Trevor de Clercq[*]

[*]*Middle Tennessee State University, Murfreesboro, TN, USA*
trevor.declercq@mtsu.edu

Pitfalls and Windfalls in Corpus Studies of Pop/Rock Music

ABSTRACT

In recent years, corpus studies have become a popular methodology for research on pop/rock music, afforded especially by recent developments in computing technology. Pop/rock music presents its own particular challenges for corpus work, though, since no official score typically exists. Thus in contrast to Western art music, any symbolic representation of pop/rock inherently requires an intermediary analytical stage, whether done by a human or a computer. Through a critical investigation of extant pop/rock corpora as well as reported findings from these corpora, this paper examines not only the benefits of this type of work but also many of its potential pitfalls. Two main aspects are examined: 1) the nature of data collection and representation, and 2) the modes of data analysis and interpretation. Although corpus studies ostensibly provide an objective measure of musical features, it is shown that a great deal of subjectivity exists within the creation of a pop/rock corpus and its analysis. In order to refine our corpus-based analytical methods, I argue that we must refine how we understand our own perception and intrinsic preconceptions. Corpus work on pop/rock music is important, ultimately, because it can shed light not only on the music under study but also on our own analytical presumptions and theoretical frameworks.

1. INTRODUCTION

This paper addresses corpus studies of pop/rock music—specifically, the benefits or 'windfalls' of this type of work as well as some of the problems. I call these problems 'pitfalls' because they are traps that we should do our best to recognize, either in our own work or in the work of others, and avoid whenever possible.

At a basic level, a corpus study is any methodological investigation of some body of work, although in its modern meaning, it implies some sort of statistical analysis of encoded music using a computer. Corpus studies thus purport to give a more objective view of theoretical insights that were previously gleaned primarily through intuition alone. For these reasons and others, corpus work has, perhaps not surprisingly, emerged as a promising subfield in music research, as shown by recent multiple-volume special journal issues (Temperley and VanHandel 2013; Shanahan 2016).

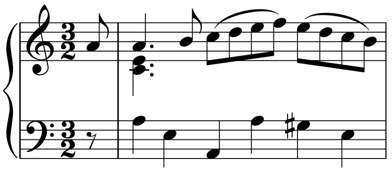
As this type of work has become more commonplace in music research overall, so too has the use of corpus methods to study pop/rock music. Existing corpora of pop/rock music include those created by a single author, such those by Summach (2012) and Tough (2013). Pop/rock corpora also include those with multiple authors, such as the corpus created by Burgoyne and his collaborators of top songs from the *Billboard* charts (Burgoyne, Wild, and Fujinaga 2011), as well as the corpus I created with Temperley of the top songs reported by *Rolling Stone* magazine (2013). Finally, some corpora of popular music have been created using computer algorithms, such as the 'Million Song Dataset' (Bertin-Mahieux et al. 2011).

One of my central points in this paper is that a corpus created by a single author or a computer algorithm has built-in shortcomings, and even a corpus created by multiple authors can fall short in similar ways. The primary problem derives from the subjective nature of music analysis, which can strongly shape the corpus and its results. The good news for music analysis, as I will argue, is that we can turn this pitfall into a windfall by using multiple independent annotators to assess subjectivity. In what follows, I discuss these issues via two main aspects of corpus work: 1) data collection and representation; and 2) data analysis and interpretation.

2. COLLECTION AND REPRESENTATION

In terms of data collection and representation, one could argue that choosing which pieces of music to include in a corpus is a type of musical analysis, since it involves judgments on the timespan or styles under study. Because Covach (2017) addresses this issue, I will not steal any of his thunder here, except to say that this is not an issue unique to pop/rock music.

What is more unique to a corpus study of pop/rock is that, unlike classical music, none of the musical content is explicitly given. In a classical work, such as the opening to the Courante by J. S. Bach shown in Example 1, the score shows discrete pitch and rhythmic information. We can thus faithfully encode the work, such as in the music21 format (Cuthbert and Ariza 2010) shown in Example 2.

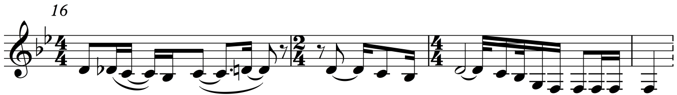


Ex. 1. First bar (with pickup) from J. S. Bach, English Suite no. 2 in A minor, BWV 807, Courante.

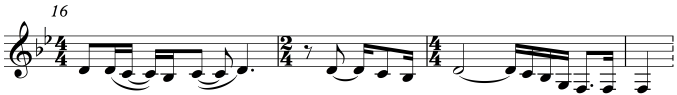
offset	Chord
0.3.4	<music21.chord.Chord A4>
1.1.1	<music21.chord.Chord C4 E4 A4 A3>
1.1.3	<music21.chord.Chord C4 E4 A4 E3>
1.1.4	<music21.chord.Chord B4 E3>
1.2.1	<music21.chord.Chord C5 A2>
1.2.2	<music21.chord.Chord D5 A2>
1.2.3	<music21.chord.Chord E5 A3>
1.2.4	<music21.chord.Chord F5 A3>
1.3.1	<music21.chord.Chord E5 G#3>
1.3.2	<music21.chord.Chord D5 G#3>
1.3.3	<music21.chord.Chord C5 E3>
1.3.4	<music21.chord.Chord B4 E3>

Ex. 2. Encoding in music21 format of Example 1 (J. S. Bach, English Suite no. 2 in A minor, BWV 807, Courante).

With popular music, however, no official score usually exists aside from the original audio recording. Thus in contrast to Western art music, any symbolic representation of pop/rock music inherently requires an intermediary analytical stage. For example, Example 3 shows my transcription of the melody prior to the first chorus (mm. 16–19, with the lyrics 'don't patronize, don't patronize me') from the song 'I Can't Make You Love Me', as recorded by Bonnie Raitt. My transcription is a good approximation of the melody, but it is an approximation nonetheless. For comparison, an alternate transcription of the melody to this same short passage, as published by Hal Leonard (2008), is shown in Example 4.



Example 3. Transcription of the vocal melody from Bonnie Raitt's 'I Can't Make You Love Me' (1991), mm. 16–19.



Example 4. Alternate transcription of the vocal melody from Bonnie Raitt's 'I Can't Make You Love Me' (1991), mm. 16–19.

The two versions of the melody shown above are very similar, but some differences can be found, such as the tuning of the second note in m. 16 and the timing of the last note in this same measure. There is also a difference in the timing of the pentatonic descent from the D to the low F in m. 18, prior to the beginning of the chorus. Admittedly, these are small differences, but they show the difficulty of separating performance from composition with pop/rock music. My version leans more towards trying to capture the subtle tuning and timing aspects of Bonnie Raitt's performance, whereas the alternate version leans more towards better reflecting the underlying composition of the melodic line.

The subjective aspect of transcription is even greater with harmony, since harmonic analysis involves reducing a texture of many notes to a single chord symbol. For example, the Hal Leonard version (2008) has a Cm7 chord (ii7 in this key) at the end of m. 16 through m. 17 (the bar of 2/4), but might the D-natural in the melody make it a Cm9 chord? The answer depends on the disposition of the analyst. Even if we agree on the notes in the chord, we may disagree on its label. For example, the Hal Leonard version (2008) has a Bb/F chord in m. 18, which leads to an F chord in m. 19. There are at least two ways to analyze the Bb/F chord using Roman numerals. One way would be to call it a tonic chord in second inversion, namely a I6/4 chord. The second way would be to call it a cadential 6/4 chord, namely a V chord with a suspended 6th and 4th above the bass. Both are standard practices in music theory, but each results in a different set of statistics. In the first reading, we find that a ii chord (in m. 17) moves to a I chord (in m. 18). In the second reading, we find that a ii chord moves instead to a V chord. Any statistics on root motions will thus simply reflect back to us our own analytical predispositions.

This reflection of our own analytical framework back to us in our statistics is an important trap to avoid or acknowledge, but how? As a first step, it is critical to have more than one encoder involved in a corpus study of pop/rock music, and moreover, for each of those encoders to do their analyses independently. Doing so, we have the chance to assess subjectivity. For example, in my corpus study with Temperley of harmony in 100 rock songs, we agreed on the key or pitch center about 97% of the time (2011, 59); with regard to the absolute root of a chord, such as whether it was an A or a D chord, we agreed about 94% of the time; and in terms of the function or Roman numeral, we agreed about 92% of the time. The fact that our agreement was consistently above 90% seems good, but it is not 100%. Because there is currently no other corpus of pop/rock music that has multiple independent analyses by different annotators, we do not know yet whether these figures are typical or atypical.

To be clear, I admit that some aspects of a pop/rock song can be objectively measured. For example, Schellenberg and von Scheve (2012) conducted a corpus study of Top 40 songs from the *Billboard* charts spanning 1965 to 2009. They found that pop songs from the latter half of the 1980s were the longest, while songs from the late 1960s were the shortest.[1] Indeed, song length seems like an entirely objective parameter (although disagreement could theoretically occur for songs with fade-ins or fade-outs).

We must be careful, though, not to overestimate how objectively any parameter can be measured. For instance, Schellenberg and von Scheve also report that average tempo decreased during this same period (2012, 200). To the casual observer, this finding may seem unproblematic. After all, deciding whether a song is fast or slow may seem to many people like a straightforward process. But recent perceptual studies of tempo assessment in popular music show a great deal of variation between listeners (Moelants and McKinney 2004; Levy 2011). To illustrate this, consider two versions of the song 'Teardrops on My Guitar' by Taylor Swift (2008). The first is the remix release called the 'pop' version. Most listeners would presumably hear the tempo of the 'pop' version as 100 BPM, with the kick and snare drum corresponding to each beat in 4/4. Now consider the original version, which is the first track on Taylor Swift's self-titled debut. (I leave it to the reader to access a recording of this song.) Note that the harmony and melody in the original version are moving at the same rate as in the 'pop' version, but the drums are going half as fast (i.e., the kick and snare drum are occurring half as often). If a listener attends to the pacing of the harmonic and melodic content, they may feel the beat around 100 BPM. But if a listener attends to the pacing of the drums, they may feel the beat around 50 BPM. Our notion that any given song has a single tempo, therefore, may itself be problematic.[2]

To be fair, Schellenberg and von Scheve admit that measuring the tempo of a song was 'complicated' (2012, 198). Their solution was to have two musicians each independently rate the tempo of every song, and if the tempo ratings did not agree, a

[1] This finding is reported in Table 2 under "Mean duration" (Schellenberg and von Scheve 2012, 200).

[2] For more on how the disbursement rate of harmonic and melodic material in pop/rock music may differ from the tempo implied by the drums, see de Clercq (2016).

third musician would resolve the disagreement. Any ambiguity in creating the corpus was thus completely removed from the final version. This is not the only corpus to use such a method. The creators of the McGill *Billboard* corpus also employed two musicians to independently analyze the chords for each song, and then brought in a third 'meta-annotator' to compare the two versions and decide what would be the final transcription (Burgoyne, Wild, and Fujinaga 2011). Here again, all information with regard to ambiguity was removed. The final corpus is thus presented as if its contents are entirely objective—that there is a single 'best answer'. Instead, I believe we should capture and investigate the extent to which and the situations in which analysts disagree. Untangling that, I would argue, is as important if not more important than any statistical results we obtain, especially given the nascent state of corpus research on pop/rock.

Embracing ambiguity is important because many researchers believe that a central benefit of corpus work on pop/rock music is to the lay the foundation for automated computer analysis of music. A human-annotated corpus is meant to act as the 'ground truth' from which a computer will learn so as to be able to automatically analyze new music. This approach has practical uses for the music industry. For example, Spotify currently has a feature called 'Sort Your Music' available through its web interface. With this feature, users can sort their music on various parameters, such as release date, loudness, and BPM. The BPM rating that Spotify identifies may not correspond to most listeners' ratings, however. For example, Spotify rates 'Out of Mind' by Colbie Caillat as having a tempo of 180 BPM. Comparing this tempo to the recorded version, it seems that the algorithm does a good job synchronizing with the music, but there is an 'octave error', in that—at least to my ears—the more obvious tempo rating for the song is around 90 BPM, i.e., half as fast as the algorithmically-derived tempo value. In other words, the software is having trouble determining whether a song is fast or slow, which is a very basic aspect of how we traditionally think about tempo.

So work remains to be done, but how much? Research by Levy (2011) offers some insight. As indicated on the 'Sort Your Music' web page,[3] the software driving Spotify's music analysis algorithm is the EchoNest API. In his 2011 article, Levy compares BPM ratings from the EchoNest API to crowd-sourced ratings of the same songs. Levy finds that the EchoNest API generates the same BPM value as the crowd-sourced value only about 40% of the time.[4] In fact, about a quarter of the time, the EchoNest value is completely unrelated to the crowd-sourced value.[5]

As it stands today, it seems that we should be wary of any corpus generated by a computer algorithm. For example, the 'Million Song Dataset', which I mentioned previously, promises to be a great resource. Unfortunately, it was created using the EchoNest API (Bertin-Mahieux et al. 2011). So even though there are a million BPM ratings in the corpus, only about 40% probably correspond to a human listener's rating. The computer algorithm, in other words, reflects back not the perception of a human listener but rather the mechanics of its programming framework, much in the same way as the encodings of a human reflect back their own analytical framework.

3. ANALYSIS AND INTERPRETATION

I will now move to the section concerning data analysis and interpretation. Because time and space is limited, I will avoid discussing the standard statistical fallacies we might find with any corpus study, such as confusing statistical significance with practical significance or correlation versus causation, since these types of statistical errors are well documented elsewhere (DeGroot and Schervish 2012; Huron 2013). Instead, I address some pitfalls more specific to studies of pop/rock music, many of which derive from the underlying subjectivity in the encoding process.

For example, Summach conducted a corpus study of the *Billboard* charts from 1965 to 1989 for his doctoral dissertation (2012). As reported in his *Music Theory Online* article, which draws from this work, he found that verse-chorus songs—both those without a prechorus and those with a prechorus—have on average been getting longer from 1965 to 1989 (2011, Ex. 27). With that in mind, consider that he also found that the distribution of verse-chorus songs either with or without prechorus sections changed over this same period, such that most verse-chorus songs in the early 60s and 70s did not have prechorus sections whereas most verse-chorus songs in the late 1980s did (2011, Ex. 26). Remember, though, that Summach found that the average length of a song increased during this same period. If, therefore, an analyst has a threshold for how long a passage must be in order to be classified as a prechorus, we should not be surprised to find prechorus sections to be more common in longer songs. And in fact, Summach states directly in this same article that a 4-bar passage is not long enough, in his opinion, to be a prechorus (2011, [22]). His finding that prechorus sections are more common in the late 1980s than in earlier years is thus arguably predicated on Summach's personal interpretation of what constitutes a prechorus.

Ultimately, the analysis of form is probably the most idiosyncratic element in music theory.[6] We should be especially cautious, therefore, when we find statistics on form that derive from the analysis of only a single listener. Yet these sorts of statistics are fairly common in published articles, such as in the corpus study by Tough on recent *Billboard* songs (2013). As Tough reports, 68% of the songs in his corpus have intro sections that last 10 seconds or less.[7] For the sake of argument, let us assume that these values reflect the hearing of all listeners. What are we to make of this finding? Tough posits that if you want a song to become commercially successful in today's market, the song should have a short intro section, because that is how most modern hits are structured (2013, 111). This line of reasoning exemplifies a field known as 'Hit Song Science' (Ni et al. 2011). The presumption is that if a song is like other songs

[3] The URL for Spotify's 'Sort Your Music' feature is <http://sortyourmusic.playlistmachinery.com> (accessed August 22, 2017).

[4] This finding is reported in Table 2 (which is located after Table 4) in the 'correct' column (Levy 2011, 322).

[5] This finding is also reported in Table 2, in the 'unrelated' column (Levy 2011, 322).

[6] For some attempts to measure the extent of subjectivity in the analysis of form, see Bruderer, McKinney, and Kohlrausch (2006) and Smith (2014).

[7] This value can be gleaned from combining the red and blue portions of the pie chart shown in Figure 1 (Tough 2013, 105).

that are hits, then the song has a higher likelihood of becoming a hit itself. Data scientists have disproven the viability of this hypothesis (Pachet and Roy 2008), at least given the current state of research, but it is a philosophy that still underpins much work on popular music (as Tough's 2013 article exemplifies). The nagging question is whether a song achieves success because of its typicality or, conversely, because of its atypicality. Certainly, Nirvana's 'Smells Like Teen Spirit' (1991) did not sound anything like the *Billboard* hits that preceded it, for example.

As we near the conclusion, I want to reaffirm some of the windfalls of corpus work on pop/rock music, since I have highlighted many pitfalls. Traditionally speaking, corpus studies have been considered a subfield of or adjunct to music cognition. A corpus study is seen as a way to understand human perception—a way to understand our built-in subjectivity and how that subjectivity affects our analyses. For example, Example 5 shows data I presented at the last EuroMAC conference (2014). This table shows the average duration of chords, in bars, for verse and chorus sections in the *Rolling Stone* magazine corpus I created with Temperley (i.e., the RS 200, as reported in our 2013 article). As you can see, both Temperley (DT) and I (TdC) have average chord lengths in our verse sections that are longer than those in our chorus sections.

Chords	Analyst	Verse	Chorus	Effect
Overall	DT	2.08	1.55	$t(134) = -2.55, p = .01$
	TdC	2.26	1.70	$t(107) = -2.06, p = .04$
Tonic	DT	2.25	1.57	$t(134) = -3.11, p = .002$
	TdC	2.48	1.75	$t(107) = -2.37, p = .01$
Non-Tonic	DT	1.10	0.99	ns
	TdC	1.10	1.09	ns

Example 5. Average chord durations, in bars, for songs in the RS 200 corpus (de Clercq 2014).

Based on this finding, we might hypothesize that a listener's perception of a section's role is affected by the durations of its chords. This finding appears to be relatively intersubjective, because it is reflected in both my analyses as well as Temperley's. We must be careful, though, with averages calculated across an entire body of songs because it may not represent any broad stylistic trait but rather the generic midpoint of multiple smaller populations. In fact, the reason that chord durations appear shorter in chorus sections than verse sections seems entirely predicated on the length of the tonic chord alone. There does not seem to be any significant difference in the length of non-tonic chords when comparing verse and chorus sections. Based on this data, I speculate that one factor that makes a passage sound more like a verse than a chorus derives from the length of the tonic harmony. My study of our corpus, therefore, has been a tool for me to interrogate my own perception, and in doing so, to compare it to the perception of others.

4. CONCLUSION

In summary, corpus work offers a number of windfalls for music research. Key-finding algorithms, for example, can be trained on symbolic data (Temperley and de Clercq 2013). Additionally, if the symbolic data is time-aligned with original recordings, it can provide a 'ground truth' for computerized tempo estimation, chord extraction, and melodic transcriptions of raw audio files. We must be sure to recognize and reflect, however, the great variability in interpretation that exists between human analysts in these tasks, else we become trapped in pitfalls of our own making. In order to refine our corpus-based analytical methods, we must embrace this ambiguity, which mirrors our own perception and intrinsic preconceptions. Corpus work on pop/rock music can then and only then shed light not just on the music under study, but also perhaps more importantly on the analytical, theoretical, and perceptual frameworks that we use to encode and understand this music.

KEYWORDS

Popular music, musical models, recorded music, epistemology.

REFERENCES

Bertin-Mahieux, Thierry, Ellis, Daniel, Whitman, Brian, and Lamere, Paul, 2011. 'The Million Song Dataset', in Anssi Klapuri and Colby Leider (eds.), *Proceedings of the 12th International Society for Music Information Retrieval Conference*, 591–96.

Bruderer, Michael, McKinney, Martin, and Kohlrausch, Armin, 2006. 'Structural Boundary Perception in Popular Music', in Kjell Lemström, Adam Tindale, and Roger Dannenberg (eds.), *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR)*, 198-201.

Burgoyne, John, Wild, Jon, and Fujinaga, Ichiro, 2011. 'An Expert Ground Truth Set for Audio Chord Recognition and Music Analysis", in Anssi Klapuri and Colby Leider (eds.), *Proceedings of the 12th International Society for Music Information Retrieval Conference*, 633–38.

Covach, John, 2017. 'Analyzing Form in Popular Music: History, Style, Aesthetics, and Data', paper presented at the *9th European Music Analysis Conference* (Salzburg, France).

Cuthbert, Michael and Ariza, Christopher, 2010. 'Music21: A Toolkit for Computer-Aided Musicology and Symbolic Music Data", in J. Stephen Downie and Remco C. Veltkamp, *Proceedings of the 11th International Society for Music Information Retrieval Conference*, 637–42.

de Clercq, Trevor, 2014. 'Typical Chords in Typical Song Sections: How Harmony and Form Interact in a Corpus of Rock Music', paper presented at the *8th European Music Analysis Conference* (Leuven, Belgium).

———, 2016. 'Measuring a Measure: Absolute Time as a Factor for Determining Bar Lengths and Meter in Pop/Rock Music', *Music Theory Online* 22/3.

de Clercq, Trevor, and Temperley, David, 2011. 'A Corpus Analysis of Rock Harmony', *Popular Music* 30/1: 47–70.

DeGroot, Morris and Schervish, Mark, 2012. *Probability and Statistics*, 4th ed. Boston, M.A.: Addison-Wesley.

Hal Leonard, 2008. *100 Greatest Love Songs*. Milwaukee, W.I.: Hal Leonard.

Huron, David, 2013. 'On the Virtuous and the Vexatious in an Age of Big Data', *Music Perception* 13/1: 4–9.

Levy, Mark, 2011. 'Improving Perceptual Tempo Estimation with Crowd-Sourced Annotations', in Anssi Klapuri and Colby Leider (eds.), *Proceedings of the 12th International Society for Music Information Retrieval Conference*, 317–22.

Moelants, Dirk, and McKinney, Martin, 2004. 'Tempo Perception and Musical Content: What Makes a Piece Fast, Slow, or Temporally Ambiguous?', in Scott Lipscomb, Richard Ashley, Robert Gjerdingen, and Peter Webster (eds.), *Proceedings of the 8th International Conference on Music Perception and Cognition*, 558–62.

Ni, Yizhao, Santos-Rodríguez, Raúl, Mcvicar, Matt, and De Bie, Tijl, 2011. 'Hit Song Science Once Again a Science?', paper presented at the *4th International Workshop on Machine Learning and Music* (Spain).

Pachet, François, and Roy, Pierre Roy, 2008. 'Hit Song Science Is Not Yet a Science', in Juan Pablo Bello, Elaine Chow, and Douglas Turnbull (eds.), *Proceedings of the 9th International Society for Music Information Retrieval Conference*, 355–60.

Schellenberg, E. Glenn, and von Scheve, Christian, 2012. 'Emotional Cues in American Popular Music: Five Decades of the Top 40', *Psychology of Aesthetics, Creativity, and the Arts* 6/3: 196-203.

Shanahan, Daniel, 2016. 'Editor's Note [to the Special Issue on Corpus Studies]', *Empirical Musicology Review* 11/1: 1.

Smith, Jordan, 2014. 'Explaining Listener Differences in the Perception of Musical Structure', PhD diss., University of London.

Summach, Jay, 2012. 'Form in Top-20 Rock Music, 1955–89', PhD diss., Yale University.

———, 2011. 'The Structure, Function, and Genesis of the Prechorus', *Music Theory Online* 17/3.

Temperley, David, and de Clercq, Trevor, 2013. 'Statistical Analysis of Harmony and Melody in Rock Music', *Journal of New Music Research* 42: 187–204.

Temperley, David, and VanHandel, Leigh, 2013. 'Introduction to the Special Issues on Corpus Methods', *Music Perception* 31/1: 1–3.

Tough, David, 2013. 'Teaching Modern Production and Songwriting Techniques: What Makes a Hit Song?', *Journal of the Music and Entertainment Industry Educators Association* 13/1: 97–124.

Winkler, Peter, 1997. 'Writing Ghost Notes: The Poetics and Politics of Transcription', in David Schwarz, Anahid Kassabian, and Lawrence Siegel (eds.), *Keeping Score: Music, Disciplinarity, Culture*. Charlottesville, VA: University Press of Virginia, 169–203.