Methodologies for Creating Symbolic Early Music Corpora for Musicological Research

Cory McKay (Marianopolis College)
Julie Cumming (McGill University)
Jonathan Stuchbery (McGill University)
Ichiro Fujinaga (McGill University)

With lots of help from Nathaniel Condit-Schultz, Néstor Nápoles López and Ian Lorenz

Motivation

- Scores are increasingly being made available in machine-readable symbolic formats
  - Music XML, MEI, MIDI, Sibelius, Finale, etc.
- Software is increasingly used to carry out studies spanning hundreds of pieces (or more)
  - jSymbolic, music21, Humdrum, MIDI Toolbox, etc.
- Naïve approaches to constructing corpora can limit or bias studies performed on them
  - Can lead to erroneous results and conclusions
  - Worse, these problems may not be apparent to those conducting the studies

Goals of this work

- Propose a robust methodology for creating early music computational research corpora
  - Identification of pitfalls
  - Creation of a model workflow and templates
- Create a sample corpus using this methodology
  - Duos from Josquin and La Rue Masses
- Perform experiments to validate and learn from the sample corpus
  - Using jSymbolic features, statistical analysis and machine learning

Big problem areas

- Interpreting the original notation
  - Many ways to represent and interpret early music in modern notation
  - Essential to have all works in the corpus transcribed using a consistent methodology
- Encoding the music in a computer-readable file
  - Inconsistent encoding can result in unexpected consequences
    - Especially when machine learning is used

Problems with inconsistency and incompleteness

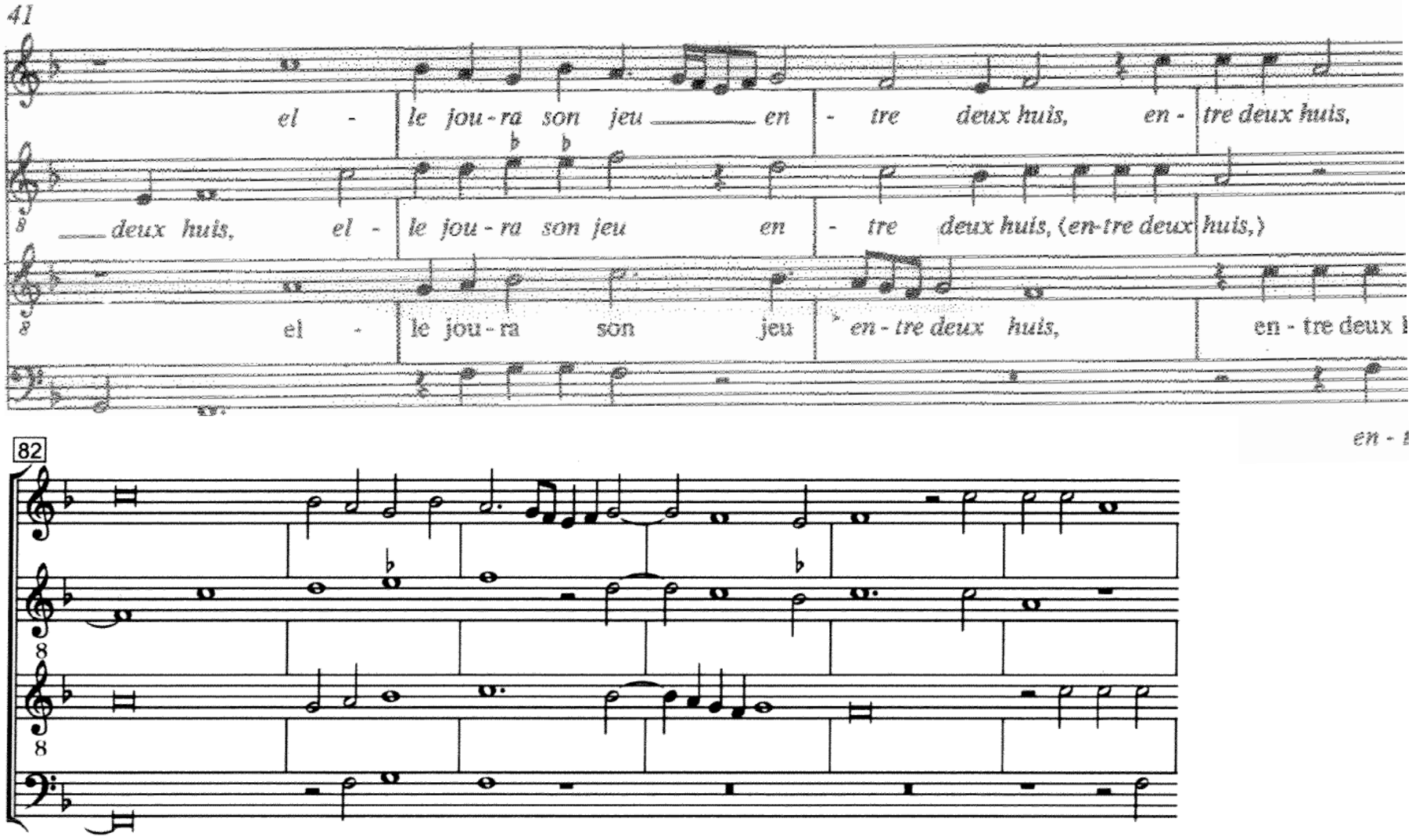
- Computers will be confused if different encoders adopt different standards or make different assumptions
  - Computers will interpret these subjective differences as real differences intrinsic to the music
- Data to be processed by a computer should explicitly specify all necessary information
  - Cannot expect computers to have the same implicit musical knowledge human experts do
  - Many automated algorithms require that information be complete and unambiguous
    - If these decisions are not made explicit in encodings, then algorithms may make their own inappropriate assumptions, or may be unable to process the music at all

Sample interpretation problems (1/2)

- Editors sometimes transpose works to different keys
  - When arranging for specific ensembles
  - Because they believe that the original proper pitch was higher or lower than specified in the source
- Performers can be expected to add accidentals without explicit instructions in the score
  - e.g. *music ficta*
  - Different performers may make different decisions

Sample interpretation problems (2/2)

- Mensuration signs indicate metrical organization
  - But are not quite the same as time signatures
  - And original parts have no barlines, ties are never used
    - Some editions use barlines, some do not
- Note values are larger than those of common Western notation
  - The beat generally falls on the semibreve (whole note)
  - Different editions may use the original, halved, quartered or smaller note values


41
el - le jou - ra son jeu en - tre deux huis, en - tre deux huis,
deux huis, el - le jou - ra son jeu en - tre deux huis, (en-tre deux huis,)
el - le jou - ra son jeu en - tre deux huis, en - tre deux
82
S
C
T
B

Overview of our approach (1/2)

- Use modern notation
  - In order to permit the use of established computational tools that can only process modern notation
- Make as few editorial decisions as possible
  - Encoders thus avoid imposing their subjective interpretations on others
  - e.g. do not add accidentals not specified in the source
    - If a given researcher wishes to add accidentals in a particular way, they can reprocess the files to be consistent in the way they feel is best

Overview of our approach (2/2)

- If an editorial decision must be made, be unwaveringly consistent
  - e.g. use barlines and time signatures, as required by modern notation, but always use the whole note as the beat if this is what is in the source
- If an editorial decision must be made, document it precisely and completely
  - And distribute the resultant workflow with the corpus
  - Those using the corpus will then be made explicitly aware of what decisions were made
    - And can reprocess the corpus to incorporate different editorial decisions if they wish

Sample encoding problems (1/2)

- Some encoding formats do not allow all information of interest to be encoded
  - e.g. MIDI cannot distinguish between a C# and a Db
- Any given piece of analysis software will only be compatible with a limited number of encoding formats
  - But one wants researchers to be able to use the software of the choice
  - MIDI is by far the closest thing to a universal format
    - But MIDI is a deeply flawed format

Sample encoding problems (2/2)

- Encoding software may make editorial decisions of its own, especially under default settings
  - These can vary across software packages
    - Or even across different versions of the same software
    - e.g. Finale and Sibelius may incorporate rubato into saved files if not explicitly told to quantize rhythm
  - Unless care is taken, the encoding software may do this without the knowledge of the encoders operating it

Overview of our encoding approach (1/3)

- Create a detailed workflow and follow it
  - Without exception!
- Use precisely the same software for all encodings (Sibelius)
  - Under the same operating system and settings
- Use pre-constructed templates
  - To maximize consistency and avoid human error
- Use automated scripts
  - To speed the process up
  - e.g. “ManuScript,” the Sibelius scripting language

Overview of our encoding approach (2/3)

- Avoid encoding methodologies that throw out information (when possible)
- Follow consistent labelling standards
  - e.g. if a piece is to be played by viola, always label it exclusively as “viola,” not as a mix of “viola” and “alto,” for example
- Encode provenance in the files
  - In case a file becomes separated from its encapsulating dataset

Overview of our encoding approach (3/3)

- Publish the corpus using multiple different file formats
  - e.g. MIDI, Music XML, Sibelius, etc.
    - Be sure to include MIDI as one of these because of its universality (and despite its flaws)
  - Offers researchers choice
  - Generate all versions from a single original master file
- Verify all final files
  - Manually
    - Labour intensive, but necessary to avoid unforeseen problems (of which there can be many)
  - Automatically
    - To detect things that were missed manually

Our corpus (1/3)

- Duos (surrounded by double bars) from Masses composed by two contemporaries:
  - Josquin Desprez
    - 33 Duos from 11 secure Masses
    - c. 1450-55 to 1521
    - Varied career in France and Italy
  - Pierre de la Rue
    - 44 Duos from 26 secure Masses
    - c. 1452 to 1518
    - Hapsburg-Burgundian chapel, Low Countries and Spain
- Meconi, *Grove:*
  - "Despite differences in style, La Rue's music was probably most strongly influenced by that of Josquin. … There are curious parallels between the works of the two."

Our corpus (2/3)

- Began with Music XML masses downloaded from the Josquin Research Project (JRP)
  - Used Sibelius to extract the duos
- Added additional duos by transcribing them directly using Sibelius
- Processed, cleaned and verified all duos from all sources using the workflow described earlier
  - e.g. restoring original note values
  - To ensure consistency, among other things

Our corpus (3/3)

- Final version will be posted publicly once the paper is accepted
  - Including Sibelius, Music XML, MIDI, MEI and PDF versions of the Duos
  - Including the detailed workflow and templates

Experiments

- We conducted a series of experiments with our Duos corpus
  - To quantitatively explore the effects of using different encoding methodologies
- Trained machine learning models to distinguish the Josquin Duos from the La Rue Duos
  - Used three different version of the corpus, encoded different ways
- I will only summarize the results here
  - Detailed results and analysis are available in the written paper . . .

Experimental conclusions

- The cleaned, consistent version of the dataset produced better results than the original files before cleaning
  - Because inconsistent encoding practices create obscuring noise
- Combining Josquin pieces consistently encoded one way with La Rue pieces consistently encoded another way resulted in grossly inflated performance
  - Because the system "cheated" by basing its classifications on encoding practice rather than the underlying music
  - An important warning not to blindly combine data from different sources

Conclusions and contributions

- Provided a set of principles and workflow for constructing proper early music research corpora
- Constructed a sample corpus of Duos from Masses using this workflow
- Showed experimentally that using consistently and systematically encoded music produces better and safer results

Thanks for your attention

- **E-mail:** julie.cumming@mcgill.ca
- **E-mail:** cory.mckay@mail.mcgill.ca

McGill

Schulich School of Music
École de musique Schulich

CIRMMT Centre for Interdisciplinary Research in Music Media and Technology

Social Sciences and Humanities Research Council of Canada
Conseil de recherches en sciences humaines du Canada

Canada

Marianopolis College

Fonds de recherche sur la société et la culture Québec

DDMAL Distributed Digital Music Archives & Libraries Lab

SIMSSA | Single Interface for Music Score Searching and Analysis

jMIR