# The Rest is Silence: Leveraging Unseen Species Models for Computational Musicology

Fabian C. Moss[1], Jan Hajič jr.[2], Adrian Nachtwey[3], and Laurent Pugin[4]

[1] Institut für Musikforschung, Julius-Maximilians-Universität Würzburg, Würzburg, Germany
[2] Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic
[3] KreativInstitut.OWL, Paderborn University, Paderborn, Germany
[4] RISM Digital Center, Bern, Switzerland

**Abstract**

For many decades, musicologists have engaged in creating large databases serving different purposes for musicological research and scholarship. With the rise of fields like music information retrieval and digital musicology, there is now a constant and growing influx of musicologically relevant datasets and corpora. In historical or observational settings, however, these datasets are necessarily incomplete, and the true extent of a collection of interest remains unknown — silent. Here, we apply so-called Unseen Species models (USMs) from ecology to areas of musicological activity. After introducing the models formally, we show in four case studies how USMs can be applied to musicological data to address quantitative questions like: How many composers are we missing in RISM? What percentage of medieval sources of Gregorian chant have we already cataloged? How many differences in music prints do we expect to find between editions? How large is the coverage of songs from genres of a folk music tradition? And, finally, how close are we in estimating the size of the harmonic vocabulary of a large number of composers?

**Keywords:** Unseen Species Models, Computational Musicology, RISM, Gregorian Chant, Corpus Studies, Chord Vocabularies, Archives, Databases

## 1 Introduction

Many research questions in the Computational Humanities rely on distributional data about cultural objects and artefacts, often gathered in observational rather than controlled experimental studies. Found distributions of these objects are thus heavily shaped by uncertainties associated with historical transmission processes, including missingness (e.g. because something lies hidden in some basement) or loss (e.g. because a library burnt down). What's worse, if there is no external record about a missing or lost item, there is no way of knowing that it had ever existed. In order to understand the representativeness of observed samples in the humanities, it is thus of great interest to be able to gauge how much we should expect to be missing.

This problem structurally resembles similar issues in species ecology, where researchers need to estimate the number of species from a limited set of incomplete samples. These belong to the class of Unseen Species models (USMs), which have recently been employed in cultural contexts as well: in computational literary studies, for estimating the true size of Shakespeare's vocabulary [9], most prominently in a comparative study of loss of medieval chivalric epics across different European cultures [24], loss in mid-19th century Russian poetry [33], or library records [25]. Beyond literary studies, the size of the Dutch East India Company [46] or the population of a specific prison in Brussels [22]. In musicology, these have been used to compare the diversities of secular and sacred late medieval Italian repertoire [7]. While awareness of this "borrowing" from ecology is growing, the potential of these models to give a better grasp of the unknown, to guide quantitative musicological research, and to provide new insights for data collection efforts is far from realised.

* These authors contributed equally.

Fabian C. Moss, Jan Hajič jr., Adrian Nachtwey, and Laurent Pugin. "The Rest is Silence: Leveraging Unseen Species Models for Computational Musicology." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 526–542. https://doi.org/10.63744/tP4bLwLkye8B.



© 2025 by the authors. Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0).

In this contribution, we probe the usefulness of USMs for musicology in four different case studies using sets of musicological data. We aim to demonstrate what kinds of questions can be addressed by applying USMs in music research, and to provoke discussions about the strengths and limits of this approach. To that end, we first introduce the methodology in Section 2, and apply it to four musicological case studies in Section 3, namely the RISM and Cantus databases (Section 3.1), a dataset of differences between 19th-century music prints (Section 3.2), a dataset of folks music sessions (Section 3.3), and, finally, a large corpus of harmonic annotations in pieces from different composers (Section 3.4). We discuss our results in Section 4 and conclude with their implications and potential for (computational) musicology (Section 5).

## 2 Methods: Unseen species models in the Computational Humanities

Translating musicological inquiries for unseen-species models means asking the following question:

- How many distinct cultural units (unknown species) did we not observe yet?

This question is critical to the representativeness of musicological data. For example, the Cantus database has "only" indexed a few hundred of the tens of thousands of extant manuscripts of Gregorian chant [16] — but chant is supposedly a highly conserved, stable tradition. How well is the entire Gregorian repertoire covered by these sources?

**Modeling the abundance of species — cultural units.** In virtually any kind of collection of cultural objects, some units are highly abundant (very common; e.g., the same composition occurs in many manuscripts, the same song is known across an entire region, the same musical patterns occurs over and over in a composer's work), while others are rare, possibly occurring only once or twice. The (unknown) probability to encounter an instance of a certain unit — observing a specimen of a species — is thus proportional to the number of times that unit occurs in the whole tradition (also generally unknown). This is called the *relative abundance* (or relative frequency) $p_i$ of a unit $i$, and probability theory ensures that $p_i \geq 0$ and $\sum_i p_i = 1$, for $i = 1, \ldots, S$, where $S$ stands for the total number of cultural species (the quantity of interest here).[1] For an observed collection of $n$ cultural units $i$ with abundances $X_i$, it holds that

$$\sum_{i=1\ldots S} X_i = n. \tag{1}$$

For some — possibly even most — species, $X_i = 0$, i.e. they are not observed. While they were at some point present, written down or perhaps sung by someone, they don't figure in the particular dataset/catalog/songbook observed. The *absolute frequency* of species with a particular abundance $r$ is defined as:

$$f_r = |\{X_i \mid X_i = r\}|. \tag{2}$$

This implies that $f_1$ is the number of species observed only once (singletons)[2] and $f_2$ is the number of species observed twice (doubletons). With $f_0$ we denote the (unknown) number of species that was not observed. It logically follows that the total number of distinct species observed in the sample, $S_{\text{obs}}$, is given by

$$S_{\text{obs}} = \sum_{r=1\ldots\infty} f_r. \tag{3}$$

[1] Relative abundances have been generalised to cover also the probability of species being detected in the first place [3].

[2] In the context of corpus studies in natural language processing (NLP), singletons are sometimes called *hapax legomena* (Greek for 'read only once') [32].

This, finally, allows us to define our quantity of interest, $S$: the true size of the musical tradition (for which we are only looking at a limited sample) measured in terms of the number of distinct musical items (songs, manuscripts, harmonies, etc.). It is simply the number of items observed plus the number of items not observed:

$$S = S_{\text{obs}} + f_0. \tag{4}$$

Since $S_{\text{obs}}$ is known (Eq. 3) and the true value of $f_0$ cannot be known, finding $S$ relies on estimating $f_0$ based on limited samples. Estimating $f_0$ is where individual Unseen Species models from ecology come into play.

**Modeling species incidence.** For some musicological scenarios, it seems more appropriate to track *incidence*, i.e., the presence or absence of some musical unit in a sample, instead of counting how many times each individual species was observed. Incidence-based models look at $m$ different samples and re-define $f_r$ to represent the number of species observed in $r$ samples (instead of $r$ times in a single sample in the case of species abundance). We only care about *whether* a cultural unit has appeared in a sample.

Decisions such as choosing between abundance and incidence are the responsibility of the experiment designer(s) and depend on factors that can better be addressed by theory than by empiricism [8].

**Chao estimators.** Popular estimators for species abundance and incidence are the Chao estimators [1], which have already been applied in the computational humanities [23; 24]. Among those, "Chao1"[3] is one particular way of estimating $S$ from the relative frequency counts $f_r, r > 0$. It is formally defined as:

$$S = S_{\text{obs}} + \frac{f_1^2}{2 \cdot f_2}, \tag{5}$$

that is, it estimates the number of species yet unseen from the numbers of species observed only once or twice.[4] From Equations 4 and 5 follows that the number of unseen species is given by

$$f_0 = \frac{f_1^2}{2 \cdot f_2} = S - S_{\text{obs}}, \tag{6}$$

and we define *species coverage* (the overall ratio of species already observed) as

$$\hat{c} = \frac{S_{\text{obs}}}{S}. \tag{7}$$

As mentioned above, the estimator is based solely on the count of singletons ($f_1$) and doubletons ($f_2$). The basic intuition behind this assumption is that distributions of counts usually have a 'long tail': few items occur very frequently and many items occur rarely [4], and the probability mass available for yet unseen species should follow from the length of the tail.[5]

Crucially, this estimator is non-parametric [1]: it can be used regardless of the underlying distribution of relative abundances or incidences $p_i$. This is crucial because this distribution is usually unknown and allows for the application of Chao1 across a wide range of research areas.

[3] Other common methods are the Abundance Coverage Estimator (ACE) [2], Jackknife [43], and Good-Toulmin estimators [15; 39]. Here, we opt for using Chao estimators because a) they provide a conservative lower bound [24]; b) because they have already been used in other applications of the unseen species model to the humanities; and c) because they are straightforward to compute; and d) because their interpretation naturally flows from their formal logic. A python impelementation is available in the `Copia` library [23].

[4] In cases where both abundance and incidence is studied, $f_r$ for incidence is commonly denoted by $Q_r$, and the incidence-based estimator is called "Chao2" [1; 3].

[5] A constructive proof derived from Good-Turing smoothing [11; 12] is provided in [3].

The Chao estimates for $f_0$ are lower bounds [3]: they provide the *minimum* expected number of unseen species (we could still be missing more). Consequently, the estimated species coverage constitutes an *upper* bound: a value of 0.5 means that we have observed *at most* half of all the musical species in some collection or repertoire.

**Relationship to Type-Token Ratio.** We also provide the numbers of types ($n_t$) and tokens ($n_T$) — corresponding to incidence and abundance data, respectively — for the respective corpora to calculate the the type-token ratio (TTR $= N_t/N_T$) that has been used in computational linguistics and computational humanities to characterise lexical diversity [34].

TTR only considers the global number of tokens and thus corresponds to the expected number of individuals per species under a uniform model. While both Chao1 and TTR characterise diversity from categorically distributed data, their relation is not straight-forward. A dataset with a few very dominant "species" but many rare ones may have a low TTR ratio but at the same time a high estimated proportion of unseen species; a dataset where everything occurs twice or a few times but rarely just once will have very low $f_0$ estimate but a high TTR. As they both do relate to underlying diversity, we do expect TTR and Chao1 to still be correlated, but one should not expect this correlation to be particularly strong.

**Accumulation curve.**

## 3 Applications for Computational Musicology

A crucial step in applying Unseen Species Models from ecology to cultural contexts in the humanities in general and to musicology in particular is to draw convincing analogies of what the concept of species corresponds to. Here, we draw an analogy between biological species and some music(ologic)al entity in four case studies relevant for different branches of musicological research. We start by looking at unseen composers and repertoire in large collections of music sources (historical musicology), move on to differences in music prints (music philology), followed by analyzing sessions of folk musicians (ethnomusicology) and harmonic vocabularies in different repertoires (music theory).

### 3.1 Case Study 1: Databases and archives

Empirical conclusions drawn about a musical tradition from a database rely on its representativeness. While one cannot know what is not represented in a database (this effort could just be spent better by adding the given items to the database!), we can use the Unseen Species models to estimate *how much* of whatever entity we define as the "species" is not covered by the data source. We can thus quantify how much "cultural diversity" has not yet been documented. We present here reports from two of the largest musicological databases: RISM, by far the largest database of musical sources, and the Cantus database of Gregorian chant.

#### *3.1.1 RISM: Counting composers*

How many composers were active in Europe since the Renaissance? How many composers are we still to discover whose works lie undetected on some attic or archive shelf? There is no better database to answer such questions than the *Répertoire International des Sources Musicales* (RISM),[6] a database of more than 1,500,000 musical sources assembled across nearly 3,000 holding institutions. In this setting, each composer (identified by a RISM authority record) can be thought of as a species, and the presence of a composer's work in a source catalogued in RISM is then an observation of that species. For a given institution, the composer observation count is the number of sources held by that institution in which the composer appears.

[6] `https://rism.online/`

The resulting dataset contains records for 48,524 composers[7] with works observed across 2,933 holding institutions, with a total of 2,009,343 observations of composers appearing in sources.[8] A composer is a singleton if they are observed in only one source in a single institution, and a doubleton if observed exactly in two sources, whether in the same institution or not.

**RISM Results.** Aggregated over the entire dataset, the Chao1 estimate gives us a lower bound of 78,432 species — composers, with 95% confidence interval widths of $(-844.7, +795.3)$.[9] With 48,524 composers observed, that implies an $f_0$ of 29,908 $(-844.7, +795.3)$ unobserved composers and a coverage upper bound of 0.619 $(\pm 0.01)$ — we have so far recorded in the RISM database at most some 62% of all composers, indicating that there might be plenty of musical diversity to discover.

If we aggregate results only over the 10 largest institutions, each of which holds 20,000+ composer records, we get an estimated $S = 32,989$ $(-515.5, +589.5)$ total composers with $S_{\text{obs}} = 20,778$ composers observed, with a similar coverage of $0.630$ $(\pm 0.01)$. For the 100 largest institutions, in turn, with $S_{\text{obs}} = 34,090$, we obtain $S = 53,561$ $(-693.2, +715.8)$ and coverage 0.635 $(\pm 0.01)$. We interpret this to indicate as sampling error: the combined largest music libraries are still not sampling the same space of composers with extant works as all the holding institutions, including the smaller ones. Otherwise we should see a similar estimate of total composers around 80,000 as for the complete dataset, with the corresponding coverage upper bound of approx. 0.26. This implies that smaller institutions play an important role in documenting the overall diversity of composers. Only when we aggregate data over the top 600 institutions do we get the Chao1 estimate of at least 70,000 composers.

We compute the TTR and Chao1 coverage $S_{\text{obs}}/S$ for each RISM institutions, and from these value pairs we measure how these metrics are related. While there is some relationship between TTR and Chao1 coverage, it is weak, and has a very high variance. For the 100 largest institutions, Pearson's $r = -0.46$ and non-correlation can be rejected ($p$-value for non-correlation using `scipy.stats.linreg`: $< 10^{-5}$), and the same holds for all institutions, though the relationship is even weaker ($r = -0.295, p < 10^{-30}$). The relationship between TTR and Chao1 is shown in Figure 1.

### *3.1.2 Cantus: "biodiversity" of Gregorian chant*

The Cantus database[10] is a large-scale project for cataloguing Gregorian chant that has been running since the mid-1980s [26; 27]. It primary mechanism is the Cantus ID, which identifies instances of the same chant — element of Gregorian repertoire — across multiple manuscripts. Cataloguing a manuscript means primarily assigning Cantus IDs to all chants recorded therein. Chant manuscripts often have more than 1,500 chants, so cataloguing even a single manuscript requires considerable effort.

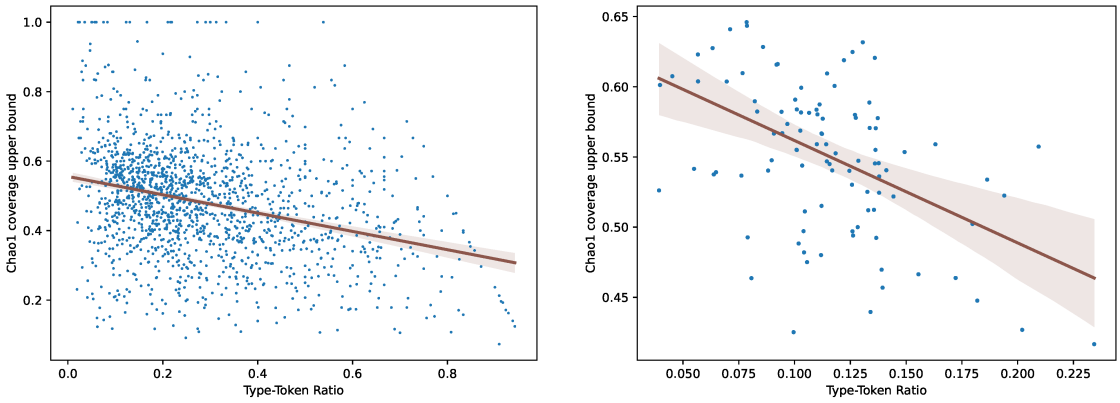
Gregorian chant was — is — an immense tradition. Despite the initial Carolingian project of Gregorian chant as a tightly controlled expression of a common identity [21][p.514–523], over the centuries and across Latin Europe, repertoire choices diversified greatly, to the extent that one can often identify the provenance of a manuscript directly from the repertoire choices. This combination of diversity and scale permits one to think of the Gregorian tradition in terms of ecology, and ask: to what extent has the Cantus database covered the existing "biodiversity" of chant?

[7] In the usual sense of the term: persons who are identified as authors of written musical works. We do not consider phenomena such as recording folk musics, or rather: we accept editorial decisions made by those who catalogued records in RISM.

[8] This includes reprints, but a reprint is in fact a valid sign of the underlying "abundance" of a composer.

[9] All confidence intervals in this paper are computed using bootstrap with 1000 iterations, as implemented in the `copia` library.

[10] `https://cantusdatabase.org/`




**(a)** Type-Token Ratio (TTR) per institution in RISM plotted against the Chao1 coverage upper bound. Linear regression shows a best fit at slope $-0.25$, with the p-value for non-correlation (zero slope) below $10^{-30}$, but the variance of the linear estimate is large.

**(b)** Type-Token Ratio (TTR) in relation to Chao1 coverage upper bound on the 100 largest institutions in RISM, to counteract the effect small data has on biasing TTR higher than in the sampled population. In this case, linear regression shows a best fit at slope $-0.69$, with the p-value for non-correlation below $10^{-5}$.

**Figure 1:** RISM results for the relationship between the Chao1 coverage upper bound and linear proxies for diversity: the Type-Token Ratio (TTR). Note that the upper right triangle plot (a) is empty: this is because at very high TTRs, the Chao1 coverage cannot be very high: even with the most uniform distribution possible, at TTR > 0.5 there will always be at least one singleton contributing to $f_1$, so coverage upper bound cannot be 1.0, and at TTR close to 1.0, nearly all tokens will contribute to $f_1$ and coverage upper bound will thus approach 0 (lower right corner). However, the correlation between TTR and Chao1 coverage is not caused by this: when we restrict ourselves to institutions where TTR < 0.6 and coverage < 0.8, where empirically this effect does not reach, we still get Pearson's $r$ of $-0.21$ and $p < 10^{-15}$.

In this abstraction, the Cantus IDs are species, and manuscripts serve as samples. We ask: how much of chant repertoire has been catalogued, and how much remains to be discovered? Chant repertoire is categorised according to *genre*, its function in liturgy. For example: antiphons are short and simple chants that are sung before and after psalms; responsories are longer and more ornate chants that are sung between blocks of psalm-antiphon pairs. Individual chant genres had a varied history as liturgy developed: for instance, the Offertory verses (a genre sung in Mass) fell out of use after the 13th century [21, p.121]. It therefore makes sense to quantify the (in)completeness of the Cantus database according to the individual main genres.

In this case study, we apply an incidence-based approach over abundance. Instead of counting how many times each Cantus ID appeared in the dataset, we count its presence or absence in manuscripts. A chant recorded in only one source, even if used twice or more times, is still considered a singleton and contributes to $f_1$. Preferring incidence follows naturally from the structure of chant data. Each manuscript acts as a sample from one site: a particular ecclesiastical community. A chant being used in more than one liturgical position in a certain church should not necessarily imply the particular chant would be more likely to be used in other churches. We use CantusCorpus v0.2 [5], a dataset derived from the Cantus database that is most widely used for computational chant research [6; 17; 28; 29].

**Cantus results.** We report the Chao1 upper bounds on coverage for individual genres on CantusCorpus v2.0 in Table 1. The genres of chant for one type of liturgy, the Divine Office (upper section of Table 1), exhibit overall lower maximum coverage than the chants for Mass (lower sec-

| Genre | CIDs | Mss. | Tokens | TTR | STR | f1 | f2 | Cov. | Conf. Int. |
|---|---|---|---|---|---|---|---|---|---|
| **A** | 11157 | 230 | 202688 | 0.055 | 0.021 | 4714 | 1542 | 0.569 | (-0.01, +0.01) |
| **R** | 5098 | 211 | 101353 | 0.050 | 0.041 | 2151 | 714 | 0.553 | (-0.02, +0.02) |
| **V** | 8162 | 213 | 93708 | 0.087 | 0.026 | 3919 | 1068 | 0.502 | (-0.02, +0.02) |
| **W** | 925 | 184 | 34983 | 0.026 | 0.199 | 292 | 127 | 0.679 | (-0.05, +0.05) |
| **I** | 599 | 180 | 9803 | 0.061 | 0.301 | 250 | 106 | 0.595 | (-0.07, +0.06) |
| ***Office*** | ***25804*** | ***240*** | ***442535*** | ***0.058*** | ***0.009*** | ***11188*** | ***3558*** | ***0.555*** | ***(-0.01, +0.01)*** |
| **In** | 206 | 49 | 1930 | 0.107 | 0.238 | 41 | 5 | 0.572 | (-0.17, +0.13) |
| **InV** | 285 | 32 | 1153 | 0.247 | 0.112 | 82 | 32 | 0.745 | (-0.09, +0.08) |
| **Gr** | 153 | 90 | 2087 | 0.073 | 0.588 | 28 | 9 | 0.731 | (-0.19, +0.15) |
| **GrV** | 206 | 68 | 1438 | 0.143 | 0.330 | 53 | 11 | 0.664 | (-0.11, +0.11) |
| **Al** | 404 | 71 | 2016 | 0.200 | 0.176 | 159 | 62 | 0.624 | (-0.07, +0.07) |
| **AlV** | 37 | 28 | 116 | 0.319 | 0.757 | 24 | 3 | 0.193 | (-0.12, +0.29) |
| **Of** | 157 | 42 | 1844 | 0.085 | 0.268 | 25 | 17 | 0.811 | (-0.14, +0.12) |
| **OfV** | 262 | 12 | 707 | 0.371 | 0.046 | 44 | 39 | 0.902 | (-0.06, +0.07) |
| **Cm** | 197 | 42 | 2059 | 0.096 | 0.213 | 27 | 8 | 0.729 | (-0.15, +0.11) |
| **CmV** | 153 | 4 | 173 | 0.884 | 0.026 | 135 | 16 | 0.172 | (-0.06, +0.07) |
| **Tc** | 46 | 21 | 272 | 0.169 | 0.457 | 10 | 6 | 0.792 | (-0.30, +0.21) |
| **TcV** | 202 | 21 | 822 | 0.246 | 0.104 | 44 | 27 | 0.844 | (-0.09, +0.08) |
| ***Mass Pr.*** | ***2267*** | ***113*** | ***14617*** | ***0.155*** | ***0.050*** | ***634*** | ***230*** | ***0.694*** | ***(-0.03, +0.03)*** |
| ***All*** | ***28056*** | ***261*** | ***457152*** | ***0.061*** | ***0.009*** | ***11809*** | ***3785*** | ***0.565*** | ***(-0.01, +0.01)*** |

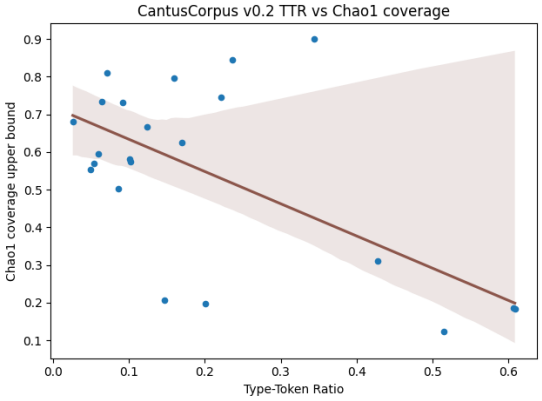
**Table 1:** Unseen species estimates for CantusCorpus v0.2 data, split by genre. We report the number of distinct chants (Cantus IDs) for each genre, the number of manuscripts (because we are using incidence data), the number of tokens (in this case: total number of chants catalogued), the Type-Token Ratio (which is computed from chant counts), additionally the Sample-Type Ratio, which is an analogy for TTR for incidence data (ratio of the sample count to species count), the singleton and doubleton counts used for Chao1 estimation, and the resulting Chao1 upper bound on coverage and the left and right widths of its 95% confidence interval. Note how the coverage varies between genres (especially those for the Mass Propers).

tion of Table 1) This is quantitative confirmation that Mass repertoire was more stable, possibly because the Mass is a public liturgy while the Divine Office is primarily a private prayer for the clergy, and thus changing Office repertoire may have been easier (though still with bureaucracy involved [14]), though interestingly the Introit genre, which starts the Mass, seems to be covered less well.

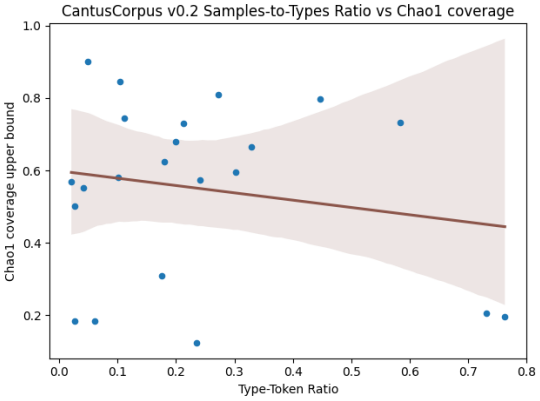
Neither the Type-Token Ratio, nor its incidence-based analogy Sample-Type Ratio, are good predictors of the Chao1 coverage upper bound. Linear regression on TTR vs. Chao1 coverage has a slope of $-0.77$ and non-correlation can just about be rejected ($p = 0.005$), but the variance is large (see Figure 2a); for STR, non-correlation cannot be rejected ($p = 0.42$; see Figure 2b).

### 3.2 Case Study 2: Ontology of differences in music prints

In this case study, we examine visual/notational differences between six editions of Beethoven's Bagatelles Op. 33, Nos. 1—5. The editions we used are the first print by Bureau d'Arts [sic] et d'Industrie (c. 1803), Zulehner (c. 1808), André (c. 1825), Schott (c. 1826), Haslinger (c.




**(a)** Type-Token Ratio (TTR) per chant genre on CantusCorpus v0.2 plotted against the Chao1 coverage upper bound. Linear regression shows a best fit at slope $-0.78$, with the $p$-value for non-correlation (zero slope) at $0.005$, but the variance of the linear estimate is large.




**(b)** Sample-Type Ratio, an analogy of TTR for incidence-based data. In this case, linear regression shows a best fit at slope $-0.20$, with the $p$-value for non-correlation at $0.42$.

**Figure 2:** CantusCorpus results for the relationship between the Chao1 coverage upper bound and linear proxies for diversity: the Type-Token Ratio (TTR), and its analogy for incidence data, the Sample-Type Ratio (STR). While the TTR has a correlation of $\rho = -0.77$ and thus non-correlation can be rejected ($p = 0.005$), predicting the Chao1 coverage upper bound still has a very large variance. STR is not correlated at all ($p = 0.42$).

1845) and Breitkopf & Härtel (1864).[11] The data for this analysis consists of files containing the results of comparisons of different MEI encodings of the six editions. Through comparisons using the Python tool `musicdiff` [10],[12] we obtained the differences between each pair of encodings[13] from which we extract the kinds and numbers of differences that occur. The Bagatelles contain a total of $38,785$ differences, summed across the six editions of each of the seven Bagatelles (for a total of $15 \times 7 = 105$ pairwise comparisons). There are $81$ different types of differences.

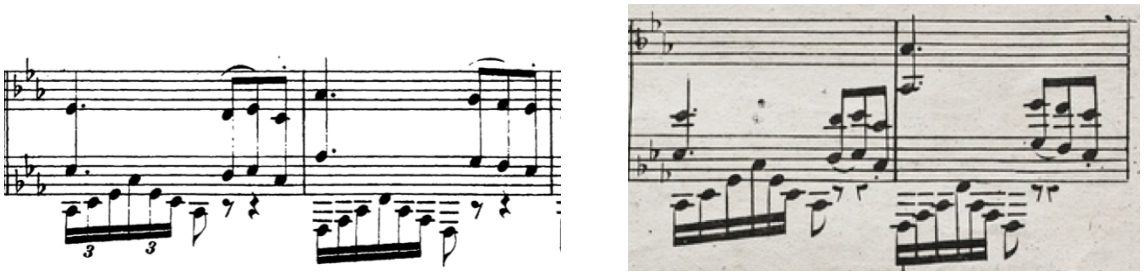
Some types differences are illustrated in Figure 3a. It shows bars 25–26 from the 5th Bagatelle in the editions of Breitkopf & Härtel and Schott. In the edition on the left (Breitkopf) the melody in the right hand is split across staffs. On the right, the same melody is printed in the lower staff (with one exception). There are also less obvious differences: the numbers to indicate triplets in the left hand in the first bar of the Breitkopf edition are omitted by Schott, slurs placed above two notes in the Breitkopf edition are below in the Schott edition, and symbols of quarter rest are different.

For most of the transmission history of these works—and music in general—, prints played a crucial role, and they heavily influenced the way the broad interested public got to know them [30]. The process of copying music throughout this history affects the musical text by intentionally or accidentally introducing variants [13]. Some printed notational variants do not affect the performance, like the one shown in Figure 3a. Despite the importance of prints for the historical reception of music, in musicological research, they are often regarded as less important than other sources like manuscripts, though the reception of music can heavily influence our own perception of these historical works today [37].

[11] All the editions can be found online at the Beethoven Haus Bonn (`https://tinyurl.com/BeethovenhausOp33`) except for the Breitkopf edition which can be found via the Petrucci Music Library (`https://imslp.org/wiki/7_Bagatelles,_Op.33_(Beethoven,_Ludwig_van)`.

[12] `https://github.com/gregchapman-dev/musicdiff.git`

[13] Find the encodings here: `https://github.com/CorpusBeethoviensis/beethoven-diff-docker.git`



**(a)** Edition by Breitkopf & Härtel. **(b)** Edition by Schott.

**Figure 3:** Example for differences between the editions of the 5th Bagatelle, bars 25 and 26, by Breitkopf & Härtel and Schott. They differ in the placement of the right hand melody, of the articulations (slur and staccato) of this melody and the numbers to indicate triplets. Also, different symbols for quarter rests are used.

The species in this case study are the types of differences found in comparing all pairs of editions of each Bagatelle. The question of unseen species in this case study is: how complete is this set of differences? Compared to the previous case study, here the unseen species problem is not the representativeness of a sample of material, but a quantitative introspection of a constructed ontology. The implications of high coverage in such a context would be that very few new categories are likely missed, and therefore the given ontology can potentially be applied to a larger corpus as-is (e.g., via a machine learning model).

**Results.** The Chao1 estimate for the combined Bagatelles data is $S = 85\ (-9.3, +26.5)$. With $S_{obs} = 81$, that means that the coverage of the differences ontology is nearly $0.947(-0.22, +0.12)$.[14] This is an upper bound, so the true coverage may be lower, but it is unlikely that the ontology still has significant blind spots, though the lower bound based on the CI does communicate some risk.

How early could we estimate how many categories we *should* first find before having a good chance of a near-complete ontology? We run the estimation with 1000 sub-samples (without replacement) of different sizes $k$ and measure the average $S$ at a given $k$. At $k = 1000$ selected out of the 38,785 differences, average $S$ is underestimated to be 67, with 50 categories observed; at $k = 5000$, $S = 76$ with average $S_{obs} = 66$, and at $k = 10,000$, somewhat above 25 % of the total differences, we obtain $S = 80$, very close to the true $S_{obs} = 81$ categories.[15]

If one estimates $S$ from all pairs of a single Bagatelle's editions, the estimates converge very quickly to the true number of difference categories for that particular Bagatelle. Using just 10 % of the total differences from each Bagatelle's edition pairs, Chao1 underestimates the true number of categories by only $5.2\%$ on average. However, the $S_{obs}$ for each complete Bagatelle never reaches more than $54$, so using a single Bagatelle to estimate the total $S$ is never going to enable reaching the true diversity of distinct editorial differences. This is expected, as each Bagatelle contains specific musical material that uses only a subset of possible music notation patterns, and therefore certain types of editorial differences do not have a chance to appear (e.g., explicit vs. implicit triplets in a composition with no triplets). This result illustrates how sampling assumptions of Chao1 estimators might be violated (each Bagatelle represents a distinct population of editorial differences, and the result should not be expected to hold for music that uses a different subset of notation than the Bagatelles), but conversely also how well the estimators work when its sampling assumptions hold, and emphasizes the value of diverse rather than large samples.

[14] The upper bound of the coverage derived from the CI on $S$ is over 100 % with respect to the true because of boundary effects in the bootstrap procedure when $S_obs$ is very close to $S$.

[15] A more principled projection would use accumulation curves or rarefaction-extrapolation curves; as this paper focuses on the breadth of applications of USMs rather than depth of methods, we leave these curves for future work.

| Genre | Types | Tokens | TTR | $f_1$ | $f_2$ | Coverage | Conf. Int. |
|---|---|---|---|---|---|---|---|
| March | 390 | 4212 | 0.093 | 110 | 63 | 0.802 | (-0.05, +0.04) |
| Slide | 269 | 5318 | 0.051 | 72 | 36 | 0.789 | (-0.07, +0.05) |
| Slip Jig | 430 | 10351 | 0.042 | 126 | 69 | 0.789 | (-0.06, +0.04) |
| Barndance | 329 | 2698 | 0.122 | 114 | 67 | 0.772 | (-0.07, +0.06) |
| Reel | 4272 | 104131 | 0.041 | 1192 | 558 | 0.770 | (-0.02, +0.01) |
| Polka | 835 | 11857 | 0.070 | 271 | 145 | 0.767 | (-0.04, +0.04) |
| Three-Two | 101 | 516 | 0.196 | 34 | 17 | 0.748 | (-0.14, +0.09) |
| Waltz | 922 | 8104 | 0.114 | 329 | 166 | 0.739 | (-0.04, +0.04) |
| Jig | 2896 | 70826 | 0.041 | 931 | 421 | 0.738 | (-0.02, +0.02) |
| Mazurka | 109 | 888 | 0.123 | 48 | 12 | 0.532 | (-0.15, +0.10) |
| **Total** | **11663** | **234330** | **0.050** | **3573** | **1747** | **0.761** | **(-0.01, +0.01)** |

**Table 2:** Repertoire coverage in different Irish folk genres represented in *The Session* dataset. Pearson correlation of coverage and type-token ratio (TTR); $\rho = .28$ ($p = .35$).

### 3.3 Case Study 3: Folk Music Sessions

Musicians all over the world gather regularly to perform traditional Irish music [45]. The platform *The Session* tracks many of these meetings, and moreover hosts a rich database of tunes that its users have added, including melodies of Irish tunes and their genres, such as Reel, Jig, Polka, Waltz, etc. Exports of the site's database are publicly available on GitHub.[16] It has been shown that population size is an important factor for melodic variety [44]. Specifically, while popular tunes recorded in the Sessions data set show higher variation of melodic complexity in their different settings, popularity is also strongly related to *intermediate* complexity of tunes. Given this mainly performer-centered view, the tunes themselves have received somewhat less attention.

The question that USMs can answer for this scenario is: given that sessions will continue to take place all over the world and that people record what was played on the website, how many tunes are likely to still be 'out there,' either in an almost forgotten tunebook, or in someone's mind? Given the partition of this tradition into relatively well-defined genres, we can also ask whether there are between-genre differences regarding the coverage of the repertoire.

Table 2 shows an overview of all genres in the Session dataset and the numbers of pieces they contain, sorted according to the coverage estimated with the Chao1 estimator. Marches have highest coverage of approx. 80.2%., whereas the Mazurka coverage is lowest, at approx. 53.2% — not too surprising, given that Mazurkas are originally a Polish dance form. Taken together, the tunes recorded in the entire Session dataset are estimated to cover about 76.1% of the entire Irish tunes repertoire. We can interpret this as showing that *The Session* project is successful at documenting and representing the living tradition of Irish music sessions.

### 3.4 Case Study 4: Harmonic vocabularies

In this case study, we use for the first time an estimator derived in the context of the Unseen Species problem for the question of the overall size of the harmonic vocabulary of Western tonal music. For many years now, corpus studies in music theory have been gaining traction, and a variety of datasets have been created and made available for computational work. Chord vocabularies have been shown to follow both Zipf's [31; 36; 40; 47] and Heaps' laws [35; 41], but these findings have not yet been extended to estimating what's still missing from empirical distributions of chords.

[16] `https://github.com/adactio/thesession-data`

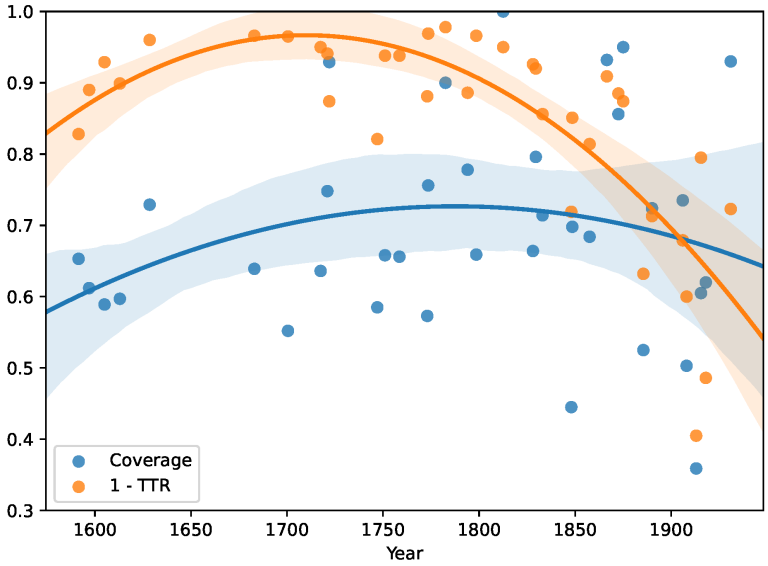



**Figure 4:** Vocabulary coverage (blue) and type-token ratio (TTR; orange) over time, with 2nd-order polynomial fit to the data points. Note that, for easier comparison, we show $1 - \text{TTR}$. Pearson correlation coefficient $\rho = .32$ ($p \approx .05$).

The recently published *Distant Listening Corpus* (DLC v2.3) [20] consists currently of 40 sub-corpora, some of which had been previously published separately [18; 19; 38]. It encompasses 1,238 score encodings by 36 composers from the extended tonal tradition (c. 1550–1945). Each piece has been analyzed by music theory experts using harmonic labels conforming to an elaborate annotation scheme based on Roman numerals.[17] In total, there are 6,015 *different* chord types (species), with a total abundance of 246,166 chords. Table 3 in Appendix A gives an overview, showing the composers' names and their birth and death dates, the numbers of chord types and tokens as well as the type-token ratio (TTR) of their works contained in the DLC. Moreover, the numbers of singletons and doubletons, and the estimated coverage based on Chao1 are shown, too. Each row shows values for a particular composer, and the last row shows these values for the aggregated corpus. Taking all DLC corpora together, the Chao1 estimator asserts that almost 70 % of the total harmonic vocabulary have been covered by this massive annotation effort, providing at the same time encouragement for its continuation.

Figure 4 compares the Chao1-based estimates of species coverage (blue) with the TTR values (orange) for each composer in the DLC.[18] In order to facilitate the visual comparison, the figure shows $1 - \text{TTR}$. For both quantities, we added a quadratic regression; error bands represent a 95% confidence interval based on bootstrap samples of the data.[19] The error for the Chao1 estimates fluctuates more because the values are more widely dispersed, especially in the second half of the timeline. The Pearson correlation between the two sets of datapoints is $\rho = .32$ ($p \approx .05$), indicating a positive but weak association, as expected.

The interpretation of TTR and Chao1 in this context is not straightforward. TTR shows the empirical fraction between the observed vocabulary size and total number of chord tokens, but the number of chord tokens depends on many non-random factors, e.g. sonatas tend to be longer than Lieder, so a higher number of tokens may stem from a composer's preference for certain genres. Moreover, while the indicated curve of the TTR over time (orange line in Figure 4) *could*

[17] `https://dcmlab.github.io/standards/`
[18] DLC sub-corpora by the same composer were merged.
[19] See `https://seaborn.pydata.org/generated/seaborn.regplot.html` for details.

be interpreted somehow to a real change of of the harmonic language over time, the curve fitted to the Chao1 estimates tells us rather something about where further encoding and annotation efforts should be directed. Apart from the general observation that many digital music corpora are heavily biased [42], the coverage of the harmonic vocabularies of composers 'at the fringes' of the represented timeline could be increased by sampling (i.e., encoding and annotating) more pieces from around that time—by the same or different composers.

## 4 General Discussion

In this study we have applied the popular Chao1 model to estimate the numbers of unseen species in a range of cultural contexts that are commonplace in musicological scholarship. We have looked at two of the largest musicological databases, RISM and Cantus, and estimated how many new composers and chants, respectively, we should still expect to encounter when continuing these cataloguing efforts. We have looked at the practice of 19th-century music prints and notational differences between them caused by to editorial intervention or pure chance, and have provided a principled answer to the question of how complete an ontology of these differences is. In the domain of music performance, we have analyzed data about Irish folk music sessions and the repertoire coverage between different sub-genres. Interestingly, across nearly all genres, coverage was higher in The Session data than it is for chant genres in Cantus, possibly indicating the strength of crowdsourcing by practitioners compared to expert efforts — or raising questions about how restrictive the ecclesiastical regulatory framework for chant in fact may have been compared to a tradition with less defined boundaries. Finally, addressing music theory, we have looked at the size of harmonic vocabularies of a great number of composers from the Renaissance onward.

What have we learned from all of this? First of all: recognizing structural similarities between vastly different fields enables the transfer of methods and can lead to opening up entirely new avenues of research. Second, using the Chao1 estimators is simple: it involves only combining two easily computable quantities, the numbers of singletons and doubletons. This methods is accessible without training in formal methods. Third, Unseen Species models can be useful proxies in assessing whether and where usually scarce resources should be put to use. Their estimates can differ significantly from other measures of dataset diversity, as we illustrate by comparing Chao1 coverage upper bounds and TTR. One must be careful in how exactly these models are applied: for instance, the population of interest is assumed to be sampled with replacement, which is not a safe approximation if the sample size approaches an appreciable fraction of the total population [46] (where Chao1 would over-estimate the species richness lower bound). The assumption that one is homogeneously sampling a single population also may not hold.

## 5 Conclusions

Our main goal in applying Unseen Species models in these case studies was to demonstrate that they can be useful additions to the methodological repertoire of computational musicologists. Surely, it will not be hard to think of other areas where one could use this model.We encourage our colleagues to engage with this kind of modeling in their own domains. However, we emphasise that the estimator relies on specific assumptions that may not hold for all scenarios, and caution has to be applied regarding the validity of the conclusions to be drawn. The coverage percentages estimated in this article may in reality lie far from the true (but possibly unknowable) achieved coverage. The strength of the methodology, however, lies in the fact that it yields a upper bound for this quantity: there is at least that much to discover. In the end, our work is meant as an invitation to constructive criticism, enabled by the explicit nature of the approach. Computational modeling and critical thinking are not opposed (as sometimes suggested), but rather are the same thing in different disguise.

**Another tip.** In some cases, it may be helpful to use `paragraph` to title individual paragraphs. For example, if a section describes features for a classifier, you can optionally title each paragraph with the name of each feature.

## Acknowledgements

We thank an anonymous reviewer for the detailed feedback that helped us further improve the quality of our contribution. This work was supported by the Social Sciences and Humanities Research Council of Canada by the grant no. 895-2023-1002, Digital Analysis of Chant Transmission, and the project "Human-centred AI for a Sustainable and Adaptive Society" (reg. no.: CZ.02.01.01/00/23_025/0008691), co-funded by the European Union.

## References

[1] Chao, Anne. "Nonparametric Estimation of the Number of Classes in a Population". In: *Scandinavian Journal of Statistics 11*, no. 4 (1984), pp. 265–270.

[2] Chao, Anne. "Species estimation and applications". In: *Encyclopedia of Statistical Sciences 12* (2004).

[3] Chao, Anne, Chiu, Chun-Huo, Colwell, Robert K., Magnago, Luiz Fernando S., Chazdon, Robin L., and Gotelli, Nicholas J. "Deciphering the enigma of undetected species, phylogenetic, and functional diversity based on Good-Turing theory". In: *Ecology 98*, no. 11 (Nov. 2017), pp. 2914–2929. ISSN: 1939-9170. DOI: 10.1002/ecy.2000. URL: http://dx.doi.org/10.1002/ecy.2000.

[4] Clauset, Aaron., Shalizi, Cosma Rohilla., and Newman, M. E. J. "Power-Law Distributions in Empirical Data". In: *SIAM Review 51*, no. 4 (Nov. 2009), pp. 661–703. DOI: 10.1137/070710111.

[5] Cornelissen, Bas, Zuidema, Willem, and Burgoyne, John Ashley. "Studying large plainchant corpora using chant21". In: *7th International Conference on Digital Libraries for Musicology*. 2020, pp. 40–44.

[6] Cornelissen, Bas, Zuidema, Willem H, Burgoyne, John Ashley, et al. "Mode Classification and Natural Units in Plainchant". In: *Proceedings of the 21st Int. Society for Music Information Retrieval Conf.* Montreal, Canada, 2020, pp. 869–875.

[7] Cuthbert, Michael Scott. "Tipping the Iceberg: Missing Italian Polyphony from the Age of Schism". In: *Musica Disciplina 54* (2009), pp. 39–74. URL: http://www.jstor.org/stable/25750547.

[8] Deffner, Dominik, Fedorova, Natalia, Andrews, Jeffrey, and McElreath, Richard. "Bridging Theory and Data: A Computational Workflow for Cultural Evolution". In: *Proceedings of the National Academy of Sciences 121*, no. 48 (Nov. 2024), e2322887121. DOI: 10.1073/pnas.2322887121.

[9] Efron, Bradley and Thisted, Ronald. "Estimating the Number of Unseen Species: How Many Words Did Shakespeare Know?" In: *Biometrika 63*, no. 3 (Dec. 1976), pp. 435–447. DOI: 10.1093/biomet/63.3.435.

[10] Foscarin, Francesco, Fournier-S'Niehotta, Raphaël, and Jacquemard, Florent. "A diff procedure for music score files". In: *Computation and visualization of the differences between two music score files. 6th International Conference on Digital Libraries for Musicology (DLfM)*. Den Haag, November 2019, pp. 7–13. URL: https://inria.hal.science/hal-02267454v2.

[11] Gale, William A and Sampson, Geoffrey. “Good-turing frequency estimation without tears”. In: *Journal of quantitative linguistics 2*, no. 3 (1995), pp. 217–237.

[12] Good, Irving John. “The Population Frequencies of Species and the Estimation of Population Parameters”. In: *Biometrika 40*, no. 3–4 (1953), pp. 237–264. DOI: 10.1093/biomet/40.3-4.237.

[13] Grier, James. “Musical Sources and Stemmatic Filiation: A Tool for Editing Music”. In: *The Journal of Musicology 13*, no. 1 (1995), pp. 73–102.

[14] Hallas, Rhianydd. *Two rhymed offices composed for the feast of the Visitation of the Blessed Virgin Mary: comparative study and critical edition*. Bangor University (United Kingdom), 2021.

[15] Hao, Yi and Li, Ping. “Optimal Prediction of the Number of Unseen Species with Multiplicity”. In: *Advances in Neural Information Processing Systems*, ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 8553–8564.

[16] Helsen, Kate, Bain, Jennifer, Fujinaga, Ichiro, Hankinson, Andrew, and Lacoste, Debra. “Optical music recognition and manuscript chant sources”. In: *Early Music 42*, no. 4 (Oct. 2014), pp. 555–558. DOI: 10.1093/em/cau092.

[17] Helsen, Kate, Daley, Mark, and Schindler, Jake. “The Sticky Riff: Quantifying the Melodic Identities of Medieval Modes”. In: *Empirical Musicology Review 16*, no. 2 (2021), pp. 312–325.

[18] Hentschel, Johannes, Neuwirth, Markus, and Rohrmeier, Martin. “The Annotated Mozart Sonatas: Score, Harmony, and Cadence”. In: *Transactions of the International Society for Music Information Retrieval 4*, no. 1 (2021), pp. 67–80. DOI: 10.5334/tismir.63.

[19] Hentschel, Johannes, Rammos, Yannis, Moss, Fabian C., Neuwirth, Markus, and Rohrmeier, Martin. “An Annotated Corpus of Tonal Piano Music from the Long 19th Century”. In: *Empirical Musicology Review 18*, no. 1 (2023), pp. 84–95. DOI: 10.18061/emr.v18i1.8903.

[20] Hentschel, Johannes, Rammos, Yannis, Neuwirth, Markus, and Rohrmeier, Martin. “A Corpus and a Modular Infrastructure for the Empirical Study of (an)Notated Music”. In: *Scientific Data 12*, no. 1 (Apr. 2025), p. 685. DOI: 10.1038/s41597-025-04976-z.

[21] Hiley, David. *Western Plainchant: a Handbook*. Oxford, United Kingdom: Clarendon Press, 1993.

[22] Karsdorp, Folgert, Kestemont, Mike, and Koster, Margo de. “Beyond the Register: Demographic Modeling of Arrest Patterns in 1879-1880 Brussels”. In: *Proceedings of the Computational Humanities Research Conference*, ed. by Wouter Haverals, Marijn Koolen, and Laure Thompson. Vol. 3834. CEUR Workshop Proceedings. 2024, pp. 265–281.

[23] Karsdorp, Folgert, Manjavacas, Enrique, and Fonteyn, Lauren. “Introducing Functional Diversity: A Novel Approach to Lexical Diversity in (Historical) Corpora”. In: *Proceedings of the Computational Humanities Research Conference, 2022*. Antwerp: CEUR-WS, 2022, pp. 114–126.

[24] Kestemont, Mike, Karsdorp, Folgert, Bruijn, Elisabeth de, Driscoll, Matthew, Kapitan, Katarzyna A., Macháin, Pádraig Ó, Sawyer, Daniel, Sleiderink, Remco, and Chao, Anne. “Forgotten books: The application of unseen species models to the survival of culture”. In: *Science 375*, no. 6582 (2022), pp. 765–769. DOI: 10.1126/science.abl7655.

[25] Koeser, Rebecca Sutton and LeBlanc, Zoe. “Missing Data, Speculative Reading”. In: *Journal of Cultural Analytics 9*, no. 2 (May 2024). DOI: 10.22148/001c.116926.

[26] Lacoste, Debra. “The Cantus Database and Cantus Index Network”. In: *The Oxford Handbook of Music and Corpus Studies*. Oxford University Press, 2022. ISBN: 9780190945442. DOI: 10.1093/oxfordhb/9780190945442.013.18. URL: https://doi.org/10.1093/oxfordhb/9780190945442.013.18.

[27] Lacoste, Debra. “The cantus database: Mining for medieval chant traditions”. In: *Digital Medievalist 7* (2012).

[28] Lanz, Vojtěch and Hajič, Jan. “Text boundaries do not provide a better segmentation of Gregorian antiphons”. In: *Proceedings of the 10th International Conference on Digital Libraries for Musicology*. 2023, pp. 72–76.

[29] Lanz, Vojtěch and Hajič jr., Jan. “Gregorian Melody, Modality, and Memory: Segmenting chant with Bayesian nonparametrics”. 2025. DOI: 10.48550/ARXIV.2507.00380.

[30] Lewis, David and Page, Kevin. “Popular musical arrangements in the nineteenth-century home: A study of The Harmonicon supported by digital tools”. In: *Proceedings of the 11th International Conference on Digital Libraries for Musicology*. DLfM ’24. Stellenbosch, South Africa: Association for Computing Machinery, 2024, pp. 32–39. DOI: 10.1145/3660570.3660575.

[31] Manaris, Bill, Romero, Juan, Machado, Penousal, Krehbiel, Dwight, Hirzel, Timothy, Pharr, Walter, and Davis, Robert B. “Zipf’s Law, Music Classification, and Aesthetics”. In: *Computer Music Journal 29*, no. 1 (2005), pp. 55–69. ISSN: 01489267. DOI: 10.1162/comj.2005.29.1.55.

[32] Manning, Christopher D and Schütze, Hinrich. *Foundations of Statistical Natural Language Processing*. 6th ed. MIT Press, 2003.

[33] Martynenko, Antonina. “Unread, yet preserved: A case study on survival of the 19th-century printed poetry”. In: *Literatura: teoría, historia, crítica 25*, no. 2 (July 2023). DOI: 10.15446/lthc.v25n2.108775.

[34] Milička, Jiři. “Rank-Frequency Relation & Type-token Relation: Two Sides of the Same Coin”. In: *Methods and Applications of Quantitative Linguistics – Selected Papers of the 8th International Conference on Quantitative Linguistics (QUALICO) in Belgrade, Serbia, April 26-29, 2012*. April. 2012, pp. 163–171.

[35] Moss, Fabian C. *Transitions of Tonality: A Model-Based Corpus Study*. PhD thesis. Lausanne, Switzerland: École Polytechnique Fédérale de Lausanne, 2019. DOI: 10.5075/epfl-thesis-9808.

[36] Moss, Fabian C., Neuwirth, Markus, Harasim, Daniel, and Rohrmeier, Martin. “Statistical Characteristics of Tonal Harmony: A Corpus Study of Beethoven’s String Quartets”. In: *PLoS ONE 14*, no. 6 (2019), e0217242. DOI: 10.1371/journal.pone.0217242.

[37] Nachtwey, Adrian and Moss, Fabian C. “Big Data = Großes Wissen? Herausforderungen der digital-vergleichenden Musikforschung”. In: *kontrovers. Debatten zur Musikwissenschaft* (2024). online. DOI: https://doi.org/10.58079/12zf8.

[38] Neuwirth, Markus, Harasim, Daniel, Moss, Fabian C., and Rohrmeier, Martin. “The Annotated Beethoven Corpus (ABC): A Dataset of Harmonic Analyses of All Beethoven String Quartets”. In: *Frontiers in Digital Humanities 5*, no. July (2018), pp. 1–5. DOI: 10.3389/fdigh.2018.00016.

[39] Orlitsky, Alon, Suresh, Ananda Theertha, and Wu, Yihong. “Optimal prediction of the number of unseen species”. In: *Proceedings of the National Academy of Sciences 113*, no. 47 (Nov. 2016), pp. 13283–13288. DOI: 10.1073/pnas.1607774113.

[40] Perotti, Juan I. and Billoni, Orlando V. “On the Emergence of Zipf ’s Law in Music”. In: *Physica A: Statistical Mechanics and its Applications* (Feb. 2020), p. 124309. DOI: 10.1016/j.physa.2020.124309.

[41] Serra-Peralta, Marc, Serrà, Joan, and Corral, Álvaro. “Heaps’ Law and Vocabulary Richness in the History of Classical Music Harmony”. In: *EPJ Data Science 10*, no. 1 (Dec. 2021), p. 40. DOI: 10.1140/epjds/s13688-021-00293-8.

[42] Shea, Nicholas, Reymore, Lindsey, White, Christopher Wm, VanHandel, Leigh, Duinker, Ben, Zeller, Matthew, and Biamonte, Nicole. “Diversity in Music Corpus Studies”. In: *Music Theory Online 30*, no. 1 (Feb. 2024). URL: https://mtosmt.org/issues/mto.24.30.1/mto.24.30.1.shea_et_al.html.

[43] Smith, Eric P. and Belle, Gerald van. “Nonparametric Estimation of Species Richness”. In: *Biometrics 40*, no. 1 (Mar. 1984), p. 119. DOI: 10.2307/2530750.

[44] Street, Sally E., Eerola, Tuomas, and Kendal, Jeremy. “The Role of Population Size in Folk Tune Complexity”. In: *Humanities and Social Sciences Communications 9*, no. 152 (2022), pp. 1–12. DOI: 10.1057/s41599-022-01139-y.

[45] Tolmie, Peter, Benford, Steve, and Rouncefield, Mark. “Playing in Irish Music Sessions”. In: *Ethnomethodology at Play*. Routledge, 2013.

[46] Wevers, Melvin, Karsdorp, Folgert, and Lottum, Jelle van. “What Shall We Do With the Unseen Sailor? Estimating the Size of the Dutch East India Company Using an Unseen Species Model”. In: *Proceedings of the Computational Humanities Research Conference 2022, CHR 2022, Antwerp, Belgium, December 12-14, 2022*, ed. by Folgert Karsdorp and Kristoffer L. Nielbo. Vol. 3290. CEUR Workshop Proceedings. 2022, pp. 189–197. URL: https://ceur-ws.org/Vol-3290/short%5C_paper1793.pdf.

[47] Zanette, Damián H. “Zipf’s Law and the Creation of Musical Context”. In: *Musicae Scientiae 10*, no. 1 (2006), pp. 3–18.

# A Tables

| Composer | Types | Tokens | TTR | $f_1$ | $f_2$ | Coverage | Conf. Int. |
|---|---|---|---|---|---|---|---|
| Béla Bartók (1881–1945) | 709 | 1191 | 0.405 | 513 | 104 | 0.359 | (-0.06, +0.05) |
| Erwin Schulhoff (1894–1942) | 251 | 488 | 0.486 | 137 | 61 | 0.620 | (-0.08, +0.07) |
| Sergei Rachmaninoff (1873–1943) | 456 | 1141 | 0.600 | 280 | 87 | 0.503 | (-0.07, +0.06) |
| Gustav Mahler (1860–1911) | 219 | 595 | 0.632 | 129 | 42 | 0.525 | (-0.11, +0.09) |
| Maurice Ravel (1875–1937) | 276 | 861 | 0.679 | 113 | 64 | 0.735 | (-0.08, +0.07) |
| Claude Debussy (1862–1918) | 291 | 1013 | 0.713 | 120 | 65 | 0.724 | (-0.07, +0.07) |
| Richard Wagner (1813–1883) | 402 | 1433 | 0.719 | 224 | 50 | 0.445 | (-0.08, +0.06) |
| Francis Poulenc (1899–1963) | 77 | 278 | 0.723 | 18 | 28 | 0.930 | (-0.14, +0.09) |
| Nikolai Medtner (1880–1951) | 1332 | 6508 | 0.795 | 669 | 257 | 0.605 | (-0.04, +0.03) |
| Clara Schumann (1819–1896) | 247 | 1326 | 0.814 | 99 | 43 | 0.684 | (-0.08, +0.07) |
| Wilhelm Friedemann Bach (1710–1784) | 314 | 1753 | 0.821 | 158 | 56 | 0.585 | (-0.08, +0.07) |
| Jan Pieterszoon Sweelinck (1562–1621) | 86 | 501 | 0.828 | 37 | 15 | 0.653 | (-0.20, +0.12) |
| Franz Liszt (1811–1886) | 755 | 5070 | 0.851 | 324 | 161 | 0.698 | (-0.05, +0.04) |
| Robert Schumann (1810–1856) | 265 | 1840 | 0.856 | 105 | 52 | 0.714 | (-0.08, +0.06) |
| Edvard Grieg (1843–1907) | 1038 | 8236 | 0.874 | 193 | 340 | 0.950 | (-0.03, +0.03) |
| Georg Friedrich Händel (1685–1759) | 44 | 350 | 0.874 | 9 | 12 | 0.929 | (-0.22, +0.10) |
| Giovanni Battista Pergolesi (1710–1836) | 141 | 1189 | 0.881 | 58 | 16 | 0.573 | (-0.13, +0.09) |
| Antonín Dvořák (1841–1904) | 177 | 1539 | 0.885 | 53 | 47 | 0.856 | (-0.09, +0.07) |
| Ignaz Pleyel (1757–1831) | 179 | 1567 | 0.886 | 67 | 44 | 0.778 | (-0.10, +0.08) |
| Jacopo Peri (1561–1633) | 316 | 2884 | 0.890 | 151 | 57 | 0.612 | (-0.09, +0.07) |
| Girolamo Frescobaldi (1583–1643) | 536 | 5318 | 0.899 | 248 | 85 | 0.597 | (-0.06, +0.05) |
| Pyotr Ilyich Tchaikovsky (1840–1893) | 278 | 3059 | 0.909 | 52 | 67 | 0.932 | (-0.06, +0.04) |
| Frédéric Chopin (1810–1849) | 726 | 9125 | 0.920 | 226 | 137 | 0.796 | (-0.04, +0.03) |
| Felix Mendelssohn (1809–1847) | 1094 | 14758 | 0.926 | 448 | 181 | 0.664 | (-0.04, +0.03) |
| Claudio Monteverdi (1567–1643) | 232 | 3289 | 0.929 | 111 | 38 | 0.589 | (-0.10, +0.08) |
| Carl Philipp Emanuel Bach (1714–1788) | 698 | 11191 | 0.938 | 290 | 116 | 0.658 | (-0.05, +0.04) |
| Johann Christian Bach (1735–1782) | 314 | 5063 | 0.938 | 132 | 53 | 0.656 | (-0.07, +0.06) |
| Domenico Scarlatti (1685–1757) | 733 | 12490 | 0.941 | 275 | 153 | 0.748 | (-0.05, +0.04) |
| Franz Schubert (1797–1828) | 308 | 6200 | 0.950 | 0 | 71 | 1.000 | (-0.03, +0.02) |
| Johann Sebastian Bach (1685–1750) | 931 | 18493 | 0.950 | 390 | 143 | 0.636 | (-0.04, +0.04) |
| Heinrich Schütz (1585–1672) | 471 | 11709 | 0.960 | 161 | 74 | 0.729 | (-0.06, +0.04) |
| François Couperin (1668–1733) | 333 | 9472 | 0.965 | 147 | 40 | 0.552 | (-0.08, +0.06) |
| Arcangelo Corelli (1653–1713) | 490 | 14314 | 0.966 | 191 | 66 | 0.639 | (-0.06, +0.05) |
| Ludwig van Beethoven (1770–1827) | 1722 | 50052 | 0.966 | 732 | 301 | 0.659 | (-0.03, +0.03) |
| Wolfgang Amadeus Mozart (1756–1791) | 466 | 15272 | 0.969 | 157 | 82 | 0.756 | (-0.06, +0.05) |
| Leopold Koželuch (1747–1818) | 361 | 16598 | 0.978 | 77 | 74 | 0.900 | (-0.06, +0.04) |
| **Total** | **6015** | **246166** | **0.024** | **2488** | **1097** | **0.681** | **(-0.02, +0.02)** |

**Table 3:** Harmonic vocabularies of composers represented in the *Distant Listening Corpus*, sorted by estimated coverage.