Heaps' Law and Vocabulary Richness in the History of Classical Music Harmony

Marc Serra-Peralta (marcserraperalta@gmail.com)
CRM: Centre de Recerca Matematica https://orcid.org/0000-0002-8000-8701
Joan Serrà
Dolby Laboratories Inc
Álvaro Corral
CRM: Centre de Recerca Matematica

Research Article

Keywords: Heaps' law, entropy, MIDI scores, harmonic richness, culturomics

Posted Date: April 12th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-392022/v1

Version of Record: A version of this preprint was published at EPJ Data Science on August 18th, 2021.
See the published version at https://doi.org/10.1140/epjds/s13688-021-00293-8.

Abstract

Music is a fundamental human construct, and harmony provides the building blocks of musical language. Using the Kunstderfuge corpus of classical music, we analyze the historical evolution of the richness of harmonic vocabulary of 76 classical composers, covering almost 6 centuries. Such corpus comprises about 9500 pieces, resulting in more than 5 million tokens of music codewords. The fulfilment of Heaps' law for the relation between the size of the harmonic vocabulary of a composer (in codeword types) and the total length of his works (in codeword tokens), with an exponent around 0.35, allows us to define a relative measure of vocabulary richness that has a transparent interpretation. When coupled with the considered corpus, this measure allows us to quantify harmony richness across centuries, unveiling a clear increasing linear trend. In this way, we are able to rank the composers in terms of richness of vocabulary, in the same way as for other related metrics, such as entropy. We find that the latter is particularly highly correlated with our measure of richness. Our approach is not specific for music and can be applied to other systems built by tokens of different types, as for instance natural language.

Full Text

Due to technical limitations, full-text HTML conversion of this manuscript could not be completed. However, the manuscript can be downloaded and accessed as a PDF.

Figures

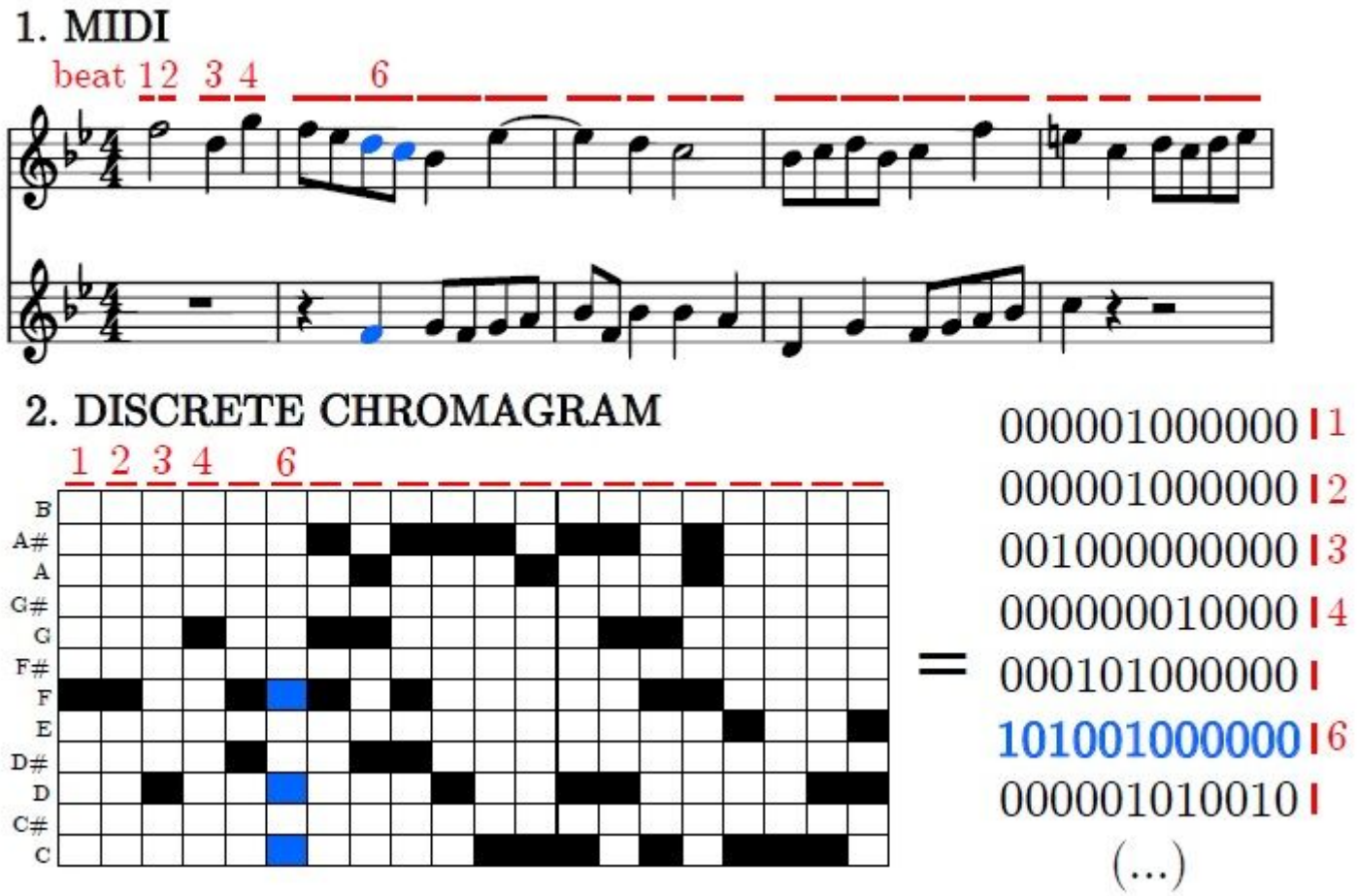

Figure 1

Example of a sequence of discretized chromas (chromagram) arising from a MIDI score. Note that the two pentagrams constituting the score have to be read in parallel.

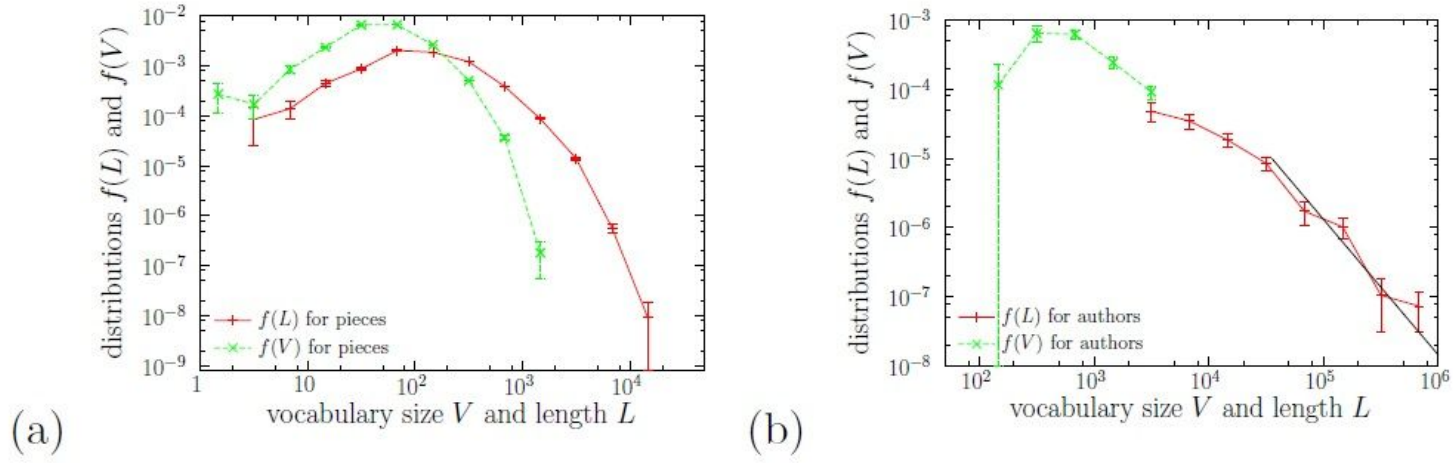

Figure 2

Probability mass functions of length L and vocabulary size V . (a) For individual pieces. (b) For individual authors. A power-law t to the tail of f(L) is shown as an indication (straight line), with an exponent 1:95 ± 0:18 (the tting method is the one in Refs. [42, 43]).

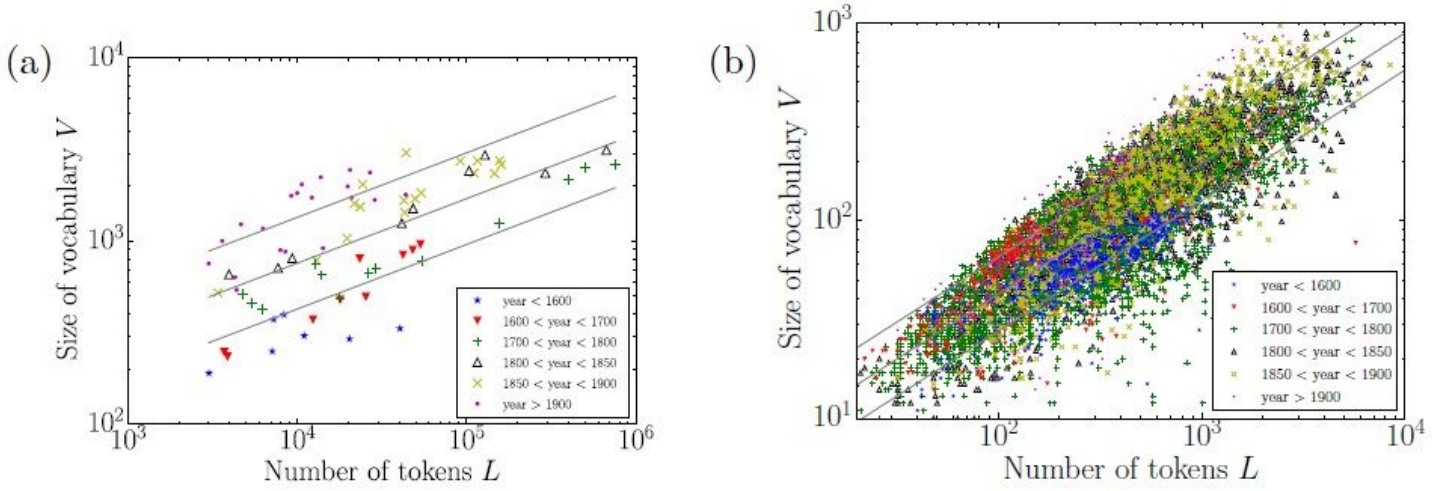

Figure 3

(a) Scatter plot of size of vocabulary versus number of tokens (length) for the 76 composers in the corpus, who are grouped chronologically, as represented by the points of different color (one point is one composer). The year of each composer is the mean between the birth year plus 20 and the death year. Heaps' law is given by the central straight line and the parallel lines denote one standard deviation oc. Notice that when a limited chronological span is considered, the scattering is reduced. (b) Analog scatter plot for the 9489 individual pieces. Year of a piece is approximated to the year of its composer. The results of the fits are in Table 3 (in Appendix I).

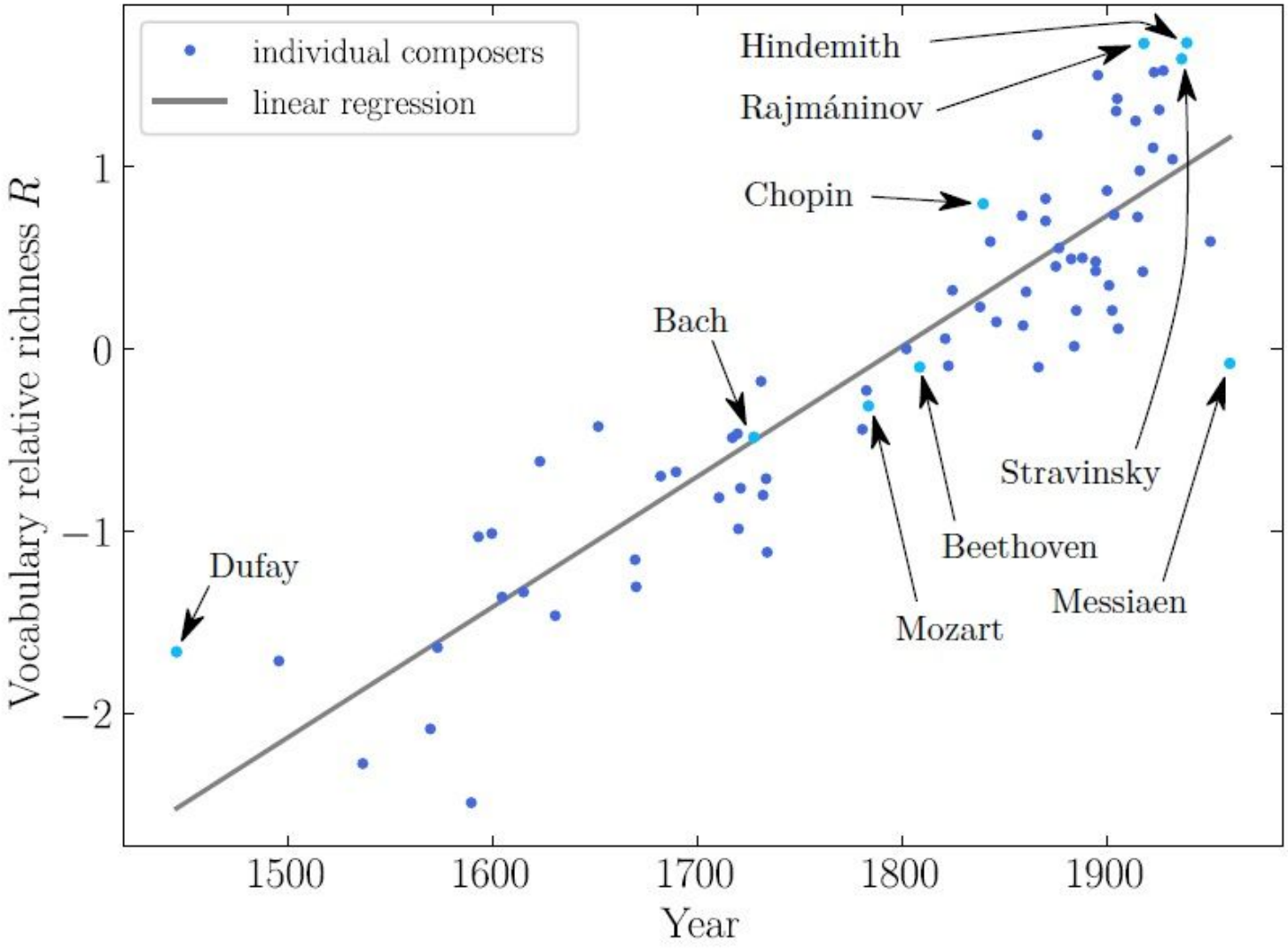

Figure 4

Vocabulary relative richness R for each composer in chronological order. Horizontal axis is birth + 20 + death, divided by 2. The straight line is a linear regression with slope 0:72 ± 0:04 \units of richness" per century and a linear correlation coefficient p = 0:90 (and intercept -12:9). Some particular composers are highlighted, for the sake of illustration.

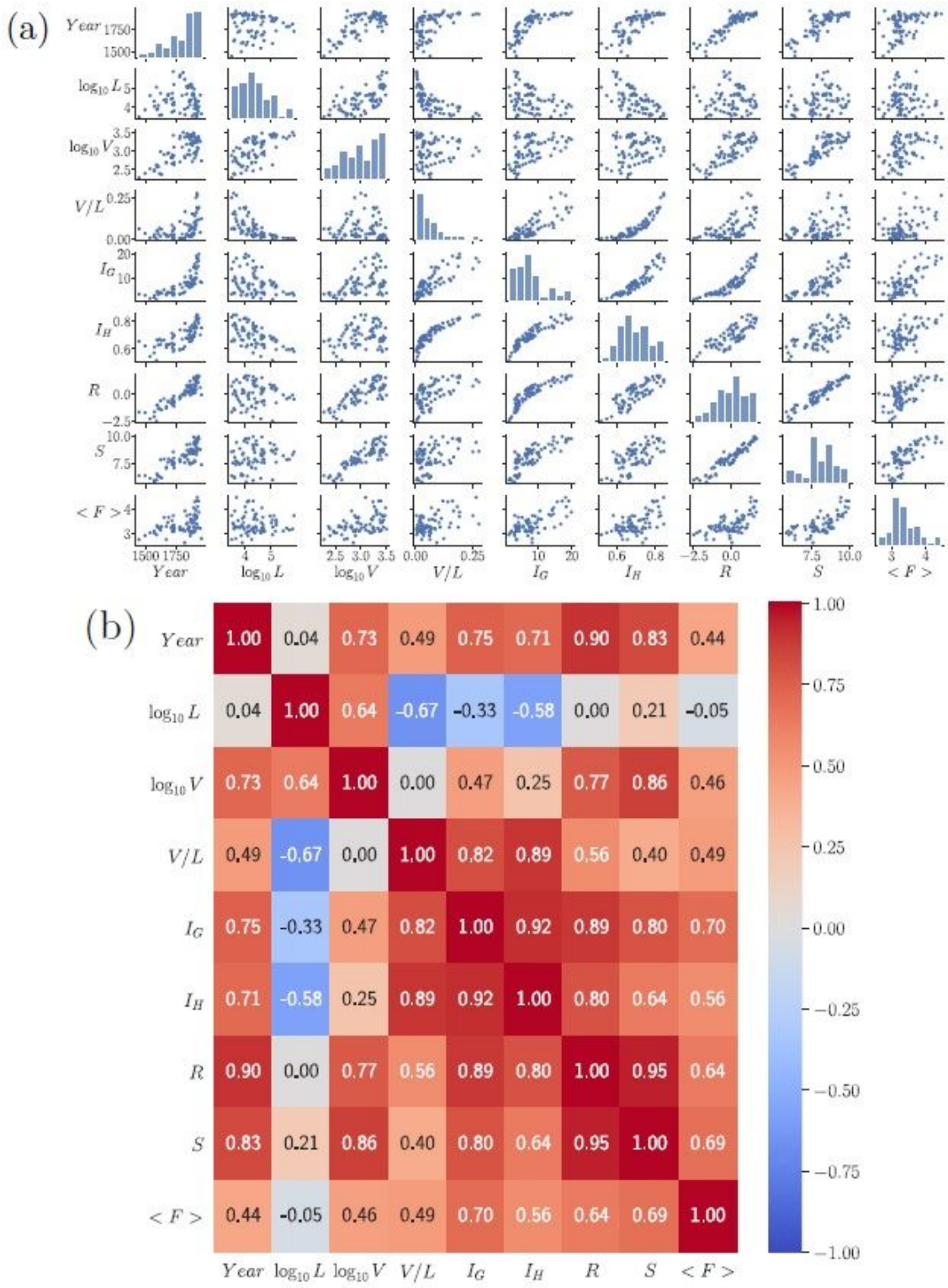

Figure 5

Comparison of the different metrics characterizing each composer. (a) Scatter plots. (b) Matrix of linear correlation coefficients.

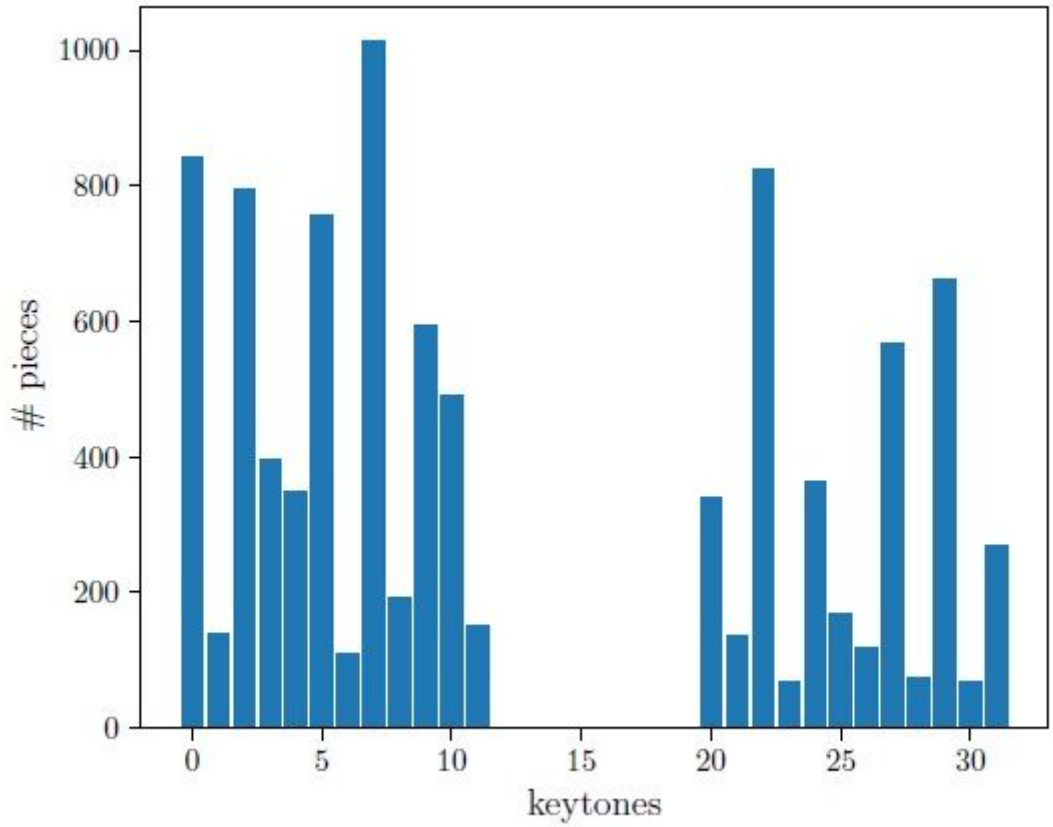

Figure 6

Absolute abundance of each key in the corpus, counted in number of pieces, before transposition, obviously. Zero corresponds to C Major, one to C#/D[Major... up to B Major; 20 corresponds to C minor and so on.