	surprise	.697
joy	.698	trust	.719
NRC	.709	**Stopwords**	.703

The local or nearby mode test, on the other hand, only considers the detected key of a window of two measures on either side of a lyric-affect instance, in an attempt to give a more fine-grained analysis of the mode in effect at the moment of the affect-carrying word. (On windowed analysis see Sapp 2001).

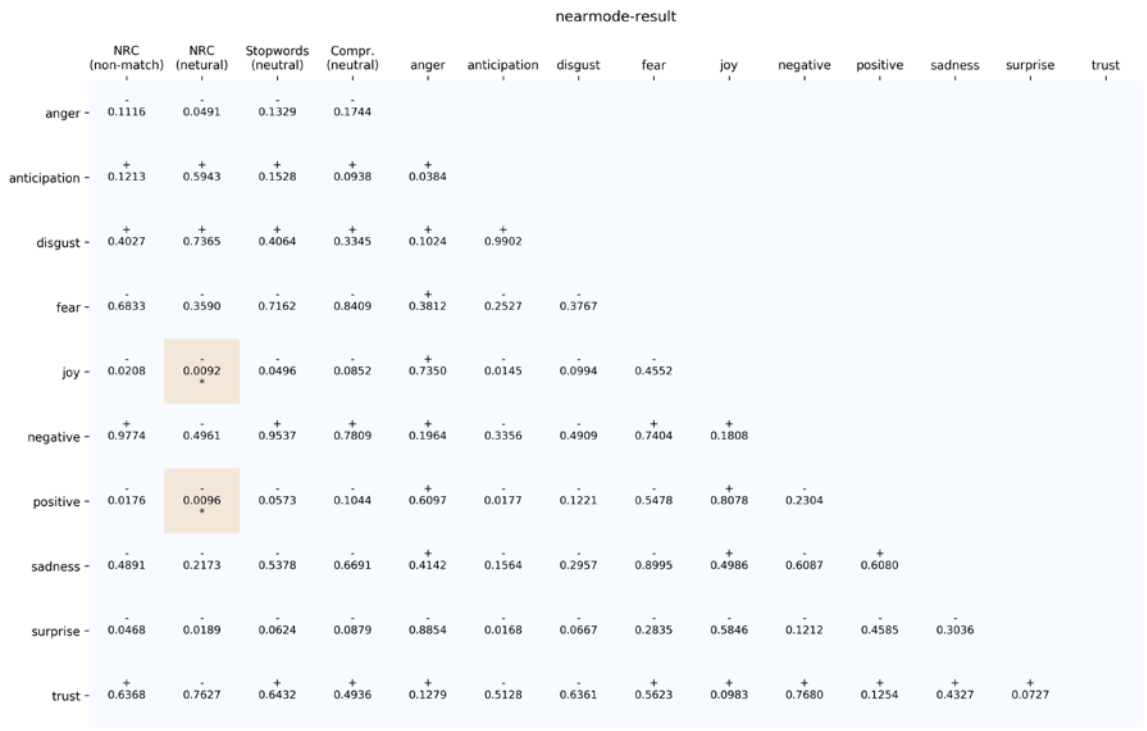

Fig. 13. Nearby mode significance matrix

The nearby mode results (Figure 13 and Table 13) showed generally the same lack of statistical significance as the global mode results. However, the two weakly significant results, connecting joyous and positive texts with minor mode, go against previous studies (G&L at 237). As these results have been repeated many times with listeners in the West of different ages and musical backgrounds, it may be that the emotional aspects of mode are decidedly not to be found in the text but are an example of a purely musical affect.

Table 13. Nearby mode results

anger	.611	negative	.626
anticipation	.633	positive	.616
disgust	.634	sadness	.620
fear	.622	surprise	.610
joy	.615	trust	.628
NRC	.630	**Stopwords**	.625

Correlators used in testing and debugging

In the process of building and debugging the corpus system for this paper, we felt it helpful to build a correlator that we could use to reliably give null results to validate our methods and statistical tests. The idea was to return a result obtained from the data itself but which was extremely unlikely to produce significant results. Our original idea was to return the seemingly meaningless metric of pitch class[9] minus the number of letters in the lyric. However, this ended up producing results that were far from meaningless, so we separated the two parts of the metric into a PitchClassCorrelator and a WordLengthCorrelator.

Although pitch class might carry semantic meaning in certain individual post-tonal works, as we expected for pieces that are almost entirely based in common-practice harmony relationships between sentiment and pitch class were non-existent, all hovering just below 5.5, the expected average for an even pitch distribution. The relationship between trust and surprise (*) should be read as a caution for over-interpreting otherwise seemingly low p-values in the context of multiple comparisons (Table 14 and Figure 14).

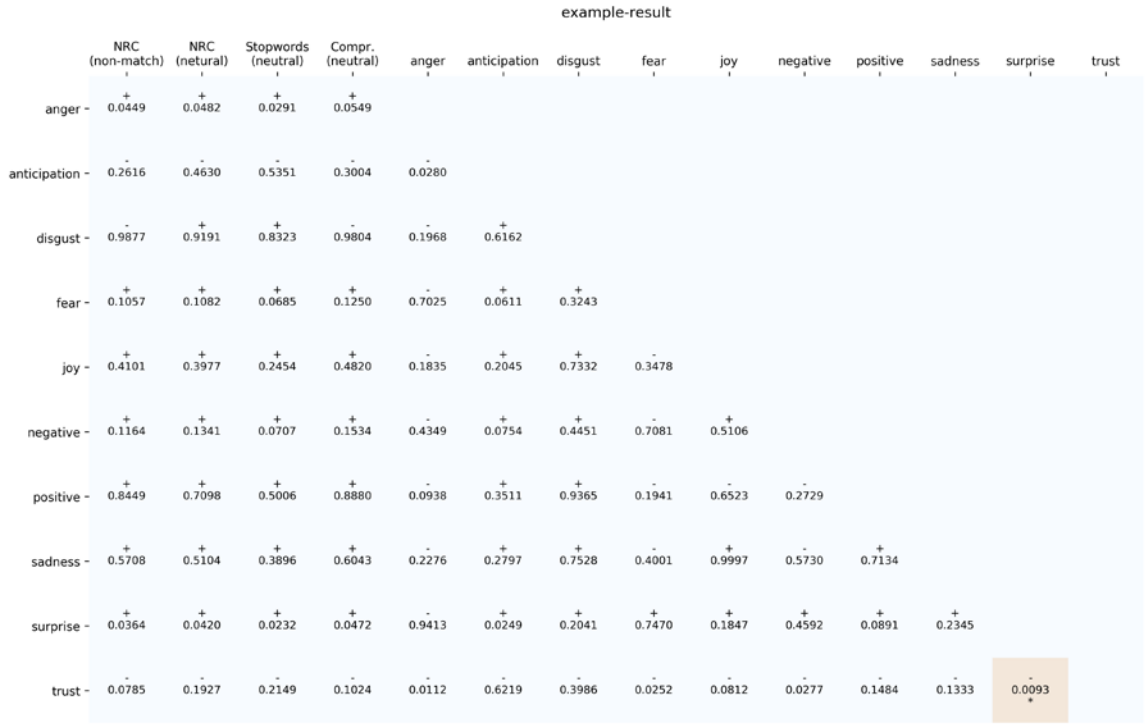

Fig. 14. Pitch class significance matrix

Table 14. Pitch class averages

anger	5.4	negative	5.4
anticipation	5.2	positive	5.3
disgust	5.3	sadness	5.3
fear	5.4	surprise	5.4
joy	5.3	trust	5.2
NRC	5.3	**Stopwords**	5.3

Word length, on the other hand, ended up having extremely significant differences among many affects (Figure 15), with the average difference between words connoting disgust and those connoting trust being nearly a full letter (Table 15). These results overwhelmed the null result for pitch class when the two correlators were combined and were the reason they needed to be separated. Because these results do not concern music per se, we leave them here for other disciplines' scholars without further discussion.

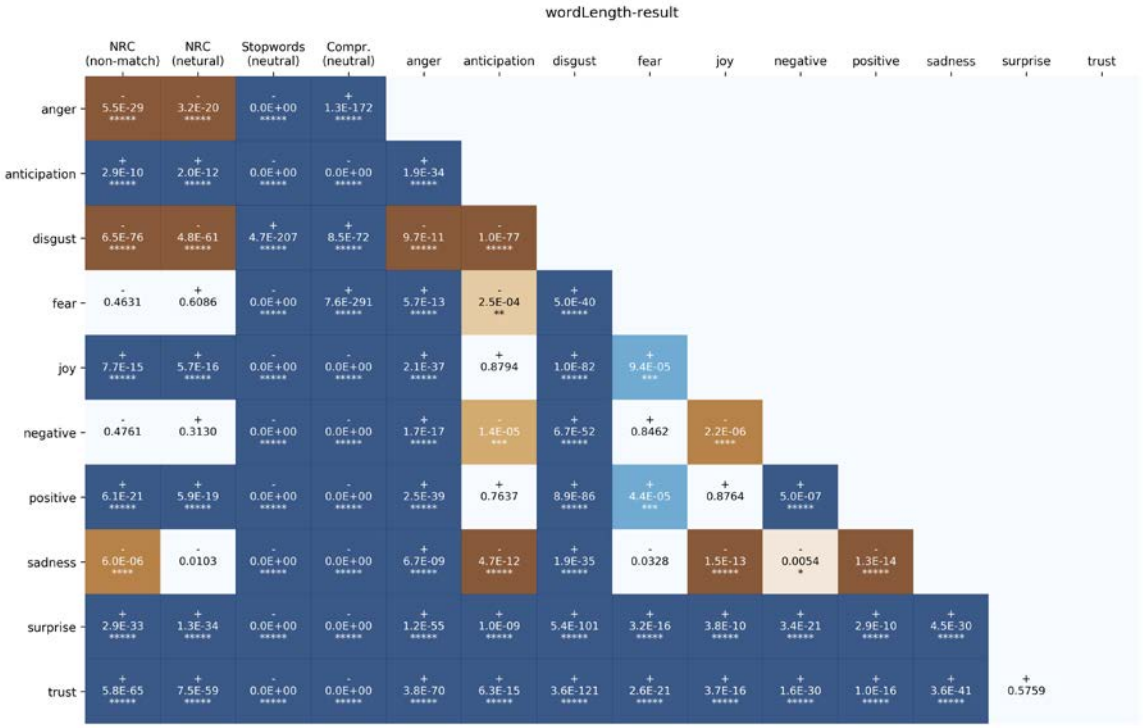

Fig. 15. Word length significance matrix

Table 15. Word length averages

anger	4.50	negative	4.77
anticipation	4.91	positive	4.92
disgust	4.23	sadness	4.71
fear	4.79	surprise	5.10
joy	4.92	trust	5.11
NRC	4.78	**Stopwords**	3.16

DISCUSSION AND FUTURE WORK

The correlators in this paper may be improved somewhat if, as noted above, the note and context of the accented syllable of a word instead of the first syllable were able to be taken into account. The corpus studied in this paper has been restricted to English-language songs, and thus its results are not necessarily applicable to other languages. We considered using the machine-translated versions of the NRC Emotion Lexicon to examine works in other languages, however, whether we had or did not have confidence in the utility of a translated vocabulary became moot when we realized that the overall corpus did not have enough works in any language except English to make a large repertory for analysis. Scholars making use of this work should also be aware that the pieces examined were selected entirely because they were available at the time. Although to us they seem to be somewhat representative of many of the most commonly heard examples of popular and folk music, the corpus can only be noted as representative of the types of pieces chosen by people who encode leadsheets on the internet. The difficulties in assembling even larger and more representative corpora are too well known to rehash, but such attempts should continue to be made. Similarly, despite the awe-inspiring amount of work that went into creating the NRC Emotion Lexicon, we were sometimes surprised by which words were not found in the NRC vocabulary and by some of the affect associations that had or not been made. Continued work on expanding resources like these should remain a priority.

In the future, we would like to look at other musical traditions to see if these results remain consistent across time and cultures. There are other features that have been associated by others with

emotional states, such as pitch direction, rhythmic regularity, and melodic consonance that would be worthy of examining on this or another corpus in the future.[10] Finally, it must be reiterated that this study is looking at the affectual content of individual words, but the emotion and sentiment of a text is determined not just by the individual words but by their interaction. (How different the sentiment of "love" is if "I love you" has a tiny "don't" interjected in it.) Recent progress in the field of natural language processing has enabled sentiment analysis on larger textual elements, and we look forward to see further studies on musical expressions of emotions aided by machine learning and other computational techniques. The current state of the art need not hamper the understanding of lyric-music relationships on the more local level similar to word painting that this paper explores.

This paper contributes results of relating text to music on ten features related to pitches, durations, meter, mode, and chord context on a corpus of nearly 2,000 pieces of folk and popular music. It demonstrates that while some received notions of music and emotion relationships remain on firm ground, others, such as the low-level connection between minor chords and sadness, need to be reexamined. It also uncovers many heretofore unknown relationships between text and music across emotions and sentiments, particularly in the domains of metrical placement and duration of notes. As corpus studies of text, music, and emotion continue to develop, so too will our understanding of how many seemingly disparate elements come together to create the human, emotional aspects of hearing sung music.

ACKNOWLEDGEMENTS

This article was layout edited by Kelly Jakubowski. The development of music21 was made possible through grants from the Seaver Institute and the National Endowment for the Humanities.

NOTES

[1] Both authors contributed equally to this paper. Order of authors determined by coin flip. Correspondence may be directed to either or both authors: ssun2@wellesley.edu, cuthbert@mit.edu.

[2] For more details on the lexicons, see http://saifmohammad.com/WebPages/lexicons.html.

[3] The list of stopwords used is an adapted version of the English Stopwords list from Ranks NL found at http://www.ranks.nl/stopwords

[4] At the time we wrote the code accompanying this paper, we were not aware of the distinction the NRC authors made between affects that are sentiments (e.g., "positive") and affects that are emotions (e.g., "joy"), so our code uses the terms "sentiment," "emotion," and "affect" interchangeably. These distinctions are otherwise maintained in this paper beyond the first word of the title.

[5] Some values in the significance matrices report different values than the chart because of rounding. For instance, 9.999E-05 might be reported as the rounded value 1.0E-04 but still receive the (***) coloring.

[6] In each experiment, the ten sentiments are each compared against four neutral or non-matching values (40 total), and there are 45 independent sentiment-to-sentiment comparisons.

[7] See https://xkcd.com/882/

[8] Number of pieces by meter: 4/4: 1412; 3/4: 266; 2/2: 154; 6/8: 37; 2/4: 16; 12/8: 7; 5/4: 2; 6/4: 1

[9] The pitch class for any given note is a number from 0 to 11 relating to pitch without regard to spelling, so C/B#=0, C#/Db = 1, D=2, etc. to B/Cb = 11.

[10] Among the potential correlators considered and not used was a TempoCorrelator, since previous studies of musical emotion had shown connections of fast tempo with happiness (Motte-Haber, 1968). However, with only 18 cases of tempo in the corpus, results were unlikely to be statistically significant.

Furthermore, while gross tempo categories would not be difficult to add to many of the pieces in the corpus, fine-grained values of tempo would vary from performance to performance.

REFERENCES

Bonte, T., et al. (2006-13) Wikifonia (internet resource), accessed 2013.

Cuthbert, M. S., & Ariza, C. (2010). music21: A Toolkit for Computer-Aided Musicology and Symbolic Music Data. *Proceedings of the 11th International Conference on Music Information Retrieval,* 637–642.

Cuthbert, M. S., Ariza, C., & Friedland, L. D. (2011) Feature Extraction and Machine Learning on Symbolic Music using the music21 Toolkit. *Proceedings of the International Symposium on Music Information Retrieval*.

Davis, H., & Mohammad, S. M. (2014). Generating Music from Literature. *Proceedings of the EACL Workshop on Computational Linguistics for Literature.* Gothenburg, Sweden: Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-0901

Gabrielsson, A., & Lindström, E.. (2001). The influence of musical structure on emotional expression, in P. N. Juslin and J. A. Sloboda (Eds.) *Music and Emotion: Theory and research,* (pp. 223–248). Oxford: Oxford University.

He, H., Jin, J., Xiong, Y., Chen, B., Sun, W., & Zhao, L. (2008). Language feature mining for music emotion classification via supervised learning from lyrics. *Proceedings of Advances in the 3rd International Symposium on Computation and Intelligence (ISICA 2008).* https://doi.org/10.1007/978-3-540-92137-0_47

Hu, Y., Chen, X., & Yang, D. (2009) Lyric-based song emotion detection with affective lexicon and fuzzy clustering method. *Proceedings of the 10th International Conference on Music Information Retrieval*. 123-128.

Hu, X., & Downie, J. S. (2010). Improving Mood Classification in Music Digital Libraries by Combining Lyrics and Audio. *Proceedings of the 10th Annual Joint Conference on Digital Libraries - JCDL '10*. https://doi.org/10.1145/1816123.1816146

Krumhansl, C. L. (1990). *Cognitive Foundations of Musical Pitch*. Oxford: Oxford University Press.

Malheiro, R. (2017). *Emotion-based Analysis and Classification of Music Lyrics.* PhD Dissertation, University of Coimbra, Portugal.

McVicar, M., Freeman, T., & De Bie, T. (2011). Mining The Correlation Between Lyrical And Audio Features And The Emergence Of Mood. *Proceedings of the 12th International Society for Music Information Retrieval Conference*, 783-788.

Mohammad, S., & Turney, P. (2013). Crowdsourcing a Word-Emotion Association Lexicon, *Computational Intelligence*, *29* (3), 436-465, 2013. https://doi.org/10.1111/j.1467-8640.2012.00460.x

Motte-Haber, H. de la. (1968). *Ein Beitrag zur Klassifikation musikalischer Rhythmen.* Cologne: Arno Volk.

Rijsbergen, C .J. van, Robertson, S. E., & Porter, M. F. (1980). *New models in probabilistic information retrieval.* London: British Library. (British Library Research and Development Report, no. 5587).

Sapp, C. (2001). Harmonic Visualizations of Tonal Music. *Proceedings of the International Computer Music Conference.* 423–430.

Sapp, C. (2016). Suggestions for Future Corpus-Based Text Painting Analyses: A Response to Strykowski. *Empirical Musicology Review*, *11*(2), 120-123. https://doi.org/10.18061/emr.v11i2.5472

Scherer, K. R., & Oshinsky, J. S. (1977). Cue utilization in emotion attribution from auditory stimuli. *Motivation and Emotion* 1, 331–346. https://doi.org/10.1007/BF00992539

Strykowski, D. R. (2016). Text Painting, or Coincidence? Treatment of Height-Related Imagery in the Madrigals of Luca Marenzio. *Empirical Musicology Review*, *11*(2), 109-119. https://doi.org/10.18061/emr.v11i2.4903

APPENDICES

Textual and electronic results files, high quality images of significance matrices, and a list of titles of all pieces in the corpus are available at: http://hdl.handle.net/1811/85873. Code used to produce these results is hosted on GitHub at https://github.com/cuthbertLab/emotionPainting.