The MeloSol Corpus

DAVID JOHN BAKER
Louisiana State University

ABSTRACT: This data report introduces the *MeloSol* corpus, a collection of 783 Western, tonal monophonic melodies. I first begin by describing the overall structure of the corpus, then proceed to detail its contents as they would be helpful for researchers working in the field of computational musicology or music psychology. In order to contextualize the *MeloSol* corpus in relation to other corpora in the literature, I present descriptive statistics of the *MeloSol* corpus alongside the *The Densmore Collection of Native American Song* and *The Essen Folk Song Collection*. I suggest possible future uses of this corpus including extending research investigating Western tonality, perceptual experiments needing novel ecological stimuli, or work involving the musical generation of monophonic melodies in the style of Western tonal music.

Submitted 2020 Apr 30; accepted 2005 XXXX X.

KEYWORDS: *melodic corpus, Tonal music, kern, sight singing, aural skills*

SUMMARY

THIS data report introduces the *MeloSol* corpus, a collection of 783 monophonic melodies taken from *A New Approach to Sight Singing: Fifth Edition* (Berkowitz, Fontrier, Kraft, Goldstein, & Smaldone, 2011). The title *MeloSol* derives from a combination of the corpus' content-- *Melo*dic data-- and the first name of the original author of the collection, *Sol* Berkowitz.

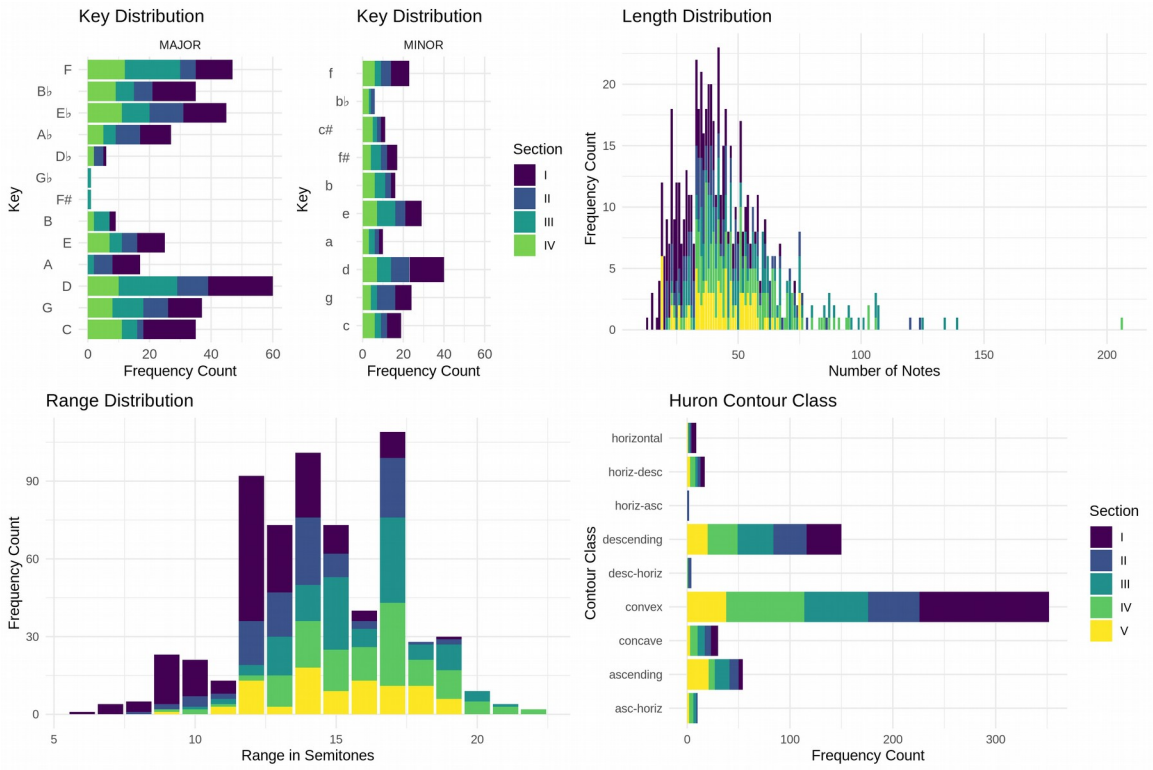
The corpus is divided into two major sections: a collection of sight-singing melodies composed specifically for pedagogical purposes (n = 629) taken from Chapter One and examples from Western classical literature (n = 154) taken from Chapter Five. The original text also contains materials for practicing rhythm (Chapter Two), Singing Duets (Chapter Three), Sing and Plays that incorporate a melody and piano accompaniment (Chapter Four), and Supplementary Exercises that are not included here. Within each of the larger sections exists five further subdivisions. These five subdivisions are mapped in conjunction with the trajectory of many aural skills classrooms.

For example, the first section of both the sight-singing melodies and the examples from Western classical literature align with melodies a first semester undergraduate student in a music degree program might be expected to learn during their first semester of university in an aural skills classroom. As the original book was designed as a pedagogical text, each section of the book and consequently each melody within each section is meant to increase in complexity as new topics are introduced. The fifth and final section of both the sight-singing melodies and examples from the literature contains melodies which break from Western tonal practice. These melodies contain either modal, atonal, or tonally ambiguous melodies. A visual depiction of the breakdown of melodies from the two larger sections is presented in Figure 1.

In terms of analyzable data, the 783 melodies are encoded in **kern format (Huron, 1994), with each individual file containing metadata listing the unique identifier, the chapter from which the melody originates, the section within that chapter of the larger text, its page number, as well as what mode the encoder labeled the melody as. Modes were only noted for a small subset of the corpus, and the vast majority of these melodies are either major (Ionian) or minor (Aeolean).

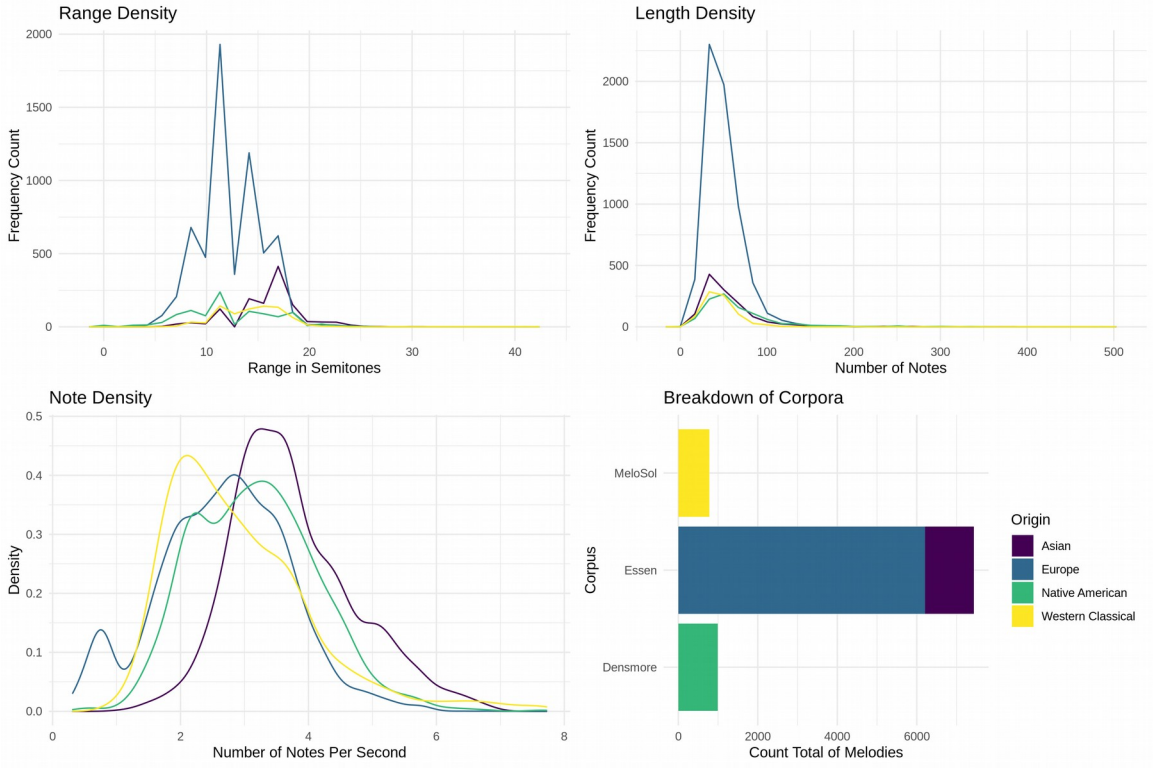
Overall, the corpus consists of 49,730 **kern tokens, a subset of which are 36,641 note heads. All melodies in the corpus were encoded by hand using the software MuseScore (Werner, Nicholas, & Bonte, 2019), initially saved as XML, then converted to **kern using the humdrum extras xml2hum tool (Sapp, 2008) with the current metadata added using the metadata_adder.R script found in the repository. Further addition to the metadata can be added with modifications to metadata_adder.R found in the scripts/R directory of the data repository.

The figures presented in this data report describe the *MeloSol* corpus from a macro perspective. I note that Section V was removed from the top left portion of Figure 1 as the majority of melodies in the atonal section of the corpus are encoded with a zero flat, zero sharp key signature and including those in the figure would skew C major and A minor's representation. The bottom right panel describing the Huron Contour Class is computed using the FANTASTIC toolbox described below and was first presented by Huron (1996).



COMPARISON

In order to further contextualize the *MeloSol* corpus with others found in the literature on musical corpora, I briefly compare descriptive statistics from the *MeloSol* corpus with both *The Densmore Collection of Native American Songs* (Neubarth, Shanahan, & Conklin, 2018; Shanahan and Shanahan, 2014) as well as the European and Asian subset of the *Essen Folk Song Collection* (Schaffrath, 1995). I chose both the *Densmore* as well as the *Essen* collection seeing as both corpora contain monophonic melodies. Further, I compare the *MeloSol* with the *Essen* collection as the *Essen* collection has been used as a proxy for representing the implicit understanding of the structure of Western, tonal music in computational models that depend theoretically on the concept of implicit, statistical learning (Demorest & Morrison, 2016; Huron, 2006; Pearce, 2018).

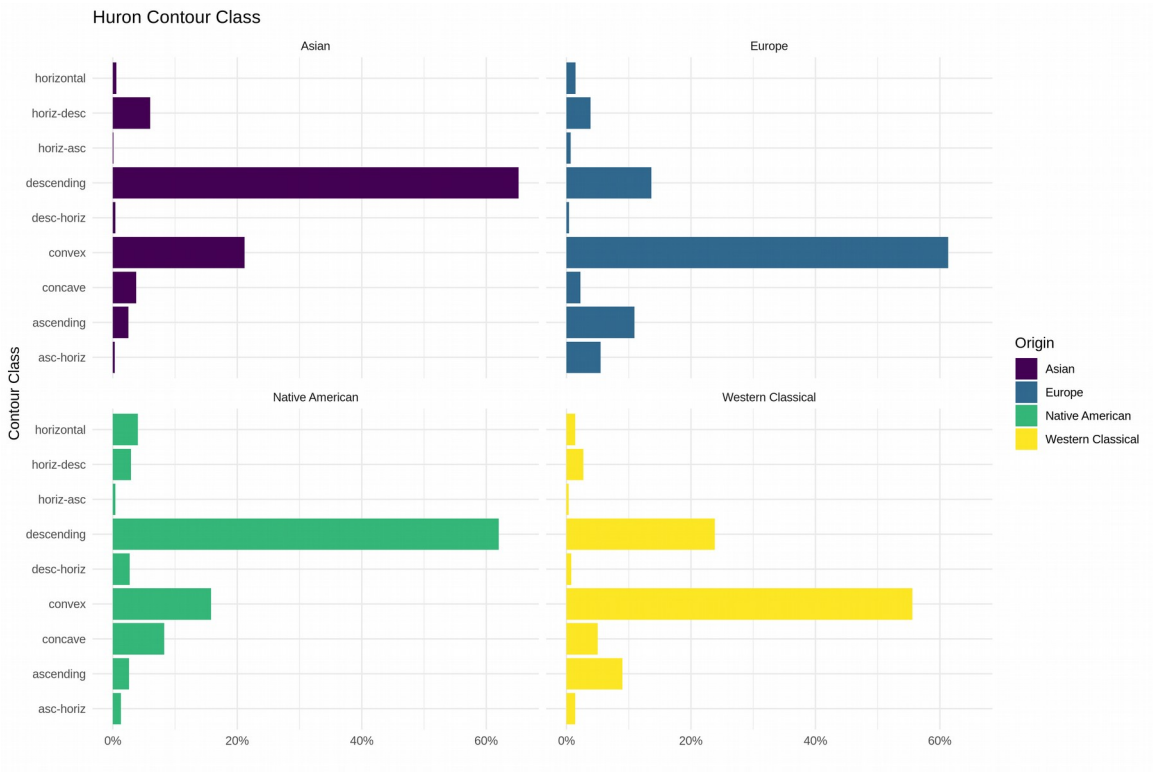


As evident in Figure 2, the *MeloSol* corpus— shown in yellow in all plots— is relatively smaller than the *Essen* and the *Densmore* by a factor of 9.5 and 1.3 respectively. Though given this large difference in total size, the distribution of some summary features exhibit similar features. For example, as evident in the top row of Figure 2, all corpora tend to fall within the same bounds of both the range and length metrics, but exhibit different patterns within those bounds. The range density of the Asian and European subsets of the *Essen* corpus have median pitch ranges of 17 and 12 semitones respectively. The *Densmore* appears to have a distinct peak at 11 semitones, contrasted with no clear peaks emerging from the *MeloSol*. This pattern could possibly reflect the compositional and didactic constraints of the *MeloSol* melodies compared with melodies that do not generate from such a constrained system. While some melodies in all corpora extend far beyond 100 notes, this might actually reflect choices on the encoder in terms of repeated musical material rather than reflecting a through-composed musical form. Future users of this corpus may be interested in performing their own exploratory data analysis with the extreme outliers excluded, as more granular patterns are visible when explored within using smaller bounds on the axes.

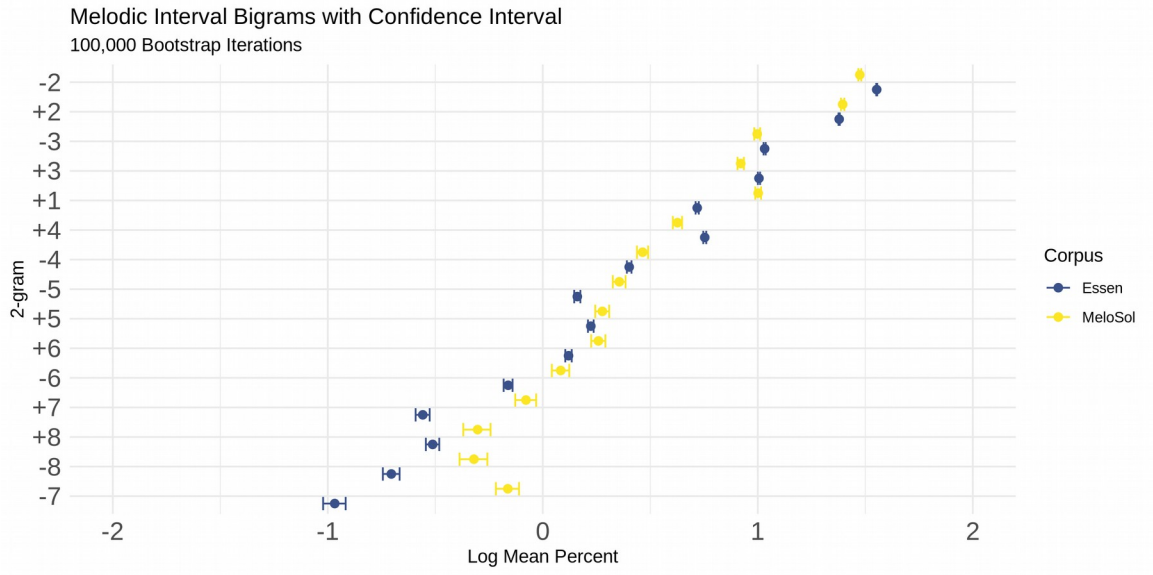
Figure 3 moves from data derived from count features of analyzable tokens and instead computes various summary statistics features derived from the FANTASTIC toolbox (Müllensiefen, 2009). The FANTASTIC toolbox, an acronym for Feature ANalysis Technology Accessing STatistics (In a Corpus), is a computational implementation of summary features of melodies inspired by work in computational linguistics, music theory, and music psychology. Brief descriptions of each feature are provided in the captions of the figures. For a complete description of each of the features, interested readers should consult the FANTASTIC documentation, but I note that all of the FANTASTIC computations for the *MeloSol*, *Essen*, and *Densmore* are found in the *MeloSol* repository for reproducible and future research.

Of interest to readers specifically working with the *MeloSol* corpus might be the generally higher distribution of interval entropy within *MeloSol* and the lower degree of durational entropy (which can be roughly understood as rhythmic entropy) when compared to the the *Densmore* collection, again reflecting the unique pedagogical nature of this corpus. These higher interval entropy calculations most likely reflect

the inclusion of the more highly chromatic and atonal melodies that are included in this corpus. Users of this corpus should be aware of the very clear stylistic diversity that is included in this corpus when considering which sections to decide to include in their analyses.

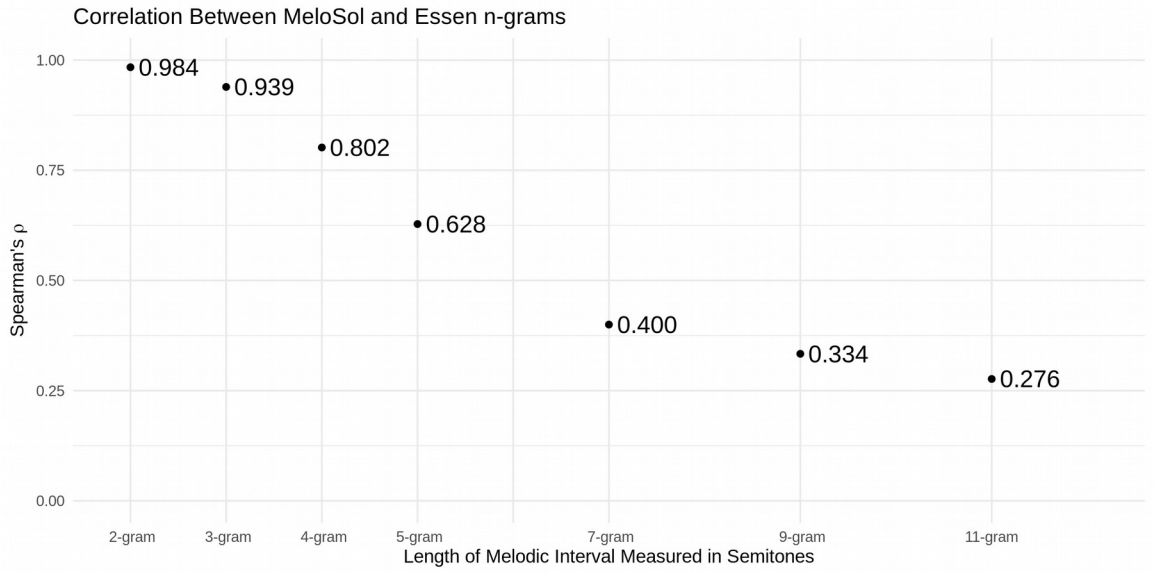


As the *MeloSol* corpus comprises Western, tonal music, this corpus might be used in order to continue research investigating empirical claims regarding patterns found in Western, tonal music. For example, work from Huron (1996) investigated the distribution of melodic arches in 6,251 melodies using the European subset of the *Essen* corpus. In Huron's first analysis of these arches, Huron proposes nine separate contour classes that are readily available for computation via the FANTASTIC toolbox. Since the *MeloSol* corpus does not have the phrase markings needed to exactly replicate the the first analysis of Huron (1996), Figure 4 recreates the analysis of the nine proposed contour classes over the entire melodies from the combined corpus. Despite this, the contours of the *MeloSol* corpus' melodies tend to behave similarly to that of the *Essen* collection with a predominance of convex contours, followed by ascending, then descending contours as initially demonstrated by Baker (2019). This melody-level, as opposed to phrase-level, analysis demonstrates the need for future versions of the *MeloSol* corpus to incorporate finer granularity of annotations for future phrase level analyses.



As the *MeloSol* corpus contains music associated with Western, tonal music, the corpus could also be used in further work in lieu of the *Essen* collection as a dataset in which to train computational models of melodic expectation (Pearce, 2018) within music cognition research. In order to demonstrate the similarity between corpora, I provide estimations of the proportion of how frequently certain absolute intervals occur in both the *Essen* and *MeloSol* corpus. These melodic intervals, or bigrams, are calculated by using the melodic interval tool in the humdrum toolkit. Despite the relative difference in size of the corpora, Figure 5 displays the stability of these intervals. The intervals in this corpora using both the *Essen* and *MeloSol* corpora and are presented with bootstrap confidence intervals (R = 100,000) for each proportion parameter. Said another way, Figure 5 demonstrates that the frequency of interval patterns in the corpora share very similar properties in terms of how frequently they occur. In interpreting this figure, if each interval from each corpus appeared with the same relative frequency, the pairings would match exactly. The presentation of intervals is arranged so that more frequently occurring intervals appear higher in the visualization. The horizontal axes represent the base 10 logarithm of the percentage of that bigram of all the bigrams in the corpus.

The similarity of intervals as indicated by the relative close proximity of pairings from each bigram suggests a high degree of stability given shorter transitional probabilities. In order to explore how the relationship between frequency of n-grams breaks down as a function of length of n-gram, Figure 6 plots the Spearman correlation coefficient reflecting the relationship between the ranked frequency of each size n-gram. For example, moving beyond a bigram that would just comprise two notes (one interval), a 4-gram would be a string of absolute consecutive intervals generated from five notes. Even at the 4-gram level, the *Essen* and the *MeloSol* corpus are strongly correlated at $\rho(1402) = .802$, although this pattern breaks down with longer n-grams. Scripts to reproduce these analyses can be found in the corpus repository.



As expected, an increased n in the n-gram results in a weaker relationship between the two corpora. Despite the *MeloSol* corpus being designed for didactic purposes, the *MeloSol* corpus appears to reflect many of the global characteristics of the *Essen*. Given many of the theoretical assumptions that come along with using a musical corpus as the basis for an individual's latent understanding of a musical structure either now (Huron, 2006) or the in the past (Byros, 2009), perhaps other corpora like the *MeloSol* could be chosen as the basis for these models rather than deferring to the *Essen* on account of size alone. Further research that combines perceptual assumption of statistical learning with corpus studies might further explore how different corpora either do or do not lead to different model fits when modeling perceptual data.

I finally note that as this corpus was initially developed in order to investigate how to make pedagogical improvements in aural skills classrooms, using *MeloSol* for this purpose would be a logical extension to this program of research (Baker, 2019).

CORPUS

The corpus, accompanying data, scripts to reproduce all analyses, as well as the documentation for the *MeloSol* corpus can be found at www.github.com/davidjohnbaker1/melosol.

Acknowledgments

I would like to thank Adam Rosado, Elizabeth Monzingo, and Connor Davis in their help encoding some of the melodies for this corpus as well as Peter Harrison and Kework Kalustian for help on the bootstrap analysis. Additionally, I would like to thank Daniel Shanahan and Craig Sapp for technical support while working with and always teaching me things about humdrum as well as the two anonymous peer reviewers for their ideas that helped improve this manuscript.

Figure Captions

Figure 1. Descriptive Statistics of *MeloSol*

Figure 2. Distributional Comparison of Features of Corpora

Figure 3. Distributional Comparison of Features of Corpora

Figure 4. Huron Contour Class Comparison

Figure 5. Melodic Interval Bigrams

Figure 6. Correlation between *MeloSol* and *Essen* n-grams

REFERENCES

Baker, D. J. (2019). *Modeling melodic dictation* (PhD Thesis). Louisiana State University.

Byros. V. (2009). *Foundations of tonality as situated cognition. 1730-1830: An inquiry into the culture and cognition of eighteenth-century tonality with Beethoven's "Eroica" symphony as a case study.* (PhD Thesis). Yale University.

Berkowitz, S., Fontrier, G., Kraft, L., Goldstein, P., & Smaldone, E. (2011). *A new approach to sight singing* (5th ed). New York: W.W. Norton.

Demorest, S. M., & Morrison, S. J. (2016). Quantifying culture: The cultural distance hypothesis of melodic expectancy. *The Oxford Handbook of Cultural Neuroscience*, 183.

Huron, D. (1994). The Humdrum Toolkit: Reference Manual. Center for Computer Assisted Research in the Humanities.

Huron, D. (1996). The Melodic Arch in Western Folk Songs. *Computing in Musicology*, *10*, 3–23.

Huron, D. (2006). *Sweet Anticipation*. MIT Press.

Müllensiefen, D. (2009). Fantastic: Feature ANalysis Technology Accessing STatistics (In a Corpus): Technical Report v1.5.

Neubarth, K., Shanahan, D., & Conklin, D. (2018). Supervised descriptive pattern discovery in Native American music. *Journal of New Music Research*, *47* (1), 1–16. https://doi.org/10.1080/09298215.2017.1353637

Pearce, M. T. (2018). Statistical learning and probabilistic prediction in music cognition: Mechanisms of stylistic enculturation: Enculturation: Statistical learning and prediction. *Annals of the New York Academy of Sciences*, *1423* (1), 378–395. https://doi.org/10.1111/nyas.13654

Sapp, C. (2008). Humdrum Extras. [Computer Software].

Schaffrath, H. (1995). The Essen Folk Song Collection, D. Huron.

Shanahan, D., & Shanahan, E. (2014). The Densmore Collection of Native American Songs: A New Corpus for Studies of Effects of Geography, Language, and Social Function on Folk Song. In *Proceedings of the Fourteenth Annual International Conference for Music Perception and Cognition*. San Francisco.

Werner, S., Nicholas, F., & Bonte, T. (2019). MuseScore.