Cross Entropy as a Measure of Coherence and Uniqueness

Christopher Wm. White(✉)

The University of Massachusetts Amherst, Amherst, USA
cwmwhite@umass.edu

Abstract. Cross entropy, a measurement of the complexity/predictability of a series of observations given a probabilistic model, has been used in a variety of domains in music scholarship for decades. This paper presents a novel application of this metric to musical corpus analysis. Given a series of divisions to a larger corpus, a sub-corpus is relatively "unique" if a probabilistic model derived from its pieces better predicts its constituent pieces than do models derived from other sub-corpora. A sub-corpus is relatively "coherent" if its own model describes its pieces better than a model derived from the entire corpus. The Yale-Classical-Archives corpus was used to illustrate several strategies for sub-corpus division, each of which are tested for uniqueness and coherence. Some broader interpretive applications are also described.

Keywords: Computation · Corpus analysis · Cognitive modeling · Style

1 Introduction

Music researchers have been experimenting with concept of *entropy* almost since the field of informatics began in the mid 20th century [1, 2]. Since the 1950s, scholars have connected entropy, or the relative complexity of some signal, to musical style, communication, normativity, meaning, and compositional modeling [3–7]. As shown in Eq. 1, the entropy H of an observed series O measures the complexity of a signal by calculating the log-probability of an event o to occur within some observed series O (here, $\log P(o_i)$), weighting that value by the relative frequency with which that observation occurs in O (here, $P(o_i)$, and summing all such values. The negative sign turns the negative value resulting from the logarithm into a positive value, such that the higher the value, the more randomness – or more entropy – as series has. A very redundant signal – one in which a particular event happens most of the time – will have low entropy since those events are highly predictable given the rest of the signal, while a series of wildly unpredictable events would have a high entropy. (Here, the logarithm's base can be chosen as appropriate for the situation: this study uses base 2 in order to report entropy in bits.)

$$H(O) = -\sum\nolimits_{i=1}^{n} P(o_i)\log P(o_i) \tag{1}$$

O. A. Agustín-Aquino et al. (Eds.): MCM 2017, LNAI 10527, pp. 324–334, 2017.
https://doi.org/10.1007/978-3-319-71827-9_25

In the past several years, work by David Temperley [8] has introduced a particular modification of this technology to music: *cross entropy.* In this framing of the general concept, the probability of each event is judged by some other model rather than by some probabilistic distribution drawn from the observation itself. As shown in Eq. 2, this formula takes the negative log-probability of an event o given some probabilistic model m, again weighting each value by its probability mass within the observation series O, and then summing for all n events. This essentially captures how well some probability distribution m accounts for the series of observations O.

$$H(O,m) = -\sum\nolimits_{i=1}^{n} P(o_i)\log m(o_i) \quad (2)$$

This paper proposes several novel ways of applying this modeling technique to the analysis of musical data. I will show how cross entropy can capture the *coherence* and the *uniqueness* of musical corpora. Because, in one sense, using a composer's identity to build a corpus creates an unassailably coherent and unique dataset: using this framework, the composer's identity provides the desideratum as to whether a piece is included in some corpus. But, one might also wonder whether a composer writes pieces that are distinct from their contemporary colleagues, or if a composer's style is basically interchangeable with that of their contemporaries. If the former were true, the composer's pieces would exhibit notably divergent statistical properties from those of their colleagues; but, if two composers have made virtually identical decisions surrounding some musical parameter, then the same statistical model could represent both corpora.

This paper agues that cross entropy can shed light onto these sorts of questions by manipulating which models are used to assess the corpus. Given some corpus with potential smaller divisions (or, *sub-corpora*), if the individual pieces within some sub-corpus are predicted by the overall corpus better than any other sub-corpus, that sub-corpus is *unique* as compared to other sub-corpora. If that sub-corpus contains pieces that are more statistically similar to one another than to the overall corpus, that sub-corpus is *coherent*. Below, I show a computational model that exploits these properties to test the coherences and uniqueness of several different divisions of a large corpus of Western-European common-practice MIDI files. I end by discussing the interpretive potentials that this modeling provides.

2 The Corpus, the Sub-corpora

This experiment relied on data from the Yale-Classical-Archives Corpus [9]. This corpus collects MIDI files from classicalarchives.com (a website of user-sourced MIDI files), each associated with metadata that specifies the file's opening key, meter, composer, date of composition, instrumentation, composer's nationality, genre, and so on. Given that this study was interested in dividing corpora by composer, the 19 composers listed as "The Greats" on the website were used: Bach, Beethoven, Brahms, Byrd, Chopin, Debussy, Handel, Haydn, Liszt, Mendelssohn, Mozart, Saint-Saens, Scarlatti, Schubert, Schumann, Tchaikovsky, Telemann, Vivaldi, and Wagner. The overall corpus was more than 5,000 pieces, and the average composer's dataset contained 231 pieces, with the smallest corpus – Wagner's – containing only 33, and the largest – Scarlatti's – containing 554. The corpus is divided into "salami slices" – every verticality where the pitch-class content changes. The average composer's sub-corpus had 339,185 such slices, with Wagner's again being by far the smallest (67,538), and Mozart's being the largest (1,322,716). The corpus also contains tonal annotations, which were used to convert the corpus's pitch material of each slice into scale degrees.

These scale-degree sets were used to create Markov (n-gram) chains designed to probabilistically model how surface harmonies progressed to one another. Different sizes of n-grams within these tonal passages were then tallied, and after initial experimentation it was determined that trigrams (i.e., $n = 2$) seemed to balance between precise and sparse data. (An n-gram model involves contiguous sequences of n items from a sequence of observations. When $n = 2$, the observation at the current timepoint is conditioned on the two previous observations. The model is therefore concerned with three-chord trigrams – the current and previous two chords – at every observed timepoint.) In order to remain as theory-neutral as possible, the meter metadata was used to gather trigrams at three metric levels; these three levels were then combined. Repeating data collection at several levels and agglomerating the resulting trigrams allows for patterns that recur at several durational or metric levels to become more dominant in a distribution while remaining agnostic as to the relative importance of different surface divisions. The three metric levels were (1) the salami slices themselves, (2) the contents of each beat as defined by the corpus's metric data (i.e., the quarter-note in 4/4), and (3) the contents of the beat's primary division (i.e., the eighth note in 4/4; this division is also recorded by the corpus). NB: this process recognizes not only traditional chords (like triads and seventh chords) but also less traditional chords (like passing chords and dissonances): this study therefore assumes that any surface structure is a legitimate "chord," following [10–12]. The tallying and organization of the YCAC's trigrams was implemented with Python version 2.6 using the music21 software package [13].

In order to compare the uniqueness and coherence of various different divisions of the larger dataset, several different divisions of the larger corpus were undertaken. Most basically, each individual composer's output will first be considered a sub-corpus. Next, chronological divisions were used, grouping pieces in the corpus by their date of publication, first arranged by the half-century beginning in 1650 and ending in 1900, and then by 30-year epochs (now beginning in 1680 because of the sparse data between 1650 and 1679). Finally, to introduce machine-learned groupings into the corpus, the groupings found in [14] were used. Here, the identical dataset and modeling as described above were used, and composers' trigram frequencies were submitted to a k-means cluster analysis to group composers whose surface harmonic progressions were statistically similar. The study used values $k = [0\ldots10]$; peaks in silhouette widths were used to identify optimal k values; and, such peaks values were identified for 7 and 10 clusters. The groupings – used here as sub-corpora are reproduced in Table 1.

Table 1. *k*-means clusters drawn from White (2014)

K-*means clusters*	
k = 7	k = 10
Bach	Bach
Byrd	Byrd
Beethoven, Mozart, Haydn, Schumann, Mendelssohn, Brahms, Schubert, Wagner	Beethoven, Mozart, Haydn, Mendelssohn, Schubert
Tchaikovsky, Liszt, Chopin, Saint-Saens	Tchaikovsky, Liszt, Chopin, Saint-Saens
Telemann, Vivaldi, Handel	Telemann, Vivaldi
Debussy	Debussy
Scarlatti	Scarlatti
	Wagner
	Brahms, Schumann
	Handel

3 Modeling Coherence and Uniqueness

The *coherence* and *uniqueness* of each division was corpus quantified by determining the cross entropy of each piece given every other piece (exclusive of the piece under question) in some sub-corpus. In terms of Eq. 2, for each piece, the observations O would be those trigrams within an individual piece within the sub-corpus, and the model m would be the probability distribution of all trigrams within the remaining pieces in that corpus. As a baseline, each piece within the sub-corpus was judged in relation to the entire corpus (here, the entire YCAC becomes the model m). The average and standard error of these cross entropies across the sub-corpora is tallied, as well as the pieces in each sub-corpus as judged by the entire corpus. A sub-corpus is *unique* if its standard error is sufficiently low to not overlap with the window of any other corpus's standard error ("does this sub-corpus predict its own pieces better than any other sub-corpora above chance?"). A sub-corpus is *coherent* if the standard error of its self-assessments is outside the standard error of the overall corpus ("does this sub-corpus predict itself better than it would be predicted by the entire corpus?").

NB: As cross entropy is itself a relative measurement, so too are uniqueness and coherence. Each of these numbers must only be judged in relation to other numbers: a piece is only unique *in relation to other sub-corpora* or only more coherent *than the overall corpus*.

Importantly, both these ideas have conceptual overlaps with the central idea of "entropy." When applied to a single dataset (i.e., using the format of Eq. 1), entropy rises when each event is more random in terms of the other events, and falls when each event it is more predictable. Uniqueness and Coherence manipulate these relationships by comparing a dataset's randomness not simply to the dataset itself, but to other potential datasets with which the original dataset has some relationship. In other words, these ideas capitalize on the original informatic structure of entropy to draw out additional relationships between datasets. Note also that the difference between uniqueness and coherence is not mathematical in nature (indeed, they are mathematically identical), but in the relationships between the models used, with uniqueness quantifying relationships between sub-corpora and coherence quantifying relationships between a corpus and its sub-corpora.

4 Sub-corpora

4.1 Dividing by Composer

By dividing the corpora by composer, 74% composer-by-composer comparisons were significantly unique. The median proportion of unique comparisons was 83%. 88% of these sub-corpora predicted themselves better than the overall corpus. Only two composers – Byrd and Handel – registered perfect results: the trials produced significantly lower cross entropy when comparing these composers' own pieces to their own corpora than when comparing them to any other composer's corpus. In other words, these results show the models to be "sure" these composers' pieces were significantly more likely on average to be composed by themselves than by someone else. Example 1a shows Handel's sub-corpus compared to that of each other composer. The cross entropy of the composer's pieces when compared to the other composers' sub-corpora are shown as the clear bar, other composers are shown by solid bars, the self-wise comparison is shown by the white bar, and the dashed bar shows Handel's average cross entropy judged by the entire corpus. Handel's own pieces are judged statistically significantly better than they are judged by other corpora– the corpus is therefore unique. The corpus also judges itself better than it is judged by the overall corpus– it is therefore coherent (Fig. 1).

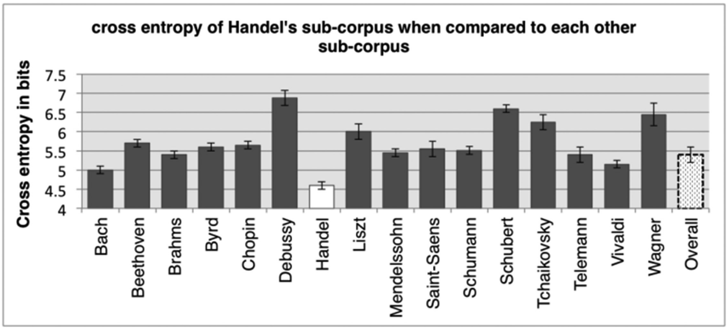

Fig. 1. Comparative cross entropies using Handel's sub-corpus model

However, more than a quarter of the time, these trials judged other corpora to predict a composer's pieces with either a lower or insignificantly different level of cross entropy when compared to the composer's own corpus. Mendelssohn's corpus, for instance, performed around the average with two non-unique comparisons: the cross entropies of the Brahms, Handel, and Schubert sub-corpora were not significantly different from the cross entropy resulting from a self comparison. However, the corpus does predict itself significantly better than the agglomerated corpus predicts its pieces. This result indicates that the Mendelssohn model is coherent insomuch as it predicts its own pieces well; however, it also shows that the model is not sufficiently unique, as other models predict Mendelssohn's corpus virtually identically to Mendelssohn's own model.

On the whole, it seems that these results suggest that grouping corpora by composer tends to create coherent corpus models, although these models are often not sufficiently unique from one another.

4.2 Dividing by Chronological Epochs

Here, pieces were divided into sub-corpora based on their date of composition, first into fifty-year epochs, and then in thirty-year epochs. The fifty-year model performed worse than that using composer-defined trials, the former returning a 68% uniqueness rate. However, the median success rate was higher, registering an 80% uniqueness rate. This rate stems from the fact that one time period, 1751–1800, did not have a single successful trial; this epoch also did not predict itself better than did the overall corpus. Example 2a shows the offending epoch's results. Not only can the late 18th-century corpus not be significantly distinguished from the late 17th-century corpus, but the other three corpora produce significantly *lower* cross entropies, indicating that these corpora predict the trigrams within the late 18th-century corpus better than they predict the trigrams of their own time periods. The fact that the overall corpus predicted its pieces better than did this sub-corpus also indicates this sub-corpus to not be coherent.

Example 2b shows the case of the 1801–1850 corpus, a relatively successful example representing this test's median. While a self-comparison yields the lowest average cross entropy, the average cross entropy when compared with the 1851–1900 corpus is not significantly different than the self-wise average. (Interestingly, 75% of the unsuccessful judgments throughout the 50-year-epoch test (i.e., incorrect/insignificant comparisons) involved time periods adjacent to one another; if one removes the late 18th-century results from the percentage, this number rises to a complete 100%. In other words, with the exception of the problematic late 18th-century, the models generally become "confused" as to a piece's time period only when comparing that piece to a chronologically adjacent corpus.)

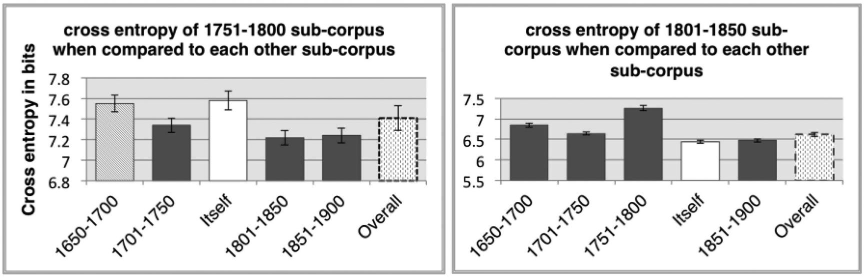

Fig. 2. Comparative cross entropies of (a) the 1751–1800 sub-corpus, and (b) the 1801–1850 sub-corpus

Dividing the corpora into 30-year segments produced similar results. The overall average success rate was 68.75%, and the median success rate was 75%. Half of the unsuccessful returns involved adjacent time periods, and 80% were within two time periods (i.e., within 60 years). As with the 50-year segments, the remaining 20% were not evenly distributed throughout the results, but centered in two particularly unsuccessful epochs. Two trials were not relatively coherent: the 1801–1830 and 1891–1920 trial. Example 8 shows a median example, the 1741–1770 corpus, while Example 9 shows the largely unsuccessful 1801–1830 results. (While it may be satisfying that the only significant positive results involve corpora that are maximally chronologically distant from the 1801–1830 corpus, note that the two adjacent corpora register a significant but lower cross entropy.)

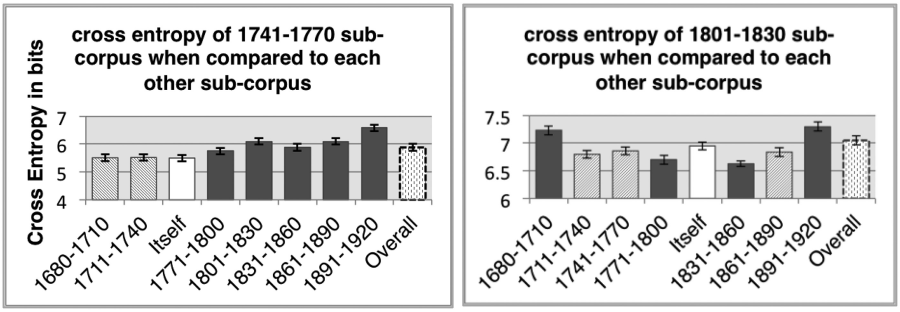

Fig. 3. Comparative cross entropies of (a) the 1741–1770 sub-corpus, and (b) the 1801–1830 sub-corpus

These results suggest that dividing a corpus by chronological epochs may be successful in some respects – it creates a high median success rate – but it also generates several corpora that are incoherent. From a modeling perspective, this incoherence could be explained by the presence of multiple and distinct chord-progression practices within a single corpus. For instance, the 1801–1830 corpus seems to have properties that are better modeled by the corpora surrounding it, perhaps indicating that this era contains practices that overlap those of its two surrounding epochs. If dividing corpora by composers seemed to create too many divisions, dividing by chronology is too broad, creating incoherent corpus models. Also, these tests seem to indicate a connection between chronological proximity and models' similarities.

4.3 Machine-Learned Sub-corpora

Using the *k*-means clusters produced markedly better results, although somewhat unsurprising as it used the same metric – chord-progression probabilities – both to divide the corpora and to judge the success of those divisions. (However, the results of this test do confirm the power of harmonic transition probabilities to classify groups of composers into unique and coherent corpora.) The seven clusters provide nearly perfect results, with only Debussy's corpus providing insignificant/non-unique comparisons, likely due to its small membership (n = 60). Figure 4a shows a typical perfect 7-cluster trial, using the "Romantic" (Tchaikovsky, Chopin, Liszt, Saint-Saens) sub-corpus. The ten clusters performed slightly worse, with an 88% success rate. If, however, one discounts the insignificant results of the two smallest corpora – now adding Wagner's corpus (n = 32) to Debussy's – the results rise to a 97.22% success rate. Figure 4b shows one of the two remaining insignificant results, the other being the average cross entropy of Vivaldi/Telemann's pieces given Handel's corpus.

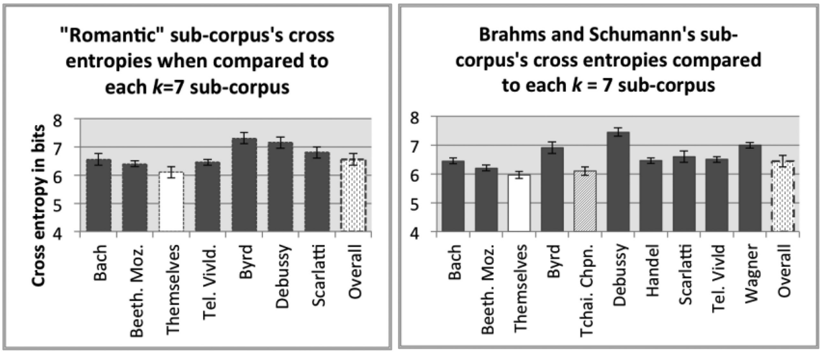

Fig. 4. Comparative cross entropies of (a) the Tchaikovsky, Chopin, Liszt, and Saint-Saens sub-corpus, and (b) the Brahms and Schumann sub-corpus

5 Applications

This type of modeling has various applications in how we think about and interpret musical corpora and the works contained therein. In what follows, I outline four potential applications of this kind of modeling, showing ways it can be used to interpret stylistic trends and compositional schools, how it can be used to identify points of innovation, how it can be used to broach the (admittedly thorny) topic of authorship, and how it might be used to formalize models of historical styles.

5.1 Describing Stylistic Trends and Compositional Schools

When the model identifies several composers whose sub-corpora and not unique, but – when grouped together – create a unique and coherent sub-corpora, this potentially identifies a compositional cohort operating within a similar compositional school. Here, we imagine that the compositional trends and norms used by these composers are sufficiently similar that the variation within their outputs makes them statistically indistinguishable (at least within the tested parameters). Furthermore, non-unique comparisons can show other potential avenues of influence. For instance, in Fig. 4, Brahms and Schumann's sub-corpus is coherent and unique in all but the comparison to the Tchaikovsky–Liszt–Saint-Saens–Chopin sub-corpus. This suggests that the output of these two composers comprise a distinct style that influences the output of these later Romantic composers. The fact that chronological adjacencies within the epoch-based models frequently accounted for non-unique comparisons also suggests stylistic trends. Here, this non-uniqueness captures the chronological developments of historical styles: historically proximate sub-corpora share statistical tendencies.

5.2 Moments of Innovations

Non-unique and incoherent findings also provide an opportunity for interpretation. At these junctures, the lack of similarity within the pieces constituting the sub-corpus begs for some kind of explanation: why would pieces written within such chronological proximity be so different?

Consider the case of the 1751–1800 sub-corpus: its constituent pieces are better predicted by the statistics of neighboring historical epochs than the pieces in its own time period. Looking inside that dataset, one finds groups of composers who would seem to be drawn from divergent compositional practices. It is not only a time period that saw the late works of Telemann and Scarlatti, but also the complete works of Mozart and Haydn, and ended with the mature works of Beethoven. One could similarly describe the 1801–1830 corpus: such a division groups middle-period Beethoven not only with Schubert, but with Schumann's early works. The incoherence of these sub-corpora, then, supports the idea that these moment host more of a stylistic shift than their surrounding eras. As indicated in Figs. 2a and 3b, it seems that significant portions of these groupings are better predicted by the surrounding epochs than contemporary compositions, further suggesting that these eras feature dramatic shifts between the previous and following styles.

5.3 Authorship

This modeling technique also allows for potential evaluations of reproductions, completions, or potentially spurious compositions. In each of these instances, one could take the piece(s) in question and treat them like their own sub-corpus, comparing its coherence and uniqueness with other sub-corpora in the piece(s) historical orbit. For instance, Fig. 5 compares a famous example of forgery to the 10-cluster sub-corpora. The forgeries here are those of Nicolas Chedeville publishing Vivaldi's fictitious "Opus 13" in 1737. The "X" above each of the bars shows the corpus' self-wise cross entropy, each constituent bar shows the forgeries' cross entropy compared to other sub-corpora, and the final bar again shows the agglomerated corpus's assessment of the forgeries. The sub-corpus is coherent, but it is not unique. Many other sub-corpora fall within the standard deviation of the average assessment: as before, these are shown as lined bars. Bars outside of the average standard deviation are shown as solid. There are two below the average: the Handel and Telemann-Vivaldi clusters. This means that these reproductions do indeed adequately imitate Vivaldi, but do so in a way that they could potentially also be passed off as composed by Handel!

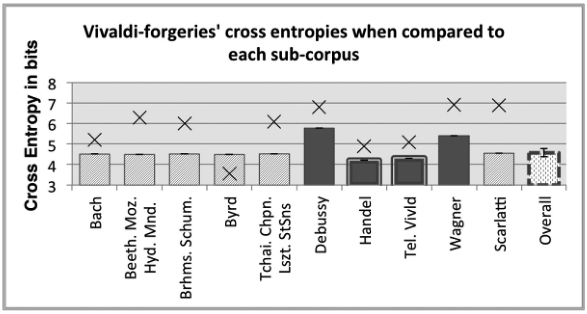

Fig. 5. Comparative cross entropies of Nicolas Chédeville's Vivaldi forgeries compared to the 10-cluster sub-corpora

5.4 Generative Modeling

Using these metrics to identify relatively unique and coherent statistical systems can potentially create well-formed generative models of some style. For instance, if one used the chord-progression (i.e., Markov-chain) probabilities embedded in the, say, Brahms-Schumann sub-corpus to generate sucessions of harmonies, one could reasonably argue that this models aspects of that style's compositional norms. The same cannot be said of, say, a model drawn from the 1751–1800 sub-corpus: because of its incoherence, it is not clear what such a generative model would capture outside of manifesting the era's stylistic heterogeneity. These metrics, then, can be imagined as ways to isolate statistical systems that can express some historically, culturally, or compositionally independent style.

6 Future Work

Of course, this work is incomplete. It relies entirely on simple Markov chains drawn from the very surface of a musical corpus. It is possible, for instance, that judging the similarity of two systems using something like a Context Free Grammar or at least some hierarchical system would better represent similarities and differences in chord progression usage. Additionally, other surface events rather than chord progressions may capture salient differences between sub-corpora: melodic figuration, recurrent bass lines, orchestration, or ornamentation may all contribute to stylistic differences better than (or in addition to) surface chord progressions. However, regardless of these potential avenues for future investigation, this study has identified a general method of using cross entropy to identify the uniqueness and coherence of various datasets, quantifying overlaps and consistencies within musical corpora.

References

1. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Techn. J. **27**, 379–423, 623–656 (1948)
2. Shannon, C.E., Weaver, W.: A Mathematical Model of Communication. University of Illinois Press, Urbana (1949)
3. Meyer, L.: Meaning in music and information theory. J. Aesthetics Art Criticism **15**(4), 412–424 (1957)
4. Cohen, J.E.: Information theory and music. Behav. Sci. **7**(2), 137–163 (1962)
5. Youngblood, J.E.: Style as information. J. Music Theor. **2**, 24–35 (1958)
6. Mendel, A.: Some preliminary attempts at computer-assisted style analysis in music. Comput. Humanit. **4**(1), 41–52 (1969)
7. Duane, B.: Agency and information content in eighteenth- and early nineteenth-century string-quartet expositions. J. Music Theor. **56**(1), 87–120 (2012)
8. Temperley, D.: Music and Probability. The MIT Press, Cambridge (2007)
9. White, C., Quinn, I.: The yale-classical archives corpus. Empirical Musicology Rev. **11**(1), 50–58 (2016)
10. Quinn, I.: What's 'Key for Key': A Theoretically Naive Key–Finding Model for Bach Chorales. Zeitschrift der Gesellschaft für Musiktheorie 7 (2010)
11. Quinn, I., Mavromatis, P.: Voice-leading prototypes and harmonic function in two chorale corpora. In: Agon, C., Andreatta, M., Assayag, G., Amiot, E., Bresson, J., Mandereau, J. (eds.) MCM 2011. LNCS, vol. 6726, pp. 230–240. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21590-2_18
12. White, C.W.: An alphabet-reduction algorithm for chordal *n*-Grams. In: Yust, J., Wild, J., Burgoyne, J.A. (eds.) MCM 2013. LNCS, vol. 7937, pp. 201–212. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39357-0_16
13. Cuthbert, M.S., Ariza, C.: music21: a toolkit for computer–aided musicology and symbolic music data. In: Proceedings of the International Symposium on Music Information Retrieval, pp. 637–42 (2011)
14. White, C.: Changing styles, changing corpora, changing tonal models. Music Percept. **31**(2), 244–253 (2014)