The MeloSol Corpus

David John Baker[1]

[1] Louisiana State University

Author Note

David John Baker now works at Flatiron School in London, England.

Correspondence concerning this article should be addressed to David John Baker, .
E-mail: davidjohnbaker1@gmail.com

Abstract

This paper introduces the *MeloSol* corpus, 783 monophonic melodies that represent Western, tonal music. We first begin by describing the overall contents of the corpus, then examine at a mirco level. In order to contextualizs MeloSol, we show descriptive stats comparing it to the Essen and Densmore.We suggest that it might be able to help research that investigates western tonality, difficulty in perceptual experiements, or possibly musical generation in the style as subset of Western tonal.

Keywords: corpus studies, FAIR data, kern

Word count: X

The MeloSol Corpus

Introdution

This data report introduces the *MeloSol* corpus, a collection of 783 monophonic melodies taken from A NEW APPROACH TO SIGHT SINGING FIFTH EDITION. CITE. The title *MeloSol* derrives from a combination of the corpus' content– *Mel* odic 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 = XXX) and examples from the Western Classical Music canon (n = XXX). Within each of the two larger sections exists FIVE further subdivisions. These five subdivisions tend to be mapped in conjunction with aural skills classroom. For example, the first section of both the sight singing melodies and the first section of the Literature align with melodies that a first semester undergraduate student would be expected to learn in their first semester of college in an aural skills classroom. Each section is meant to increase in difficulty. The fifth and final section of both the sight singing melodies and examples from the literature contains melodies either meant to be atonal or have some sort of unstable tonality (bi-tonality/modality). A visual depitction of the breakdown of melodies from the two larger sections in terms of count data is presented IN FIGURE HERE.

- FIGURE HERE

In terms of analyzable data, the 783 melodies are all encoded in `kern` format with each file containing metadata listing the excerpt's LIST HERE. Overall, the corpus consists of XXXXX digital tokens, a subset of which are XXX note heads. All melodies in the corpus were encoded by hand by the author using MUSE SCORE 3, initially saved as XML, then converted to kern using the HUMDRUM EXTRAS `xml2hum` with the current

meta data added using the `name-of-script.R` file. Further addition to the metadata can be added with modifications to `name-of-scrpit.R`.

From a more meaningful point of view, the descriptive statistics of the corpus are displayed in FIGURE TWO and FIGURE THREE.

- FIGURE TWO (subset out Section Five)
- FIGURE THREE (Section Five)

Comparison

Further descriptive statistics of the corpus generated from MULLENSIEFEN'S FANTASTIC TOOLBOX can help contextualize the *MELOSOL* corpus in context with other corpora commonly used in the literature. One of the most cited corpora in the field of computational musicology is ESSEN. ESSEN contains XYZ and is often taken as proxy for representing implicit understanding via statistical learning (HURON) and PEARCE and OTHERS. Essen also has Chinese songs. The *MeloSol* corpus also falls under umbrella of Western Music and as discussed below, may be helpful novel corpus for continuing to investigate claims. Since publication of ESSEN there has also been DENSMORE CITE. DENSMORE is collection of melodies encoded by Shanahan and Shanahan. From musicological point of view, both DENSMORE and CHINESE are expected to be different for reasons of both location as well as style. Here we compare the two to get high level reduction of idea.

First in FIGURE FOUR we compare high level descriptive statistics between WESTERN ESSEN, CHINA, DENSMORE, and MELOSOL. Figure contains comparative overlap of LIST OF FEATURES HERE. Note there is a very large difference in the size of MELOSOL (and others) compared to ESSEN.

- FIGURE FOUR

Second in FIGURE FIVE we compoare MORE ABSTRACT FEAGURES

- FIGURE FIVE

Here is some small comparison on the differences in features.

Useful

As the *MeloSol* corpus is made of Western music, can be used to continue research that has made claims about certain features of Western music if need proxy. For example there are a lot of claims made by HURON about contour class that have been initiall explored by BAKER. There also have been lots of modeling of expeectation using IDyOM by Pearce that have used ESSEN. If buy the idea of sample population as generation, this could be taken forward in that area.

Also note that now have dataset was initially generated using pedagogical materials and might be helfpul in that domain. For example, extending work of MY DISSERATATION could look at proxies of difficulty using FANTASTIC. Could also see if enough data here can be used for generative data analyses using LSTM.

Data analysis

We used R (Version 3.6.2; R Core Team, 2019) and the R-package *papaja* (Version 0.1.0.9942; Aust & Barth, 2020) for all our analyses.

References

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/