Studying large plainchant corpora using chant21

Cornelissen, B.; Zuidema, W.; Burgoyne, J.A.

DOI
10.1145/3424911.3425514

Publication date
2020

Document Version
Final published version

Published in
Proceedings of DLfM 2020

License
Article 25fa Dutch Copyright Act (https://www.openaccess.nl/en/in-the-netherlands/you-share-we-take-care)

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)
We present chant21, a Python package to support the plainchant for-

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
corpus 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 inte-
nation in Byzantine chant, and Biró et al. [2] studied cadences in
Torah trope. Closer to the present work, Van Kranenburg and

 Permission to make digital or hard copies of all or part of this work for personal or
classroom use is granted without fee provided that copies are not made or distributed
for profit or commercial advantage and that copies bear this notice and the full citation
on the first page. Copyrights for components of this work owned by others than the
author(s) must be honored. Abstracting with credit is permitted. To copy otherwise,
or republish, to post on servers or to redistribute to lists, requires prior specific permission
and/or a fee. Request permissions from permissions@acm.org.


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 understand-
ing 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

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

ACM Reference Format:

Permission to make digital or hard copies of all or part of this work for personal or
classroom use is granted without fee provided that copies are not made or distributed
for profit or commercial advantage and that copies bear this notice and the full citation
on the first page. Copyrights for components of this work owned by others than the
author(s) must be honored. Abstracting with credit is permitted. To copy otherwise,
or republish, to post on servers or to redistribute to lists, requires prior specific permission
and/or a fee. Request permissions from permissions@acm.org.

© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.
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.
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.

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]. 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.

**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.
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 (pauses) 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 differentiae 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 differentiae is the very end of
A question dividing chant scholars is whether there is a 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, uses mostly one differentia, but this leads to three possible antiphon openings.

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.

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 differentiae ($t \leq -4$) 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 differentiae, 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 differentiae and antiphon openings is less predictable than the connection between notes within differentiae, 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


