Introduction to Musical Corpus Studies

Fabian C. Moss

13 November 2020

Musikwissenschaftliches Seminar // Universität zu Köln // WS 2020/21

Universität zu Köln

Today

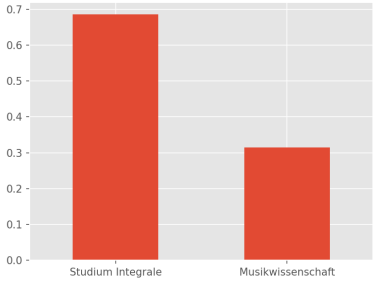
Introduction (16:00–17:20)

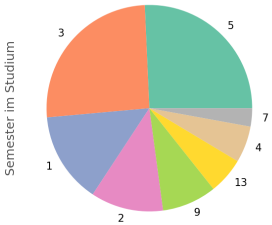
I. What are Musical Corpus Studies? Potential Issues.
II. Case study
III. Organization of the course
IV. Questions

— *Break* —

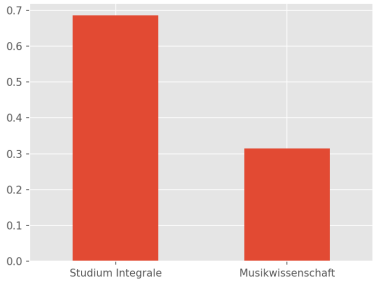
Melody I (17:40–19:00)

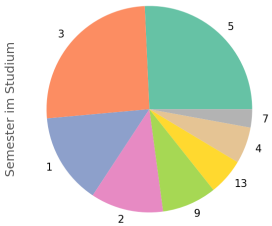
Who are we?


0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Studium Integrale
Musikwissenschaft


Semester im Studium
5
7
4
13
9
2
1
3

Who are we?





1. Who are you?
2. Why did you choose this course?
3. Do you have any musical / technical background?

I. What are Musical Corpus Studies

A possible definition

Corpus studies. *Corpus studies are possibly the most common type of project in computational musicology. A corpus study uses software to analyze statistical patterns in a large collection — corpus — of musical works. It is, essentially, descriptive statistics for musical data. Like text-based corpus studies, musical corpus studies often use n-gram and cluster analysis methods. Unlike text-based corpus studies, musical corpus studies often involve Markov models — probability analyses for progressions in time, such as how likely is [a] C-major chord to progress to a D-minor chord in a piece in the key of A minor. (Schaffer, 2016)*

A possible definition

> ***Corpus studies.*** *Corpus studies are possibly the most common type of project in computational musicology. A corpus study uses software to analyze statistical patterns in a large collection — corpus — of musical works. It is, essentially, descriptive statistics for musical data. Like text-based corpus studies, musical corpus studies often use n-gram and cluster analysis methods. Unlike text-based corpus studies, musical corpus studies often involve Markov models — probability analyses for progressions in time, such as how likely is [a] C-major chord to progress to a D-minor chord in a piece in the key of A minor. (Schaffer, 2016)*

Which of these terms are not clear?

History of Musical Corpus Studies

Short overview (after Temperley and VanHandel, 2013):

- Jeppesen (1927): counts of contrapuntal features in Palestrina
- Cohen (1962), Youngblood (1958): statistics and information theory
- Meyer (1956, 1967, 1989): information theory and relation to psychology
- Krumhansl (1990): algorithm for key finding, relation to music perception
- Huron (2001, 2006, 2016): corpus studies on melodies and voice leading

History of Musical Corpus Studies

Short overview (after Temperley and VanHandel, 2013):

- Jeppesen (1927): counts of contrapuntal features in Palestrina
- Cohen (1962), Youngblood (1958): statistics and information theory
- Meyer (1956, 1967, 1989): information theory and relation to psychology
- Krumhansl (1990): algorithm for key finding, relation to music perception
- Huron (2001, 2006, 2016): corpus studies on melodies and voice leading

Important tools:

- Humdrum (Huron, 1999)
- music21 (Cuthbert and Ariza, 2010)

Data-related issues

- London (2013): representativity
- Pugin (2015): encoding vs OMR
- Neuwirth, Harasim, Moss, and Rohrmeier (2018): balancedness and biases

II. Case study

Example

Example of our most recent research:

- Harasim, D., Moss, F. C., Ramirez, M., & Rohrmeier, M. (in press). Exploring the foundations of tonality: Statistical cognitive modeling of modes in the history of Western classical music. *Humanities & Social Sciences Communications*

Research questions

1. How can we find modes automatically?
2. How can the concept of a mode be operationalized?
3. Can we do it without knowing how many modes there are and what they look like (unsupervised learning)?
4. How do modes change historically?

Corpus

- 21'000 pieces from https://classicalarchives.com
- MIDI format
- user-generated (quality?)
- biases
- metadata: composer names, keys, composition date, …
- representativeness?
- almost no early music examples ⟶ add from other projects
  1. *Citations: The Renaissance Imitation Mass Project* (CRIM)
  2. *The Lost Voices Project*

Corpus

- 21'000 pieces from https://classicalarchives.com
- MIDI format
- user-generated (quality?)
- biases
- metadata: composer names, keys, composition date, ...
- representativeness?
- almost no early music examples $\longrightarrow$ add from other projects
  1. *Citations: The Renaissance Imitation Mass Project* (CRIM)
  2. *The Lost Voices Project*

$\Longrightarrow$ in total 13'402 pieces (ca. 55 million notes) with given composition year (but not key)

Corpus

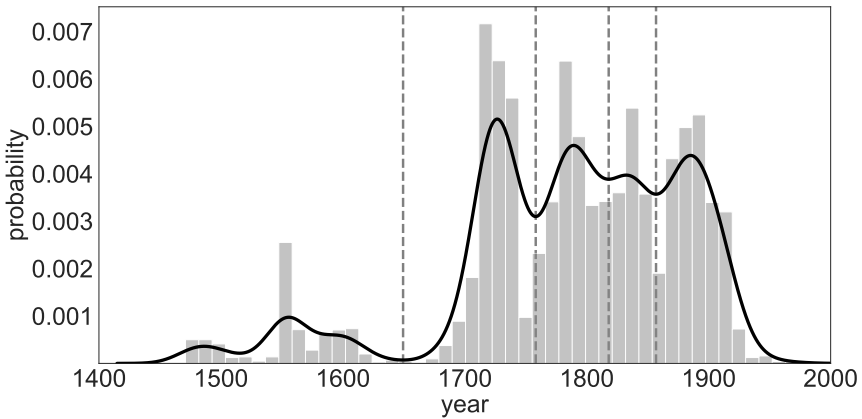


Figure 1: Historical distribution of pieces in the corpus.

Assumptions

1. pieces can be represented by pitch-class counts
2. enharmonic equivalence
3. transpositional invariance

Assumptions

1. pieces can be represented by pitch-class counts
2. enharmonic equivalence
3. transpositional invariance

All of these assumptions are highly questionable, especially on a large historical scale!

Assumptions

1. pieces can be represented by pitch-class counts
2. enharmonic equivalence
3. transpositional invariance

All of these assumptions are highly questionable, especially on a large historical scale!

$\Longrightarrow$ explitic modeling

An example

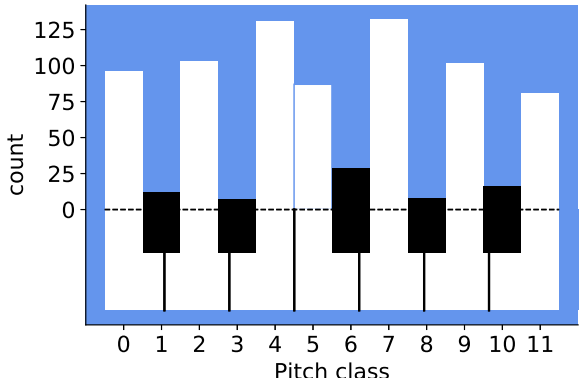


Figure 2: Pitch-class counts of an example piece in C major.

The model

Abstand zwischen zwei Stücken p und q im *key space* $\mathbb{K}$:

$$d_{\mathbb{K}}(p, q) = \left\| \frac{p}{\sum_i p_i} - \frac{q}{\sum_i q_i} \right\|_2, \tag{1}$$

The model

Abstand zwischen zwei Stücken p und q im *key space* $\mathbb{K}$:

$$d_{\mathbb{K}}(p, q) = \left\| \frac{p}{\sum_i p_i} - \frac{q}{\sum_i q_i} \right\|_2, \tag{1}$$

Abstand zwischen zwei Stücken p und q im *mode space* $\mathbb{M}$:

$$d_{\mathbb{M}}(p, q) = \min_{i \in \mathbb{Z}_{12}} d_{\mathbb{K}}(\sigma_i(p), q) = \min_{i \in \mathbb{Z}_{12}} d_{\mathbb{K}}(p, \sigma_i(q)), \tag{2}$$

The model

The optimal mode:

$$(R^*, M^*) = \underset{(R,M) \in \mathbb{Z}_{12} \times \{1,\ldots,m\}}{\operatorname{argmax}} p(R, M \mid T, P, D). \tag{3}$$

In words: Given a piece P in a time period T in the corpus D, the best (mode, root) pair (M^*, R^*) is the one that maximizes the probabilty p.

Automatically finding modes

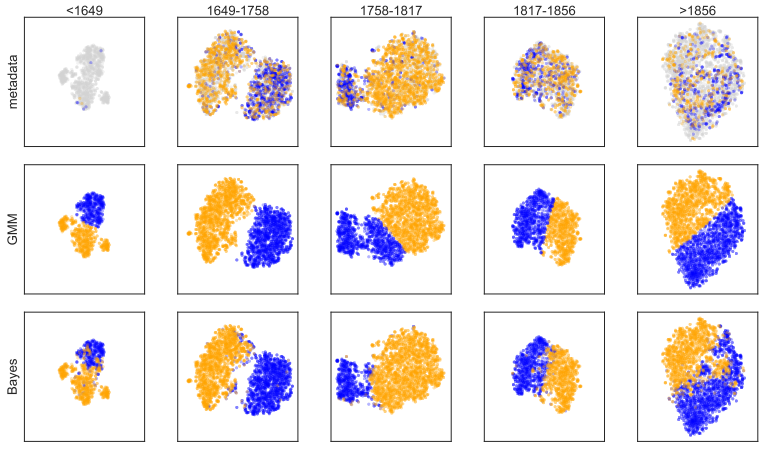


Figure 3: Three models for automatic mode finding.

Quality of the model

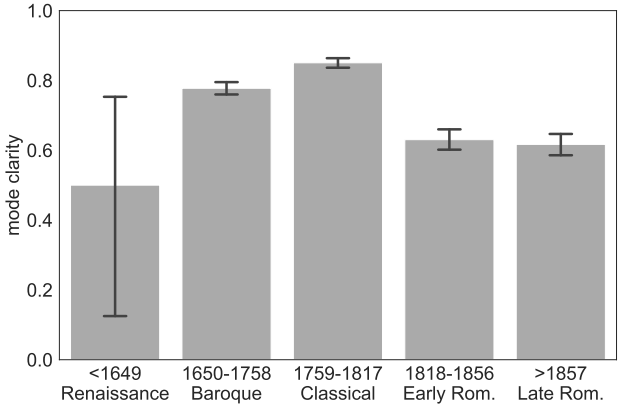


Figure 4: Accuracy scores of our model in five historical periods.

The major and minor modes

Pitch-class distributions of all pieces in the Baroque and Classical periods:

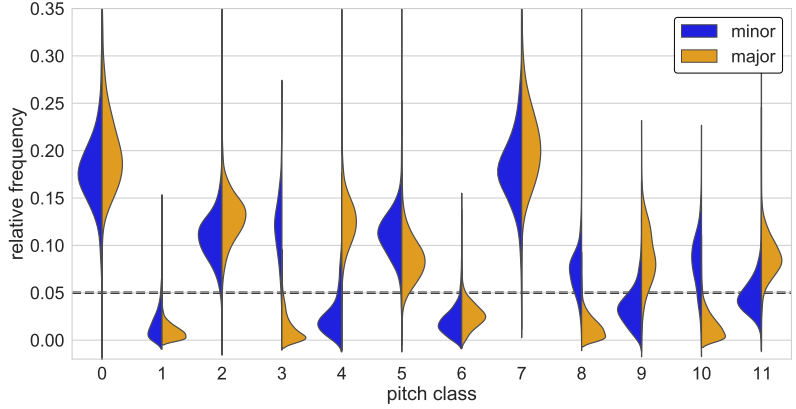


Figure 5: Pitch-class distribution of the major and minor modes.

Modes in the Renaissance

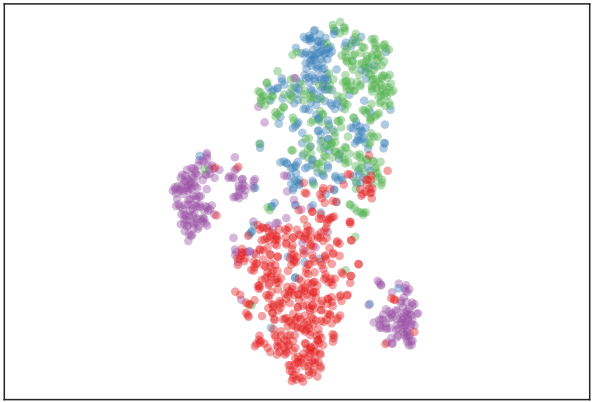


Figure 6: Clustering into four modes in the Renaissance.

Modes in the Renaissance

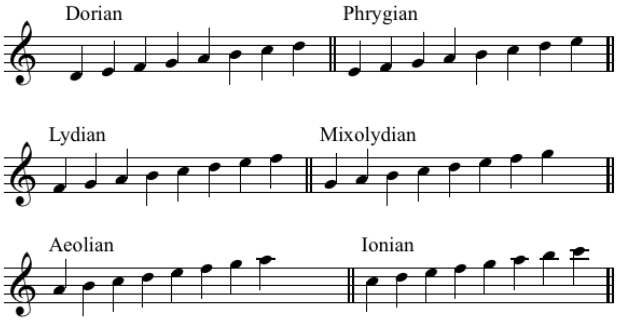
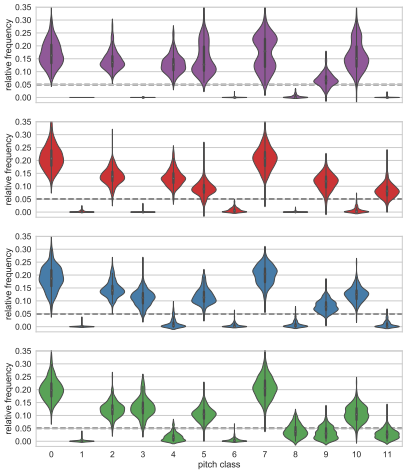


Figure 7: Six modes in early music.

Modes in the Renaissance



Four modes emerge in the Renaissance

- Mixolydian (violet)
- Ionian (red)
- Dorian (blue)
- Aeolian/Dorian (green)

Figure 8: Pitch-class distribution of Renaissance modes.

This course

- The examples in this course are much simpler!
- Quality > Quantity

III. Course organization

Resources

- main organization via ILIAS
- literature
- forum
- Zoom link (you are all here)
- external website: https://fabianmoss.github.io/intro-corpusmus
  - general info
  - course materials (updated after each session)
- HfMT students: by your group members

Credit Points

- 3 CPs = 90 SWS
  - 24 SWS presence in seminar
  - 24 SWS preparation of and follow-up on course materials
  - 42 SWS reading of literature and writing of report

Group work

- you will meet with your group in the breakout rooms
- discussions
- exercises
- Let's test the breakout rooms! (5–7 min for contact info exchange)

Report

- report due on **31 January 2021, 23:59h**
- 6–8 pages
- suggested structure
  1. Introduction
  2. Discussion
  3. Issues
  4. Various
  5. Contributions

Questions?

References

Cuthbert, M. S., & Ariza, C. (2010). Music21: A Toolkit for Computer-Aided Musicology and Symbolic Music Data (J. S. Downie & R. C. Veltkamp, Eds.). In J. S. Downie & R. C. Veltkamp (Eds.), *11th International Society for Music Information Retrieval Conference (ISMIR 2010)2*.

Harasim, D., Moss, F. C., Ramirez, M., & Rohrmeier, M. (in press). Exploring the foundations of tonality: Statistical cognitive modeling of modes in the history of Western classical music. *Humanities & Social Sciences Communications*.

Huron, D. (2001). Tone and Voice: A Derivation of the Rules of Voice-Leading from Perceptual Principles. *Music Perception: An Interdisciplinary Journal*, *19*(1), 1–64. https://doi.org/10.1525/mp.2001.19.1.1

Huron, D. (2006). *Sweet anticipation: Music and the psychology of expectation*. MIT Press.

Huron, D. (2016). *Voice-Leading: The Science Behind a Musical Art*. MIT Press.

London, J. (2013). Building a Representative Corpus of Classical Music [Publisher: University of California Press Journals]. *Music Perception: An Interdisciplinary Journal*, *31*(1), 68–90. https://doi.org/10.1525/mp.2013.31.1.68

Neuwirth, M., Harasim, D., Moss, F. C., & Rohrmeier, M. (2018). The Annotated Beethoven Corpus (ABC): A Dataset of Harmonic Analyses of All Beethoven String Quartets. *Frontiers in Digital Humanities*, *5*(July), 1–5. https://doi.org/10.3389/fdigh.2018.00016

Pugin, L. (2015). The Challenge of Data in Digital Musicology. *Frontiers in Digital Humanities*, *2*, 4. https://doi.org/10.3389/fdigh.2015.00004

Schaffer, K. (2016). What is computational musicology? https://medium.com/@krisshaffer/what-is-computational-musicology-f25ee0a65102

Temperley, D., & VanHandel, L. (2013). Introduction to the special issues on corpus methods. *Music Perception: An Interdisciplinary Journal*, *31*(1), 1–3. https://doi.org/10.1525/MP.2013.31.1.1