Studying Large Plainchant Corpora Using chant21

Bas Cornelissen, Willem Zuidema, and John Ashley Burgoyne
{b.j.m.cornelissen,w.h.zuidema,j.a.burgoyne}@uva.nl
Institute for Logic, Language and Computation
University of Amsterdam
Amsterdam, The Netherlands

ABSTRACT

We present chant21, a Python package to support the plainchant formats gabc and Volpiano in music21, and two large corpora of plainchant. The CantusCorpus contains over 60,000 medieval melodies collected from the Cantus database, encoded in the Volpiano typeface. The GregoBaseCorpus contains over 9,000 transcriptions from more recent chant books in the gabc format. Chant21 converts both formats to music21, while retaining the textual structure of the chant: its division in sections, words, syllables and neumes. We present two case studies. First, we report evidence for the melodic arch hypothesis from the GregoBaseCorpus. Second, we analyze connections between differentiæ and antiphon openings in the CantusCorpus, and show that the systematicity of the connection can be quantified using an entropy-based measure.

CCS CONCEPTS

• **Applied computing** → **Sound and music computing**.

KEYWORDS

plainchant, datasets, gabc, volpiano, melodic arch, differentia

ACM Reference Format:
Bas Cornelissen, Willem Zuidema, and John Ashley Burgoyne. 2020. Studying Large Plainchant Corpora Using chant21 . In *7th International Conference on Digital Libraries for Musicology*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3424911.3425514

1 INTRODUCTION

If one thing stands out about our species' musical behaviour, it is its ubiquity: all cultures seem to make music [12]. Yet, our understanding of music from corpus studies is almost entirely based on Western classical or popular music [15]. Part of the explanation might be the scarcity of large corpora from other traditions. Recent efforts have been addressing this, often under the header of *computational ethnomusicology* [19]. We contribute to the efforts to diversify by converting two existing databases of Christian plainchant into a form suitable for corpus analysis in popular tools: the medieval CantusCorpus[1] and the more recent GregoBaseCorpus.[2] We also release the Python package Chant21 for working with these corpora in music21.[3] Finally, we present two case studies illustrating their usefulness. First, we show that melodic phrases have arch-shaped contours in the GregoBaseCorpus, confirming the general *melodic arch hypothesis* [9]. Second, we focus on a particular problem in chant scholarship and revisit the relation between so-called differentiæ and antiphon openings [17] in the CantusCorpus.

The plainchant on which we focus is, indeed, another European tradition. But it is sufficiently distant from Western classical and popular music, if not in time then certainly in its musical language, to be studied as a separate tradition [10]. The music goes back well over a thousand years, to the ninth century, when the first melodies appear in manuscripts. Multiple chant traditions had coexisted in Europe before then, with their own variants of music and texts, but many were (partly deliberately) displaced by what became known as Gregorian chant. The monophonic melodies are rooted in the recitation of sacred Latin texts which formed the backbone of the liturgy. The first manuscripts therefore only record the text, but later sketches of the melodies appear between the lines of text. These sketches consisted of so called *neumes*, figures indicating the contour of small melodic motifs, but not their exact pitches. Later, these neumes were placed on staff lines to also indicate their exact pitches. This developed into both the modern five-line notation, and the four-line *square notation* used in chant books today. The corpora we present employ both types of notation (figure 1).

The chant repertoire was, sometimes actively, organized along several lines. First of all, chants were classified into a system of eight *modes*, usually grouped in four pairs (Dorian, Phrygian, Lydian, Mixolydian). Two paired modes use the same final note, but differ in their typical range: the so-called *authentic* one moves mostly above the final, the *plagal* one around it. This already shows that modes are *melody types*, more than just the church scales to which they are sometimes associated [14]. Second, different parts of the liturgy use different chant *genres*, from the short, syllabic *antiphons* to the elaborate responsories. Some genres, like antiphons, consisted of freely composed melodies, but others, like psalms, used standard melodic formulae: a reciting tone decorated by an opening and closing gesture particular to the mode of the chant.

Most computational studies of plainchant have concerned optical music recognition of medieval manuscripts. But several recent studies have addressed more musicological questions, also in other chant traditions: Pantelli and Purwins [13] analyzed scale intonation in Byzantine chant, and Biró et al. [2] studied cadences in Torah trope. Closer to the present work, Van Kranenburg and

DlfM '20, 16 October 2020, Montréal, Canada

ACM ISBN 978-1-4503-8760-6/20/10…$15.00
https://doi.org/10.1145/3424911.3425514

[1] CantusCorpus is available at github.com/bacor/cantuscorpus.
[2] GregoBaseCorpus is available at github.com/bacor/gregobasecorpus.
[3] Chant21 can be found at github.com/bacor/chant21; or run `pip install chant21`.

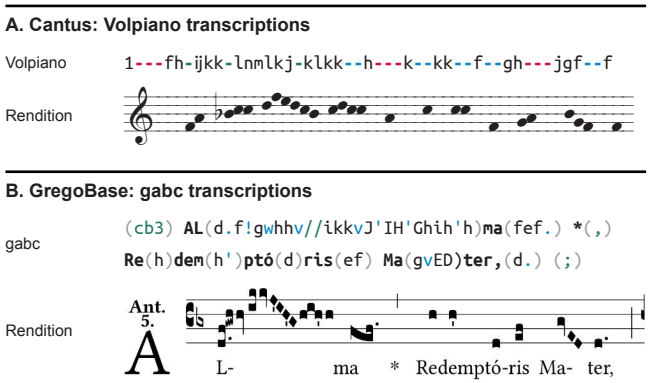

Figure 1: Two versions of *Alma redemptoris mater*. (A) The CantusCorpus contains melodic transcriptions from medieval manuscripts notated in Volpiano: a simple five-line notation. (B) The GregoBaseCorpus contains scores from recent chant books in the gabc format, an elaborate format for four-line square notation.

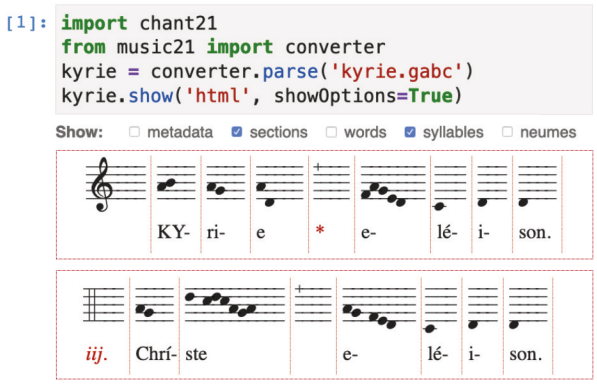

Figure 2: Chant21 in action. Chant21 improves the support for plainchant in Music21 with converters for gabc and Volpiano. It uses a chant representation that divides the chant in sections, words, syllables and neumes. This structure can be interactively explored in Jupyter notebooks.

Maessen [20] used perplexities under an *n*-gram model to classify five early Christian chant traditions and in our own work we have compared several approaches to mode classification [3]. We hope that the two corpora and software we will now present, inspire more computational studies of plainchant.

2 CORPORA

The first corpus we present, the CantusCorpus, is in essence a cleaned-up export of the Cantus database [11]. This is an online index to the many medieval manuscripts kept in libraries across the world. Currently it contains 497,071 chants; the database contains records for almost all, with information on where they are found in which manuscript, but also on things like their incipit, liturgical genre, feast, mode, and a *Cantus ID* to be able to identify the same chants across manuscripts and databases. For 63,628 chants (13%) the melody has also been (partially) transcribed using *Volpiano*.[4]

Volpiano is a typeface that renders text as notes on five staff lines, and was specifically developed for notating plainchant. Several conventions are commonly adhered to, such as the use of three, two and one hyphen(s) to indicate word, syllable and neume boundaries respectively (figure 1A). This allows the music to be aligned to the manuscript text, which is transcribed separately. Many of these conventions have been fixed in the elaborate transcription guidelines of the Cantus database and this is what we refer to as the *(Cantus) Volpiano format*. Such guidelines, and editorial reviews, ensure the high quality of the transcriptions [8].

The Cantus database is easy to use for chant scholars, but not necessarily for computational purposes: it is continuously updated, which is actually inconvenient when replication is a concern. We therefore scraped the database via its API and converted it to a set of clean CSV files which we release as the CantusCorpus. Releases are versioned as we plan to occasionally release newer versions.

Our second corpus, GregoBaseCorpus, again repackages and versions an existing database: GregoBase [1], which provides a complementary perspective on chant. Whereas the Cantus database maps the complexity of medieval manuscripts in a simplified notation (Volpiano), GregoBase consists of modern reinterpretations of the Gregorian repertoire: the one found in chant books like the *Liber Usualis*. Such books are indented for practical use and use the full scope of square notation, including things like breathing marks, different note shapes, rhythmic signs, and clef changes.

The GregoBase website currently hosts 9139 chant transcriptions from 29 books, including the complete *Liber Usualis*. The transcriptions are written in *gabc* (figure 1B), a plain text format for square chant notation, developed for the typesetting system *Gregorio*. We converted the GregoBase database to a set of easy to use CSV files, but also to separate gabc files that include metadata such as the mode, liturgical genre and all books a chant appears in.

3 CHANT21

To make it easier to work with the two corpora we present the Python package chant21 which improves the support for gabc and Volpiano in music21 [4], by now the go-to toolkit for symbolic computational musicology. Chant21 consists of parsers for (1) gabc and (2) Volpiano; (3) a way to align text to music notated in Volpiano; (4) a chant representation which retains the subdivision in sections, words, syllables and neumes; (5) a way to export this representation to HTML, which allows for fast visualization in Jupyter notebooks.

Writing parsers for the elaborate gabc syntax and the informal Volpiano guidelines is not straightforward. After experimenting with custom parsers, we decided to specify the syntax of both formats as *parser expression grammars* (PEGS) [7].[5] Specifying the syntax in a grammar makes it transparent and much easier to maintain. PEGS resemble context free grammars but use a deterministic choice operation to make parse trees unambiguous. After specifying the grammar, we delegate the actual parsing to the PEG parser *Arpeggio*

[4] Of the transcribed chants, 37% contain fewer than 30 notes and are probably incipits.

[5] This idea was borrowed from `gabc-parser`, but we had to completely rewrite the grammar as `gabc-parser` only implements the basic features of gabc and left many chants unparseable.

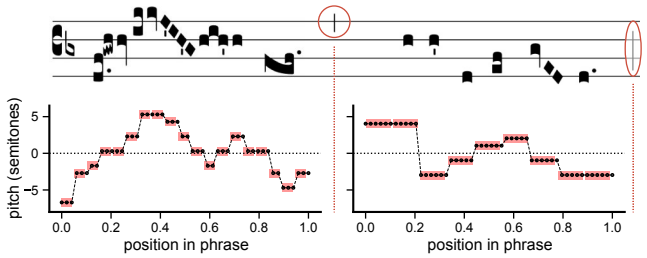

Figure 3: Contour representation. Contours consist of 50 pitches, sampled after normalizing the phrase duration and transposing the phrase by its mean pitch. This is illustrated the first two phrases of the antiphon *Alma mater redemptoris*. The plots below the score shows the contours in black over a red piano roll.

[5]. The resulting parsers are reliable: their error rates are well under 1% when evaluated on the CantusCorpus and GregoBaseCorpus and most failures are caused by syntax errors.

The parse trees of both gabc and Volpiano strings are then converted to music21 objects, but using a custom, hierarchical chant representation which groups the music in sections, words, syllables and neumes. This structure can be useful in computational studies [3], but is also needed to align Volpiano to the text. The Cantus database has guidelines for full text transcriptions: how to for example mark section boundaries, or missing pitches. We use another PEG-based parser to parse the text, and then split all words in syllables using the Latin syllabifier from the *Classical Language Toolkit* [6]. After all this, the text is divided in sections, words and syllables, which we match to their counterparts in the music.

Finally, inspired by the Cantus website, chant21 can export the hierarchical chant representation to HTML, using Volpiano to display the music. This is particularly useful in Jupyter notebooks: it results in much faster typesetting and allows you to interactively explore the structure of the chant. After installing Volpiano and running `pip install chant21`, chant21 is ready to be used (figure 2).

4 CASE STUDY I: THE MELODIC ARCH

To illustrate the usefulness of the presented corpora and software, we discuss two case studies.[6] The first concerns the *melodic arch hypothesis*: the claim that the pitch contour of musical phrases across cultures tend to be arch-shaped. David Huron [9] was the first to present quantitative support for this phenomenon, based on an analysis of 6000, mostly German folksongs from *Essen*. Later studies confirmed the hypothesis in the 2000 Chinese folksongs that were later added to *Essen* [18], and a small global sample of 35 recordings from the Garland Encyclopedia [16].

It has been suggested that the melodic arch is the result of general motor constraints [18]. Those make it easier to produce rising pitch contours at the start of a phrase, when the pressure beneath the vocal folds is rising, and falling contours when the pressure drops towards the end. These constraints could imply a weak tendency for phrases to be arch-shaped (or descending) *on average*, even though individual phrases can take many shapes.

[6] For the data and code of the case studies, see github.com/bacor/DLfM2020

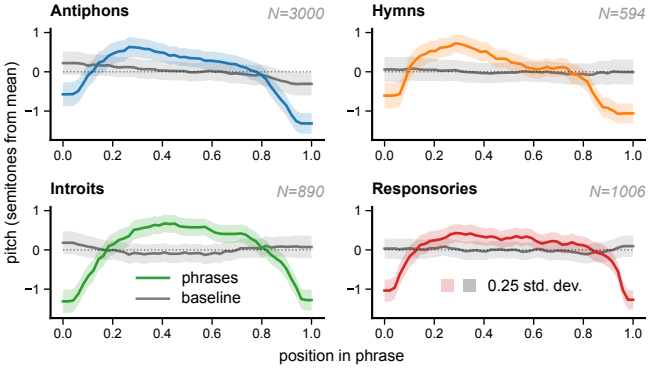

Figure 4: Average phrase contours. The melodic arch hypothesis seems to hold in Gregorian chant. Averaging all phrase contours results in arch-shaped contours (coloured), whereas averaging random segments (grey) yields more or less flat contours. This is illustrated for four chant genres.

We analyze if these findings extend to Gregorian chant and focus on the *Liber Usualis* from the GregoBaseCorpus (v0.4). We extracted phrases using the explicit breathing marks (*pausas*) in chant notation. As rhythmic interpretations of chant vary, we assigned all notes in chants equal duration. We removed duplicate phrases and phrases with fewer than 4 notes, and then randomly sample 3000 phrases per chant genre. Finally, we normalized all phrases to have duration 1 and mean pitch 0, and sampled 50 equally spaced pitches from the resulting contour [16, 18], as illustrated in figure 3.

We average the 3000 normalized contours of a given genre and compare this to the following random baseline. We randomly segment every chant by successively sampling segment lengths from a Poisson distribution approximating the actual phrase lengths. The first and final (random) segments of each chant are omitted. This results in a set of random segments whose lengths are similar to actual phrases, but whose boundaries are unlikely to overlap with actual phrase boundaries. This keeps the melody intact and only shifts phrase boundaries—rather than shuffling all pitches [16].

Figure 4 shows the average phrase contours (coloured) compared to the average random segments (grey) for four chant genres. Whereas the actual phrases are clearly arch-shaped on average, the baseline is pretty much flat. The overall size of the arch is small (around 2 semitones), but similar to earlier findings [16, 18]. The average contours appear to differ across genres, but it requires further analyses to see if these differences are significant. The comparison with the random baseline does however make clear that phrase boundaries have a noticeable and consistent effect on the shape of phrase contours. In sum, these results from this corpus of plainchant are consistent with the melodic arch hypothesis.

5 CASE STUDY II: DIFFERENTIÆ

Our second case study revisits a particular problem in chant scholarship which also figured in a recent edition of this conference: the relation between so-called *differentiæ* and antiphon openings [17]. Every week, monks would sing a cycle of 150 psalms to melodic formulae known as psalm tones. An antiphon was sung before the psalm, and repeated afterwards. The *differentiæ* is the very end of

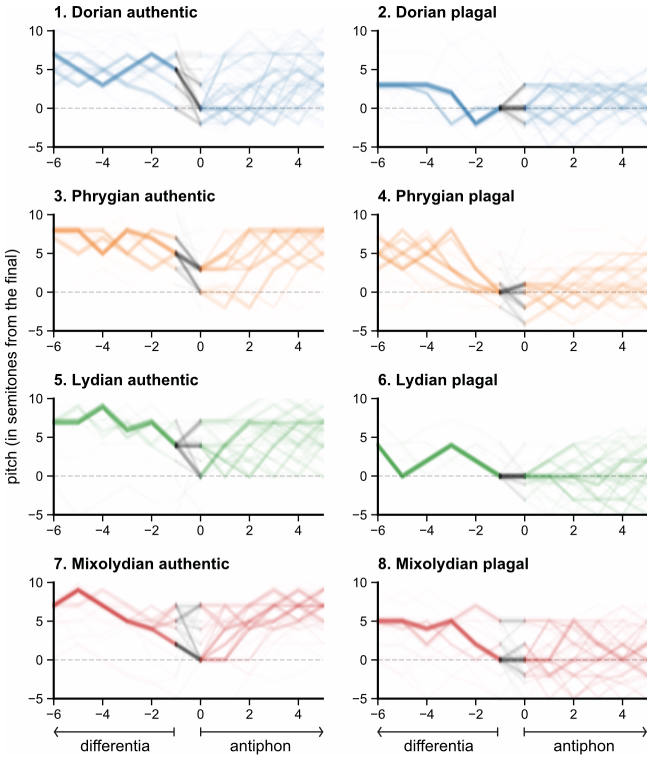

Figure 5: Differentia-antiphon connections in all modes. Each line represents the last 6 notes of the differentia (coloured), followed by the return to the antiphon (black), and 5 more notes of the antiphon (coloured). We sample and show 200 connections per mode, jittered vertically to reveal clusters of overlapping contours.

the psalm, always set to the words *sæculorum amen* (abbreviated as *euouae*) and sung directly before the repetition of the antiphon. The order, in short, was always antiphon–psalm– differentia–antiphon. A question dividing chant scholars is whether there is a systematic relation between differentiæ and antiphon openings: do certain psalm endings usually imply certain antiphon openings?

Rebecca Shaw [17] conducted the first large-scale data analysis and suggests that there is indeed a systematic connection for mode 1. Using chant21 we can extend this to all eight modes by visualizing the connections directly. We selected all 7102 antiphons from the CantusCorpus (v0.1) that had a complete Volpiano transcription, lyrics ending on variants of *aeouae*, and a 'simple' mode (e.g., not transposed). We extract the last 6 pitches of the differentia and concatenate the first 6 notes of the antiphon to obtain the *(differentia–antiphon) connections*. We transpose all connections so that the final has pitch 0.

Figure 5 shows the connections for all modes. The systematicity seems to differ between modes. For example, mode 6 exhibits a very systematic connection: only one differentia is really ever used, and this virtually always leads to the same starting pitch of the antiphon (the final, F). Mode 5, on the other hand, also uses mostly one differentia, but this leads to three possible antiphon openings.

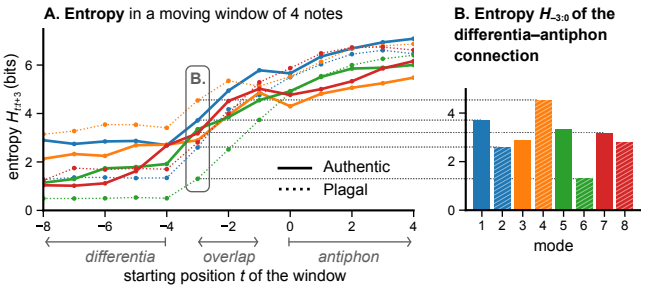

Figure 6: Entropy of the chant. (A) We move a sliding window of 4 notes across the chant and estimate the unpredictability in the window using the entropy $H_{t:t+3}$ (details in main text). This shows that differentiæ ($t \leq -4$) are more predictable than antiphons ($t \geq 0$). (B) Highlights the window containing the last 3 notes of the differentia and the first note of the antiphon, showing for example that the connection in mode 6 is more predictable than in mode 4.

This is certainly less systematic, but more predictable than a random transition.

This suggests a way to quantify the systematicity. For a given mode, consider all the segments $s_{-3:0} = (n_{-3}, n_{-2}, n_{-1}, n_0)$ spanning the last 3 notes of the differentia and the first of the antiphon. If we compute the relative frequencies $p(s_{-3:0})$ of all those segments, we find that in mode 6 only one segment is very frequent, where in, say, mode 4 multiple segments are relatively frequent. One way to quantify this is using the *entropy* $H(p(s_{-3:0}))$ or $H_{-3:0}$ for short, of those relative frequencies: this is a measure of the unpredictability of the chant in the segment from position −3 to position 0. This is what we show in figure 6B. We can repeat this starting at different positions t in the chant, and compute the entropy $H_{t:t+3}$ in all windows of four notes. We did this in figure 6A; it shows how unpredictable different parts of the chant are. It is immediately clear that the more formulaic differentia ($t \leq -4$) are indeed more predictable than antiphons ($t \geq 0$). But we also see that the moment we return to the antiphon, the entropy increases: $H_{-4:-1} < H_{-3:0}$. This suggests that across modes, differentia–antiphon connections are less predictable than differentiæ, but more predictable than antiphon openings.

6 CONCLUSIONS

We have presented two large corpora of Christian plainchant, the Python library chant21 which allows them to be used in music21, and two case studies. First, we showed that phrase contours in the GregoBaseCorpus confirm the melodic arch hypothesis. Second, we show that the connection between differentiæ and antiphon openings is less predictable than the connection between notes within differentiæ, but more predictable than within antiphons. Moreover, the relation clearly differs across modes. Both case studies only scratch the surface, and raise further questions. We hope that this work will inspire more computational studies of plainchant, and broaden the traditions studied by computational musicologists.

REFERENCES

[1] Olivier Berten and contributors. 2013-2020. GregoBase: A Database of Gregorian Scores. https://gregobase.selapa.net/.

[2] Dániel Péter Biró, P van Kranenburg, Steven Ness, George Tzanetakis, and Anja Volk. 2012. Stability and Variation in Cadence Formulas in Oral and Semi-Oral Chant Traditions – a Computational Approach. In *Proceedings of the 12th International Conference on Music Perception and Cognition and the 8th Triennial Conference of the European Society for the Cognitive Sciences of Music*. 98–105.

[3] Bas Cornelissen, Willem Zuidema, and John Ashley Burgoyne. 2020. Mode Classification and Natural Units in Plainchant. In *Proceedings of the 21th International Conference on Music Information Retrieval (ISMIR 2020)*. Montréal, Canada.

[4] Michael Scott Cuthbert and Christopher Ariza. 2010. Music21: A Toolkit for Computer-Aided Musicology and Symbolic Music Data. In *Proceedings of the 11th International Conference on Music Information Retrieval (ISMIR 2010)*. Utrecht, The Netherlands, 637–642.

[5] I. Dejanović, G. Milosavljević, and R. Vaderna. 2016. Arpeggio: A Flexible PEG Parser for Python. *Knowledge-Based Systems* 95 (March 2016), 71–74. https://doi.org/10/f79cb2

[6] Kyle P. Johnson et al. 2014–2019. CLTK: The Classical Language Toolkit. https://github.com/cltk/cltk.

[7] Bryan Ford. 2004. Parsing Expression Grammars: A Recognition-Based Syntactic Foundation. In *Proceedings of the 31st ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages - POPL '04*. ACM Press, Venice, Italy, 111–122. https://doi.org/10/ddkr4k

[8] Kate Helsen and Debra Lacoste. 2011. A Report on the Encoding of Melodic Incipits in the CANTUS Database with the Music Font 'Volpiano'. *Plainsong and Medieval Music* 20, 01 (April 2011), 51–65. https://doi.org/10/bptj5d

[9] David Huron. 1996. The Melodic Arch in Western Folksongs. *Computing in Musicology* 10 (1996), 3–23.

[10] Peter Jeffery. 1992. *Re-Envisioning Past Musical Cultures: Ethnomusicology in the Study of Gregorian Chant*. University of Chicago Press, Chicago and London.

[11] Debra Lacoste, Terence Bailey, Ruth Steiner, and Jan Koláček. 1987-2019. Cantus: A Database for Latin Ecclesiastical Chant. http://cantus.uwaterloo.ca/. Directed by Debra Lacoste (2011-), Terence Bailey (1997-2010), and Ruth Steiner (1987-1996). Web developer, Jan Koláček (2011-).

[12] Samuel A. Mehr, Manvir Singh, Dean Knox, Daniel M. Ketter, Daniel Pickens-Jones, S. Atwood, Christopher Lucas, Nori Jacoby, Alena A. Egner, Erin J. Hopkins, Rhea M. Howard, Joshua K. Hartshorne, Mariela V. Jennings, Jan Simson, Constance M. Bainbridge, Steven Pinker, Timothy J. O'Donnell, Max M. Krasnow, and Luke Glowacki. 2019. Universality and Diversity in Human Song. *Science* 366, 6468 (2019), eaax0868. https://doi.org/10/ggdvjp

[13] Maria Panteli and Hendrik Purwins. 2013. A Computational Comparison of Theory and Practice of Scale Intonation in Byzantine Chant. In *Proceedings of the 14th International Conference on Music Information Retrieval (ISMIR 2013)*. Curitiba, Brazil, 169–174.

[14] Harold S. Powers, Frans Wiering, James Porter, James Cowdery, Richard Widdess, Ruth Davis, Marc Perlman, Stephen Jones, and Allan Marett. 2001. Mode. In *Grove Music Online*. Vol. 1. Oxford University Press. https://doi.org/10.1093/gmo/9781561592630.article.43718

[15] Patrick E. Savage. 2018. An Overview of Cross-Cultural Music Corpus Studies. https://doi.org/10.31235/osf.io/nxtbg

[16] Patrick E. Savage, Adam T. Tierney, and Aniruddh D. Patel. 2017. Global Music Recordings Support the Motor Constraint Hypothesis for Human and Avian Song Contour. *Music Perception: An Interdisciplinary Journal* 34, 3 (2017), 327–334. https://doi.org/10/ggmd4f

[17] Rebecca Shaw. 2018. Differentiae in the Cantus Manuscript Database: Standardization and Musicological Application. In *Proceedings of the 5th International Conference on Digital Libraries for Musicology - DLfM '18*. ACM Press, Paris, France, 38–46. https://doi.org/10/gg572p

[18] A. T. Tierney, F. A. Russo, and A. D. Patel. 2011. The Motor Origins of Human and Avian Song Structure. *Proceedings of the National Academy of Sciences* 108, 37 (2011), 15510–15515. https://doi.org/10/b65mxj

[19] George Tzanetakis, Ajay Kapur, W Andrew Schloss, and Matthew Wright. 2007. Computational Ethnomusicology. *Journal of Interdisciplinary Music Studies* 1, 2 (2007), 1–24.

[20] Peter van Kranenburg and Geert Maessen. 2017. Comparing Offertory Melodies of Five Medieval Christian Chant Traditions. In *Proceedings of the 18th International Conference on Music Information Retrieval (ISMIR 2017)*. 204–210.