Harmonic patterns in Brazilian choro: a data-driven comparison with Western classical and Anglophone popular music

*Mauro Orsini Windholz
Julius-Maximilians-Universität Würzburg, Würzburg, Germany
mauro.windholz@uni-wuerzburg.de
Orcid: 0009-0002-6605-8400

David R. W. Sears
University of Michigan, Ann Arbor, MI, USA
drwsears@umich.edu
Orcid: 0000-0002-3745-6621

Fabian C. Moss
Julius-Maximilians-Universität Würzburg, Würzburg, Germany
fabian.moss@uni-wuerzburg.de
Orcid: 0000-0001-9377-2066

DOI: 10.46926/musmat.2026v10.54-71

Abstract: *Data-driven research on different harmonic styles has increased substantially since the early 2000s. However, such studies have largely focused on one style at a time, while also rarely addressing styles from the Global South. Sears and Forrest (2021) present a method for comparing Western classical and Anglophone popular harmonic styles based on large corpora of Roman numeral annotations, ranking both the conventional and characteristic chord progressions of each style at formal boundaries. The present study addresses the lack of representation of styles from the Global South by reproducing Sears and Forrest's methodology with the inclusion of Brazilian choro. Our results suggest that choro harmony, even if largely tonal and tertian, exhibits unique chord progressions at formal boundaries relative to the other two styles, such as iv – $ii^{ø}$ – i, characteristic use of the "minor 2 – 5 – 1", as well as a much higher incidence of applied dominants. At the same time, certain tri-grams seem roughly equally present in all three styles, such as vi – V^7 – I, potentially revealing tonal harmonic patterns that have more cross-stylistic applicability. These results advance our knowledge of how a previously under-studied genre deploys unique vocabularies of tonal chord progressions in comparison to more studied styles, increasing our understanding of how tonal harmony develops in different parts of the world, while also contributing to diversifying corpus studies.*

Keywords*: Brazilian music. Choro. Harmony. Information theory. Music corpus studies.*

1. Introduction

In the early 21st century, computational studies of musical harmony in different styles boomed. Several studies analyzed large corpora of chord symbols within certain styles and/or composers with the goal of quantitatively describing them [8, 4, 11, 10]. These methodologies complement previous music-theoretic and analytical approaches that typically generalize style traits from a smaller selection of pieces and composers. Two main gaps emerge from this literature, however. First, few such studies have systematically compared different styles [cf. 27]. Without such comparative analyses, a clearer understanding of what distinguishes harmonic styles on quantitative grounds is not possible [16]. Second, very few such studies have been carried out including styles that do not belong to Western classical or Anglophone popular traditions. This prevents the understanding of how musical harmony is deployed in more diverse contexts, which remains a pressing concern in the music research community today [30].

This study addresses both of these gaps by presenting a comparison of the typical harmonic vocabularies at formal boundaries of Brazilian choro, Anglophone popular (hereafter popular), and Western classical (hereafter classical) music, based on the harmonic annotations from five different corpora. Choro is a popular style of (mostly) instrumental music that emerged in Rio de Janeiro around the 1870s. At its inception, choro was performed by emerging classes of industrial and government workers that played music as entertainment for gatherings at middle-to-lower class family houses [32]. Waltzes, polkas, schottisches and mazurkas were played with a particular lamenting character, the likely origin of the style's name, which means "cry" in Portuguese [14]. Some typical instruments featured in choro, then and today, are mandolin, flute, guitar, and cavaquinho. Percussion instruments such as the pandeiro, and syncopated melodies and rhythms denote more clearly the influence of Afro-Brazilian styles such as maxixe [26]. The genre ebbed and flowed in the Brazilian popular music mainstream across the decades [14], and has been "mobilized during times of pursuit of a national identity based on miscegenation and the embodiment of foreign elements as a creative basis upon which to produce Brazilianness" [3, p. 416]. As such, it is a central genre of Brazilian popular music, having been officially declared Brazilian cultural heritage in 2024 [7]. Considering choro's influence on important styles within Brazilian popular music, having been described, for example, as "the older cousin of samba" [20], understanding its unique harmonic traits may serve as a starting point for further research on some other styles of Brazilian popular harmony. What is more, Choro's harmony is overwhelmingly tertian and tonal [2], so comparing it to classical and popular styles will further our understanding of how such harmonic structures developed across diverse cultures.

In order to quantitatively compare these three harmonic styles, we draw on and extend the work of Sears and Forrest [27], who present a methodology for comparing harmonic sequences at formal boundaries between pairs of musical styles. In short, the method identifies chord sequences leading up to moments of strong formal segmentation and ranks them according to two metrics from corpus linguistics. Scaled Pointwise-Mutual Information (pMI_c) identifies those chord sequences that are most conventionally found within a single style. Relative Frequency Ratio (r) ranks the same sequences according to how characteristic they are in a style when directly

*The authors wish to thank audiences at the International Conference for Computational and Cognitive Musicology 2025, which took place in Aalborg University, and at Prof. Juniper Hill's ethnomusicology research seminar at University of Würzburg. They attended talks showing preliminary results of this work and provided excellent feedback and suggestions. We also thank Tim Eipert, Lucas Hofmann, and Adrian Nachtwey at University of Wurzburg, as well as two anonymous reviewers, who provided key feedback.

Received: December 11th, 2025
Approved: April 1st, 2026

compared to another. What is more, the method identifies formal harmonic boundaries according to the Information Content of each chord symbol, calculated using the Information Dynamics of Music (or IDyOM) model [23] (see Section 3 below for details on all of these metrics). Additionally, we extend the methodology by introducing a technique to compare the harmonic vocabularies of three styles simultaneously [18].

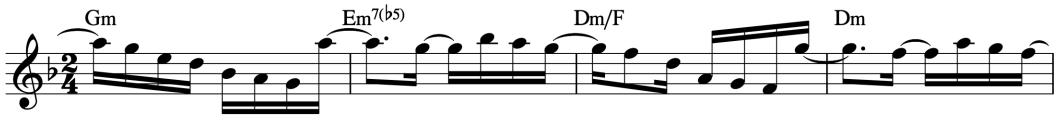
2. Data

This study uses five published corpora of harmonic annotations from the three different musical styles in question. Tonal harmony from the Western classical period is represented by the *Annotated Beethoven Corpus* (ABC) [21] and the *Theme and Variation Encodings with Roman Numerals* (TAVERN) corpus [11]. The 20th century popular harmonic idiom is represented by the *McGill Billboard Corpus* (Billboard) [5] and the *Rolling Stones Corpus* (RS200) [8]. These four corpora were also examined in [27]. Finally, harmony in Brazilian choro (1860–2000) is represented by the *Choro Songbook Corpus* (CSC) [17]. Table 1 presents some summary statistics for these datasets.

	Western Classical		**Popular**		**Choro**
	ABC	TAVERN	Billboard	RS200	CSC
no. pieces	70	27	721	200	295
no. chords	27,993	12,444	95,522	19,433	43,839
unique chords	876		701		691

Table 1: *Overview of the corpora used. The unique chords row shows vocabulary size differences.*

As each of these datasets provides chord symbols in different encoding standards, we first transformed all data into a common representation, a variant of the so-called Harte syntax [13], also employed in [27]. See Figure 1 for an example of this transformation process based on a typical choro chord sequence. Pop chord symbols appear above the staff, and the corpus and common representations appear below the staff.[1] All pre-processed data and analyses scripts can be accessed at `https://osf.io/wdbhg/overview`.



Corpus representation	IVm	IIm7(b5)	Im/3	Im
Common representation	iv	iih7	i6	i

Figure 1: *Example of transformation of chord representation for an excerpt from the piece "Murmurando" by Jacob do Bandolim, available in the CSC.*

[1] We use "h" as a shorthand for half-diminished chords in our common notation to facilitate analysis. For representing this kind of chord in tables and other figures, we use the "ø" shorthand, as is common in certain styles of chord symbol notation.

3. Method

3.1. Modeling harmonic corpora and chord sequences

We consider a corpus or dataset D of music to be a set of K pieces, $D = \{d_1, ..., d_K\}$. Each piece $d_k \in D$ is represented as a sequence of L_k chord symbols, $d_k = e_1 e_2 e_3 \cdots e_{L_k} := e_{1:L_k}$.[2] We denote a subsequence of chord symbols with length $n \leq L_k$ (a so-called n-gram) by

$$e_{i:i+n-1} := \underbrace{e_i e_{i+1} \cdots e_{i+n-1}}_{\text{length } n},$$

for $i = 1, \ldots L_k - n$. To estimate the probability of a particular event e_i (a chord token) in such a subsequence, we adopt the *Information Dynamics of Music* (IDyOM) model [22] using its R implementation [12]. IDyOM relies on the so-called Markov assumption that the probability of an event given the entire preceding context can be approximated by its relative frequency of occurrence after a shorter preceding context,

$$p(e_i \mid \underbrace{e_1 e_2 \cdots e_{i-2} e_{i-1}}_{\text{length } L_k - i}) \approx p(e_i \mid \underbrace{e_{i-n+1} e_{i-n+2} \cdots e_{i-2} e_{i-1}}_{\text{length } n-1}),$$

or, equivalently,

$$p(e_i \mid e_{1:i-1}) \approx p(e_i \mid e_{i-n+1:i-1}). \tag{1}$$

The *information content* IC of an event e_i is defined as

$$\mathrm{IC}(e_i \mid e_{i-n+1:i-1}) = -\log_2 p(e_i \mid e_{i-n+1:i-1}) \tag{2}$$

and measures the unexpectedness of event e_i given the previous context $e_{i-n+1:i-1}$. Note that, if the probability of an event is low, it is highly unexpected, and so IC is high. Conversely, if an event is highly likely, IC is low [9].

Building on earlier work [24], Sears and Forrest [27] use the information content of chords drawn from chord sequences to derive a measure of each chord's *boundary strength* on information-theoretic grounds as

$$b(e_i) = \begin{cases} \frac{1}{\mathrm{IC}(e_i)} \cdot \frac{\mathrm{IC}(e_{i+1}) - \mathrm{IC}(e_i)}{\mathrm{IC}(e_{i-1}) - \mathrm{IC}(e_i)}, & \text{if } \mathrm{IC}(e_i) < \mathrm{IC}(e_{i-1}) \wedge \mathrm{IC}(e_i) < \mathrm{IC}(e_{i+1}) \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

That is, the boundary strength of an event is defined as the ratio of differences between the information content of event e_i and that of its preceding and subsequent elements, normalized by the event's own information, but only if its information content is lower than that of the events immediately preceding and following it.[3] Following the methodology of [27], we only consider chord tri-grams ($n = 3$) in this study, but the method generalizes to all values of n.

3.2. Boundary strength, conventionality, and characteristicness

We consider a chord tri-gram $e_{i-2:i}$ to end in a strong formal boundary if its terminal member e_i has a boundary strength greater than one standard deviation above an exponentially-weighted

[2] We use the same notation for indexing unordered sets and ordered lists, and context should help disambiguate.

[3] In Equation 3, we omitted the conditionals in the expressions for better legibility, but they are meant to be implied.

moving average with a window size of 20 events. Tri-grams exceeding this threshold are included in a filtered tri-gram list. From this list, we identified *conventional* tri-grams by first estimating their respective *point-wise mutual information* (pMI_c), following [28],

$$pMI(\mathcal{T}) = \log_2 \frac{p(e_{i-2:i})}{\prod_{j=i-2}^{i} p(e_j)}. \tag{4}$$

This measure estimates the ratio of the observed probability of an n-gram type $\mathcal{T}$ consisting of events $e_{1:n}$ from the filtered n-gram list to the joint probability of it's constituent members. If all e_i were independent, it would follow per definition that $p(e_{1:n}) = \prod_i^n p(e_i)$, and thus $pMI(T) = 0$, as there would be no information shared between chords. We included an additional scaling factor c, based on [27], a coverage statistic that measures the proportion of compositions in the filtered list that contain the n-gram $\mathcal{T}$ at least once,

$$c = \frac{\mid \{d_k : \mathcal{T} \in d_k\} \mid}{\mid D \mid}, \tag{5}$$

and is used to counterbalance the known low-frequency bias of pMI:

$$pMI_c(\mathcal{T}) := c \times pMI(\mathcal{T}). \tag{6}$$

pMI_c thus represents the conventionality of a pattern with respect to a given corpus: the higher it is for a certain pattern (in our case, a tri-gram sequence of chords), the more conventional it is.

In corpus linguistics, the *relative frequency ratio* of a pattern, denoted by r, is used to assess how characteristic it is of a certain corpus when compared to a comparison corpus [15]. It is defined as the logarithm of the ratio between the probability of the pattern in one corpus and its probability in a comparison corpus (or *anti-corpus*), denoted by D'.

$$r(\mathcal{T}; D, D') = \log_2 \frac{p(e_{1:n} \mid D)}{p(e_{1:n} \mid D')}. \tag{7}$$

Importantly, it holds that

$$r(\mathcal{T}; D, D') = \log_2 \frac{p(e_{1:n} \mid D)}{p(e_{1:n} \mid D')} = -\log_2 \frac{p(e_{1:n} \mid D')}{p(e_{1:n} \mid D)} = -r(\mathcal{T}; D', D), \tag{8}$$

and for three corpora D, D', D'' we have

$$r(\mathcal{T}; D, D') + r(\mathcal{T}; D', D'') = r(\mathcal{T}; D, D''). \tag{9}$$

If and only if an n-gram occurs equally frequently in both corpora, we have $r(T; D, D') = r(T; D', D) = 0$. We can say that the higher the r value of a chord sequence, the more characteristic it is for corpus D with respect to corpus D'. Equivalently, the more negative its value, the more characteristic the pattern is in anti-corpus D' with respect to corpus D.

3.3. Analytical workflow

Following [27], we applied these calculations for the three styles under investigation. First, we obtained information-content estimates for all Roman numerals in all corpora (ABC and TAVERN were grouped together as a single classical corpus, and the same was done for the Billboard and RS popular music corpora) using IDyOM's long-term model with variable-order bounds,

escape method C, and 10-fold cross-validation at the song level (for further details, see [29]). Then, we applied boundary (b) strength calculations and retained only the tri-grams where the terminal member exceeded the established threshold. Next, we applied pMI_c calculations for each tri-gram in this filtered list. Following the original methodology, we removed all occurrences of 'non-chords', such as `N` or `X`, and removed any tri-gram that contained a repeated chord token (e.g. I – I – I or I – V – I), from the final ranking of pMI_c. Lastly, r values were calculated for each tri-gram of the filtered list of Roman Numerals across three comparisons of corpus and anticorpus: classical and popular (replicating the analysis from the original paper), choro and classical and choro and popular. Following the same methodology in [27], we adopted Laplace smoothing [15] for tri-grams that did not occur in one of the corpora in a comparison. See Figure 2 for a visual representation of the method.

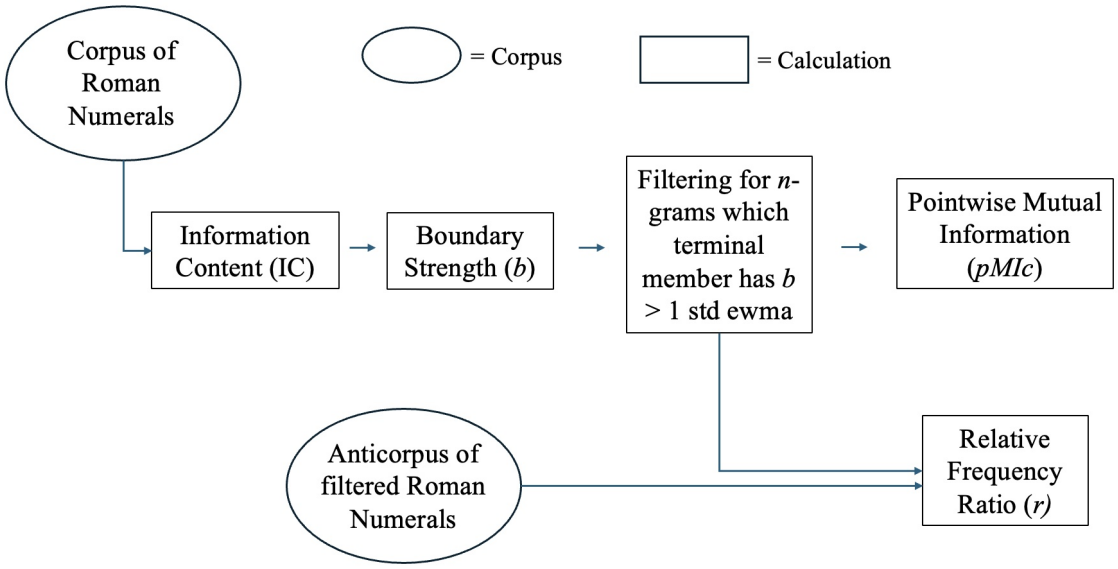


Figure 2: *Flowchart of the method for uncovering the conventional and characteristic n-grams in two corpora of Roman Numeral annotations.*

All analyses were carried out in R [25], excecpt Figure 5, which was generated in Python. In an additional step, the top 10 tri-grams from each analysis were manually classified according to a harmonic "family". This classification largely followed the guidelines in [27], with a few adaptations to account for small differences in our results, and for chord sequences in choro that featured characteristics that required their own family classification. This procedure enabled a qualitative comparison of what kinds of harmonic progressions are the most conventional and characteristic across the three styles. See Figure 3 for examples of progressions that were labeled with our family classification scheme.

4. Results

4.1. Conventional harmonic patterns per style

Table 2 presents the results of the pMI_c analyses for all three styles. The top 10 conventional tri-grams for each style are shown. As expected, the results obtained for the classical and popular

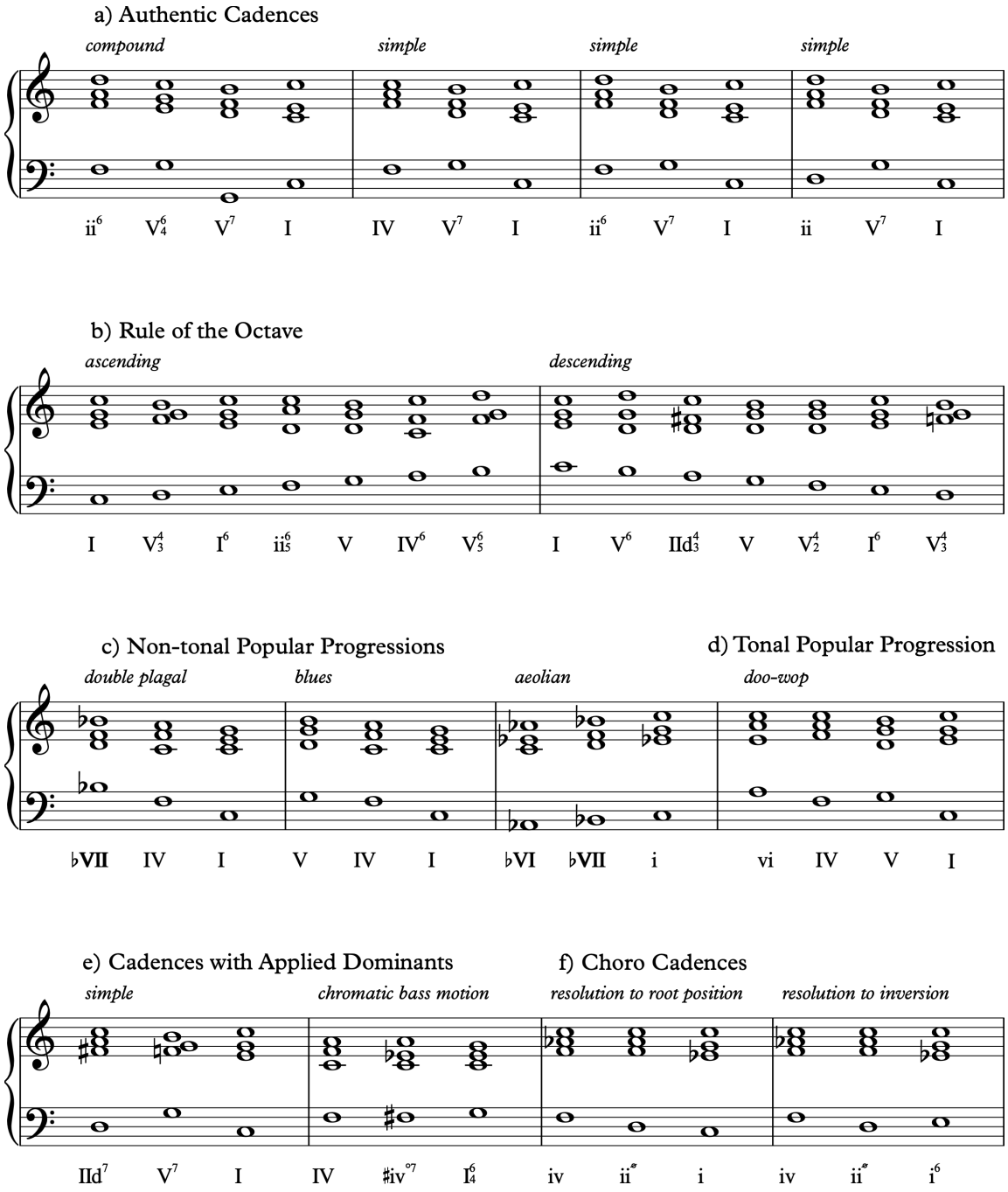


Figure 3: *Families of progressions present in the top 10 tri-grams of each analysis with specific examples in music notation. NB: to simplify reporting of results, we adopt a different notation scheme for applied dominants. If a chord is made up of a major triad with the inclusion of a minor 7th, but it is not V, it gets the shorthand "d", for "dominant" between the Roman Numeral and the 7th. For instance, a IId^7 can be read as a "II dominant 7th", and if it is followed by any V chord, it is equivalent to a V^7/V in traditional notation.*

corpora were highly correlated with those from [27] (Pearson's r = .85 for 1136 common tri-grams across the original and the present filtered lists for classical, and = .92 across 856 common tri-grams for popular, both calculated for pMI_c values), suggesting that our analysis replicated the results in the original paper to a high degree. More specifically, 8 and 6 of the tri-grams for the popular and classical corpora, respectively, shown in Table 2 were also present in the same analysis in [27], though we did not replicate the findings exactly. Despite using the same datasets and IDyOM model architecture in both studies, these differences likely reflect the choice to implement a new version of IDyOM in R for this study[12], and with different k-fold partitions of the underlying datasets. However, as we explain below, the overall characteristics of the tri-grams discussed for popular and classical in both studies were markedly similar.

Within the classical corpus, the compound cadence appeared as the most conventional closing pattern (rank 1), and its version tonicizing V also appeard in the top ten (rank 8). This corpus also included several tri-grams with stepwise bass motion, and confirmed the simple cadence progression, ii^6 – V^7 – I (rank 7), as highly conventional closing patterns. Within the popular corpora, the simple IV – V – I cadence appeared at the top of the ranked list, followed by patterns characterized by root-position triads reflecting simple tonal motion, with the addition of the aeolian, blues and double-plagal cadences as other important closing patterns. These results are closely aligned with [27].

The top ten conventional tri-grams in choro, on the right-hand side of Table 2, already exhibit several distinguishing features. The most striking is the predominance of chord sequences employing applied dominants, which received their own family label of "app. dom.". At the very top, a cadence tonicizing V before resolving to I appeared as the most conventional closing pattern. In total, 6 out of the top 10 patterns in choro employed applied dominants, compared to two in the classical corpora and none in the popular corpora.

Two harmonic patterns appeared in the choro list that defy the family classification scheme of [27]. Ranks 4 and 6 feature variations of the pattern iv – $ii^{ø}$ – i, which do not appear at all in the other styles, even before filtering. At first glance, the $ii^{ø}$ – i root motion challenges classical tonal principles, which tend to privilege root motion by step in the ascending direction (e.g., i–$ii^{ø}$, iv–V, etc.) [e.g., 31]. Such motion, however, does sometimes appear in 19th century European romantic music that remains less well-represented in corpus studies. In such cases [1] (see pp. 193, 417, 418), the iv – $ii^{ø}$ – i pattern intensifies the plagal resolution from iv to i or I, though typically not in root position as is observed in the results here for choro.

Although descending step-wise root motion is more common in the popular corpora, it rarely occurs between $ii^{ø}$ and i or i^6: across both popular corpora before filtering, this pattern appeared only 11 times. In contrast, $ii^{ø}$ – i or i^6 appeared 269 times in the choro corpus. Some well known examples in popular music feature a $ii^{ø6}$ – I motion, such as "No Surprises" by Radiohead, but the resolution to the major tonic distinguishes such cases from choro. This result is an indication that the choro corpus features at least one unique closing harmonic pattern when compared to the other two corpora, and so was labeled "choro" in the family column.

4.2. Characteristic harmonic patterns between pairs of styles

Tables 3, 4 and 5 present the results of the r analyses comparing each pair of styles, showing the tri-grams with the top ten r values. Table 3 presents the popular versus classical comparison. Pearson's r between the present and the earlier values of r in this comparison were .86, across 2175 common tri-grams, again replicating the findings from the previous study. Specifically, 8 and 7 of the top-10 most characteristic tri-grams for the popular and classical corpora, respectively, were also present in the same analysis in [27].

Table 2: *Top 10 tri-grams for each of the three styles, ranked according to their pMI_c values*

	ABC/TAVERN				RS-200/Billboard				Choro Songbook			
Rk.	ngram	pMI_c	N	Family	ngram	pMI_c	N	Family	ngram	pMI_c	N	Family
1	$V(^6_4)$ V^7 I	3.399	217	comp.	IV V I	0.541	554	simple	IId^7 V^7 I	1.951	271	app. dom.
2	I^6 V^4_3 I	1.942	61	rule	ii V I	0.157	106	simple	ii^7 V^7 I	1.367	201	simple
3	V V^4_2 I^6	1.531	48	rule	ii^7 V^7 I	0.131	75	simple	ii V^7 I	1.219	189	simple
4	ii^6 V^7 I	1.432	57	simple	i ♭VI ♭VII	0.124	120	aeolian	iv $ii^ø$ i	0.811	76	choro
5	I V^4_3 I^6	1.402	48	rule	V IV I	0.120	202	blues	IV $♯iv^o$ I^6_4	0.796	56	app. dom.
6	I Id^4_2 IV^6	1.336	22	rule	vi IV V	0.113	132	doo-wop	iv $ii^ø$ i^6	0.697	53	choro
7	ii^6_5 V^7 I	0.829	19	simple	♭VI ♭VII i	0.085	55	aeolian	I VId^7 ii	0.695	74	app. dom.
8	$II(^6_4)$ IId^7 V	0.811	42	comp.	I IV V	0.080	182	simple	I $IIId^7$ vi	0.682	68	app. dom.
9	ii V^7 I	0.751	29	simple	I ♭VII ♭IV	0.070	175	dbl. plagal	IId^7 V^7 i	0.595	84	app. dom.
10	V V^7 I	0.673	54	simple	V I IV	0.069	141	simple	$v^6(6)$ VId^7 ii	0.511	26	app. dom.

When these two styles are directly compared, the prevalence of simple tonal progressions in the top ten popular tri-grams is drastically reduced, and blues, double plagal, aeolian and other popular non-tonal progressions rank much higher when compared to the pMI_c analysis. The top ten tri-grams for the classical corpora exhibit fewer differences from the pMI_c analysis, the main one being that compound cadences rank slightly lower in this analysis. These results suggest that, when directly compared with classical harmony, popular harmony features characteristic closing patterns that do not conform to typical, classically tonal guidelines - largely a replication of the results from [27].

Table 4 presents the comparison of choro and classical harmony. The choro-type progressions rise in rank in this analysis, compared to pMI_c, namely, from ranks 6 and 4 to 5 and 2, respectively, reinforcing the interpretation that they are uniquely characteristic of choro. The high prevalence of applied dominant patterns is maintained among the top ten choro tri-grams, while the classical tri-grams only featured one pattern with an applied dominant. In the r analysis, the so-called "minor 2-5-1" progression appeared among the top 10 choro tri-grams (rank 8). This pattern was not present in the top 10 choro tri-grams in terms of pMI_c, which in turn featured two variants of the "major 2-5-1" progression (Table 2, ranks 2 and 3). This result suggests that 2-5-1 progressions are conventional closing patterns in choro, but the style makes characteristic use of the minor 2-5-1 when compared to classical harmony. In the classical portion of Table 4, the prevalence of compound cadences is even greater than in the pMI_c analysis, and in the comparison between the popular and classical corpora. This finding suggests that the double suspension above the cadential dominant, while highly characteristic of classical music, is not a typical harmonic device in choro, which instead favors descending-fifths motion.

Finally, Table 5 presents the comparison between choro and popular harmony. In this comparison, the choro pattern iv – $ii^ø$ – i rose in rank in relation to pMI_c (from 4 to 2), and applied dominants also figure prominently in the choro corpus despite being completely absent in the popular corpus. This result suggests that these kinds of closing progressions are also typical of choro when compared to the popular corpora. The minor 2-5-1 appeared in this comparison at an even higher rank (Rk. 4) than the comparison with classical, and major 2-5-1s were also absent here. This suggests that choro uniquely deploys the minor 2-5-1 as a closing pattern when compared to both popular and classical music. Blues, double plagal and aeolian cadences also figured prominently on the popular side of the table, indicating that these harmonic devices are

Table 3: *Top 10 tri-grams for popular and classical, ranked according to their r values*

	RS-200/Billboard				ABC/TAVERN				
rank	ngram	r	*N*1	*N*2	Family	ngram	r	*N*1	*N*2	Family
1	I V IV	6.941	182	0	blues	I^6 V^4_3 I	6.529	61	0	rule
2	I ♭VII IV	6.885	175	0	dbl. plagal	ii^6 V^7 I	6.433	57	0	simple
3	V I IV	6.575	141	0	simple	V V^4_2 I^6	6.189	48	0	rule
4	vi IV V	6.481	132	0	doo-wop	I V^4_3 I^6	6.189	48	0	rule
5	i ♭VI ♭VII	6.344	120	0	aeolian	$II(^6_4)$ IId^7 V	6.001	42	0	comp.
6	IV I V	6.167	106	0	blues	$V(^6_4)$ V I	5.860	38	0	comp.
7	♭VII IV I	6.140	104	0	dbl. plagal	I^6 V^6_5 I	5.433	28	0	prolong.
8	V IV I	6.091	202	1	blues	I^6 V I	5.433	28	0	simple
9	Id^7 ♭VII IV	5.818	83	0	dbl. plagal	$V(^6_4)$ V^7 I	5.343	217	7	comp.
10	ii ii(4) I	5.575	70	0	scalar	ii^6 V I	5.218	24	0	simple

Table 4: *Top 10 tri-grams for choro and classical, ranked according to their r values*

	Choro Songbook					ABC/TAVERN				
rank	ngram	r	*N*1	*N*2	Family	ngram	r	*N*1	*N*2	Family
1	IId^7 V^7 i	6.371	84	0	app. dom.	$II(^6_4)$ IId^7 V	5.465	42	0	comp.
2	iv $ii^ø$ i	6.228	76	0	choro	$V(^6_4)$ V I	5.324	38	0	comp.
3	I VId^7 ii	6.190	74	0	app. dom.	I^6 V^6_5 I	4.896	28	0	prolong.
4	I $IIId^7$ vi	6.070	68	0	app. dom.	I^6 V I	4.896	28	0	prolong.
5	iv $ii^ø$ i^6	5.717	53	0	choro	V^6 V I	4.793	26	0	prolong.
6	IId^7 V^7 I	5.464	271	5	app. dom.	ii^6 V I	4.682	24	0	simple
7	V^7 $IIId^7$ vi	5.171	36	0	app. dom.	V V^4_2 I^6	4.653	48	1	rule
8	$ii^ø$ V^7 i	5.049	67	1	simple	V^7 $V(^6_4)$ V	4.623	23	0	comp.
9	I IId^7 V^7	5.006	32	0	app. dom.	$V(^6_4)$ V^7 I	4.485	217	9	comp.
10	vi^7 ii^7 V^7	4.916	30	0	simple	ii^6 $V(^6_4)$ V	4.431	20	0	comp.

unique to popular music when compared to the choro tradition. Overall, these results indicate that roughly the same harmonic characteristics that set choro harmony apart from the classical corpora also distinguish it from the popular corpora. Therefore, across the three comparisons, choro harmony distinguishes itself mainly by its frequent use of applied dominants, the iv – $ii^ø$ – i pattern, and the minor 2-5-1 progression.

The main noticeable difference between the comparisons of choro with the classical and popular corpora is that the pattern IV – $♯iv^o$ – I^6_4 appeared as highly characteristic of choro when compared to the popular corpora (Rk. 5), but not when compared to classical. This pattern appeared in rank 5 in the choro pMI_c analysis. This is likely because classical music often features chromatic stepwise ascending motion to the dominant scale-degree in cadential progressions, just as it does in choro.

Table 5: *Top 10 tri-grams for choro and popular, ranked according to their r values*

	Choro Songbook					RS-200/Billboard				
rank	ngram	r	*N*1	*N*2	Family	ngram	r	*N*1	*N*2	Family
1	IId^7 V^7 i	6.946	84	0	app. dom.	IV V I	8.580	554	0	simple
2	iv ii^{ø} i	6.803	76	0	choro	V IV I	7.129	202	0	blues
3	I VId^7 ii	6.765	74	0	app. dom.	I IV V	6.980	182	0	simple
4	ii^{ø} V^7 i	6.624	67	0	simple	I V IV	6.980	182	0	blues
5	IV ♯iv^{o} I$^6_4$	6.369	56	0	app. dom.	I ♭VII IV	6.923	175	0	dbl. plagal
6	iv ii^{ø} i^6	6.291	53	0	choro	V I IV	6.614	141	0	simple
7	i Id^7 iv	6.209	50	0	app. dom.	vi IV V	6.519	132	0	doo-wop
8	iv V^7 i	6.028	44	0	simple	i ♭VI ♭VII	6.383	120	0	aeolian
9	iii V^7 I	5.822	38	0	simple	IV I V	6.205	106	0	blues
10	V^7 IIId^7 vi	5.746	36	0	app. dom.	♭VII IV I	6.178	104	0	dbl. plagal

4.3. Characteristic harmonies across the three styles

The conventional and characteristic tri-gram analyses do not automatically enable a quantitative comparison of the tri-grams across the three styles at once. In order to achieve this, all tri-grams featured in each of the three *r* calculations were labeled with the style they were most characteristic of, according to a comparison of the *r* values obtained in the three pair-wise analyses. If a tri-gram had an *r* value that was highest for one style in two out of three comparisons, that style received a score of 2 and was considered its main style (since each style was present only in two out of the three comparisons). In this process, Laplace smoothing was adopted again for tri-grams that were not present in a comparison. See Table 6 for an example of this labeling procedure.

Table 6: *Example of the procedure for labeling tri-grams according to which style they were most characteristic of. Raw counts of the occurrence of the tri-grams on the filtered lists for each style are also included.*

ngram	popular/ classical *r*	choro/ popular *r*	choro/ classical *r*	popular count	choro count	classical count	popular score	choro score	classical score	main style
I VId$^6_5$ ii	-3.160	4.706	1.547	0	17	5	0	2	1	choro
I I^6 ii^6	-3.160	0.536	-2.623	0	0	5	0	1	2	classical
I ii^6 V	-3.160	0.536	-2.623	0	0	5	0	1	2	classical
V$^6_5$ I V	-3.160	0.536	-2.623	0	0	5	0	1	2	classical
i V I	-3.160	0.536	-2.623	0	0	5	0	1	2	classical
V$^4_3$ V7 I	-3.312	3.039	-0.273	2	16	19	0	1	2	classical
V$^6_5$ V7 I	-3.312	1.536	-1.775	2	5	19	0	1	2	classical
IV^6 V$^6_5$ I	-3.382	-0.464	-3.846	1	0	13	1	0	2	classical
V V^6 I	-3.382	-1.049	-4.431	2	0	20	1	0	2	classical

To better evaluate which kinds of progression are most characteristic of each style, the 50 tri-grams that had the highest and lowest values were selected from each *r* analysis, resulting in 300 tri-grams of a total of 5028 (5.97%). This essentially amounted to selecting the 100 most characteristic tri-grams of each style across the three *r* analyses. These tri-grams were manually labeled according to their music-theoretic family. The labels were further adapted to account for some kinds of progressions that did not appear in [27] or in the previous tables, and to reduce the number of families to optimize understanding and visualization. All non-tonal popular progressions (e.g. "blues" and "aeolian") were grouped under "non-tonal". All simple authentic

cadences and rule-of-the-octave motions were grouped under "simple". The label "scalar" was adopted for tri-grams that only featured step-wise root motions (e.g. ii – ii(4) – I or IV^{M7} – iii^{7} – ii^{7}). "Extensions" was used for progressions that featured extensions including the 9^{th} and beyond (e.g. $ii^{7(9)}$ – $V^{(13)}$ – I^{M7}). "Jazz" was adopted for progressions that included the motion between iii^{o} – ii (and its variants including 7^{ths} and inversions), and for the progression $\flat II^{7}$ – V^{7} – i. The results of this analysis are summarized in Figure 4.

The figure shows a stark predominance of applied dominant progressions within choro, making up the largest family and almost half of all top tri-grams in this style (43%), and much smaller shares in the other two styles. This corroborates the earlier results suggesting that applied dominants are uniquely characteristic of choro when compared to the selected classical and popular corpora. The family of simple tonal progressions made up the largest share of tri-grams in the classical corpora (71%), and the second-largest in choro (25%) and popular (24%), suggesting that, while all three styles often feature typical tonal motions, the classical corpora deploy them to a much greater extent. Classical obtained the largest share of compound progression out of the three styles (16%), its second largest family, followed by applied dominant progressions (13%). This corroborates the compound cadences as uniquely characteristic of classical among these three styles. Almost half of the popular progressions were non-tonal (45%), it's largest family. This could be interpreted as aligning with previous research showing that harmonic motion in popular is largely less constrained than in classical music [31]. This feature also sets popular harmony apart from choro, which only had 17% non-tonal tri-grams. The popular corpora obtained 8% of scalar progressions versus none in the choro and classical corpora, suggesting that step-wise root motion is more typical of popular music than of choro or Western classical music. The popular corpora concentrated most cadences with extensions (at 11%), and the choro corpus, all the jazz cadences (at 11%).

Figure 5 plots all of the trigrams in all *r* analyses in a two-dimensional plane as a polar plot. This is possible because the three *r* analyses are mathematically interrelated (see Equation 9). The figure can be read as revealing a two-dimensional space of harmonic characteristicness of three-chord closing patterns in choro, popular and Western classical music. In broad terms, the upper left-hand side of the plot is the choro space, the bottom of the plot is the classical space, and the upper right-hand side, the popular space. Towards the left, a predominance of tri-grams featuring applied dominants, as well as the choro progression and the minor 2-5-1, can be found. Towards the right, the blues and double plagal progressions are located. Towards the bottom, two compound cadences are visible.

The areas of the figure in-between any two styles exhibit tri-grams that can be regarded as more characteristic of both, and less of the third style. For example, IV – V – I, towards the right, is often taught in music theory contexts as the prototypical tonal progression. Its location on the figure suggests that it is more characteristic of popular than classical music, but more characteristic of classical than choro. Accordingly, this tri-gram appeared at the top of the *r* list for popular versus choro, but not at all in the list of popular versus classical. The tri-gram IV – V^{7} – I, on the other hand, is located towards the middle of the figure, which suggests that it is similarly non-characteristic of all three styles, or that it is shared between all of them. This is in line with Almada's assertion that the V chord in choro always features the 7^{th} [2]. This result suggests that the three styles exhibit typically tonal progressions, but that choro rarely adopts the V triad on it's own, relative to the other two styles.

The tri-gram, I^{6} – V^{4}_{3} – I is in-between the classical and choro spaces, and away from the popular space. The IV – $\#iv^{o}$ – I^{6}_{4} progression also appears in between classical and choro and away from popular, but is located closer to the extreme of the choro space. These two tri-grams illustrate the larger use of inversions and stepwise bass motion in classical and choro music when

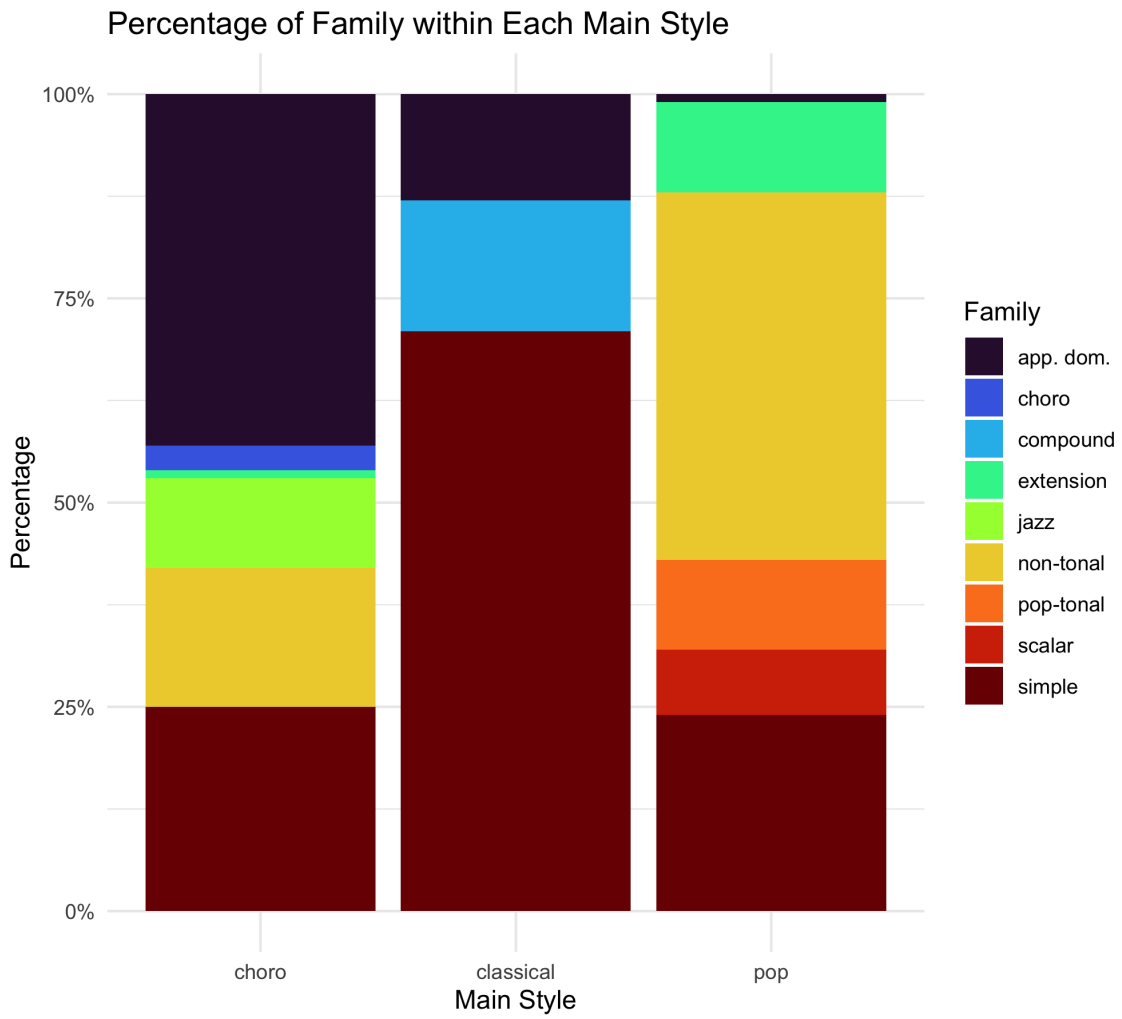


Figure 4: *Percentage of harmonic families of the top 100 tri-grams within each main style with the highest r values across the analyses (300 in total).*

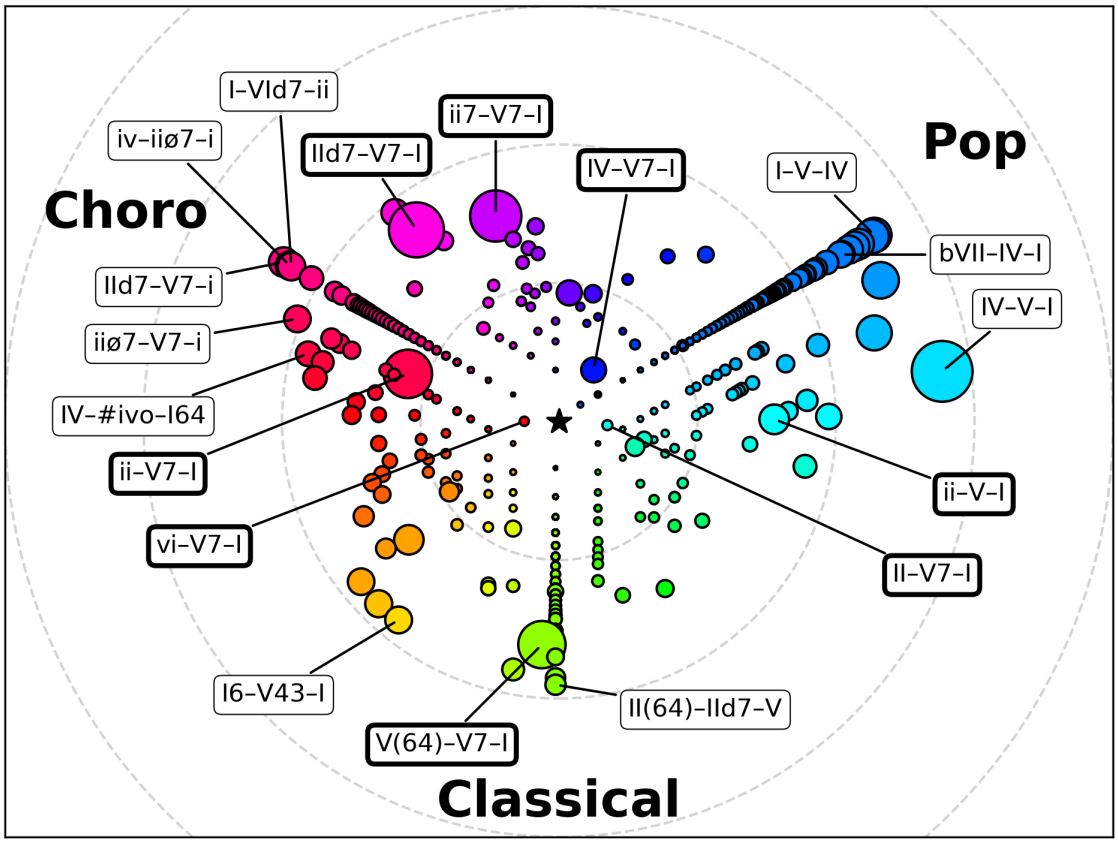


Figure 5: *Harmonic space between the three styles as a polar plot. Values come from the full dataset of 5028 tri-grams of which Table 6 is a snippet. Each circle represents a chord tri-gram occurring in all r analyses, and the size of the circles is proportional to their total TF-IDF score based on raw counts across all filtered tri-gram lists. Selected tri-grams are labeled for illustration. Tri-grams occurring in only one of the three styles appear on one of the three visible axes, all other tri-grams occur in at least two styles. Among the 18 labeled trigrams, 8 occurred in the filtered lists of all corpora and have their labels highlighted. Dashed gray circles represent r values of equal magnitude and are added for orientation.*

compared to popular music.

In the in-between spaces, it is also possible to see many varieties of the major 2-5-1 progression. The ii^7 – V^7 – I variant is located between the popular and choro spaces, while the ii – V^7 – I, between choro and classical, in both cases, further in the direction of the choro space. This suggests that these variants are slightly more characteristic of choro than the other two styles, that the variant with a 7th added to the ii chord is more characteristic of popular than classical music, and, conversely, that the variant with the ii triad is more characteristic of classical than popular music. The variant without any 7^{ths}, ii – V – I, is between the classical and popular spaces and away from choro, in line with the relative lack of the V triad in choro music.

In addition to IV – V^7 – I, the vi – V^7 – I progression also appears close to the middle of the figure. This sequence can be viewed simply as a variant of the typically tonal IV – V^7 – I, in line with text-book tonal music theory, which typically claims that a vi chord can fulfill the same predominant function as a IV chord, and thus substitute for it. Their appearance close to the nexus of all three styles suggests that these two tri-grams make up prototypical tonal progressions at formal boundaries similarly in all three tonal styles. In addition, the II – V^7 – I also appears close to the middle, suggesting that it is also shared across styles. These three progressions appear in the filtered lists of all three styles, regardless of smoothing.

5. Discussion

The results of this study point towards a number of key features distinguishing the harmonic styles of choro, Western classical and popular music at moments of formal segmentation, as well as some features that they share. First, the harmonic vocabulary of choro, while overwhelmingly tonal and tertian, makes much more use of applied dominants and chromatic descending-fifths motion than the other two styles. It also employs unique harmonic patterns such as iv – ii^{ø} – i, indicating ways of articulating form with harmony that are not predicted by typical tonal theory. The minor 2-5-1 pattern also features prominently in choro, while variants of the major 2-5-1 are shared to a greater degree with the other two styles. Choro also deploys more 7^{ths} in its 2-5-1 variants than the other styles. The results related to the 2-5-1 patterns could be explained by the influence that jazz exerted on choro in the latter half of the 20th century, a period strongly represented in the data [19]. Jazz typically employs more 7^{ths} than classical and popular music [4], which could be reflected in the higher r values for choro of patterns containing ii^7 and V^7, among others. It may be the case that the use of applied dominants can also be traced to this influence, although this hypothesis would require further investigation.

It should be noted, however, that the inclusion or absence of 7^{ths} in choro chords is, to a significant extent, a choice of the performer, and the notated 7^{ths} in our data are a reflection of particular editorial decisions of the scores that make up the choro corpus [17]. While the inclusion of 7^{ths} in the overwhelming majority of V chords in choro is relatively safe to assume [2], the same is not necessarily true for other chords such as ii. Future work adopting chord estimation from audio analysis of a significant corpus of choro recordings might provide a more robust assessment of the presence of 7^{ths} in choro.

Stepwise bass motion and the use of inversions appeared as uniquely classical characteristics, a replication of the results in [27]. These features also figured to some extent in choro harmony, especially with respect to the IV – #iv^o – I$^6_4$ pattern, but were notably absent in the popular corpus. However, the nature of the data likely underestimates the use of inversions in choro, which typically features highly active and often improvised bass lines [14]. The inversions that eventually make their way into notated chord symbols, which are the basis of the choro songbook corpus, are likely only a fraction of the inversions that are actually heard in the style, which, similarly to

7^{ths}, are often at the discretion of the performer. To address this issue, further work employing corpora of transcribed or composed bass lines, or audio analysis, will be necessary.

Some characteristics of choro resemble nineteenth-century Western tonal traditions, which have yet to feature prominently in harmonic corpus studies. Much like the popular and choro corpora discussed here, nineteenth-century harmony tends to distinguish itself from previous common-practice eighteenth-century traditions by its increased use of stepwise root motion and chromaticism at points of cadential closure, for example [6]. Moreover, the characteristic choro pattern, iv – $ii^{ø}$ – i, could stem from the influence of European romantic music on choro, since the European dances that were performed in the early days of choro were likely contemporary, nineteenth-century pieces[14]. Further research comparing choro and European romantic music could further illuminate the relationships between choro harmony and the European tradition.

Our results suggest that popular harmony at formal boundaries exhibits greater use of simple triads in root position, along with more non-tonal progressions such as the double plagal or aeolian progressions. These features distinguish popular from classical music, which largely replicates [27], as well as with the choro corpus examined here. In particular, the typical asymmetries in root motion that distinguish classical harmonic progressions from those in popular music [31] are shared with choro: both classical and choro privilege falling to rising 5^{ths} root motions, while these motions are used relatively equally in popular.

Lastly, our results revealed a few harmonic patterns that are present but not characteristic of any of the three styles. In particular, the prototypical tonal motion IV – V^7 – I and its variation vi – V^7 – I are shared across these styles. This result suggests that the music-theoretic notion that these progressions provide the basis for cadences in tonal music might have some empirical grounding. However, evidence from an even more diverse set of traditions would be required to better assess this claim.

This study relies on certain assumptions. First, we assume that the relationship between information content and boundary estimation, which was empirically established in studies with listeners familiar with Western tonal music [24], also holds for choro due to the latter's strong ties with Western tonal harmony. Second, we also assume that a window size of 20 events is also appropriate for identifying formal boundaries in choro. Future studies should examine these assumptions empirically either computationally or with choro listeners.

6. Conclusion

This study revealed unique harmonic characteristics at formal boundaries of choro music when compared with traditions associated with Anglophone popular and Western classical music. In so doing, it also revealed aspects of shared tonal grammar across the selected corpora. As the first study to attempt a direct, data-driven comparison of the harmonic vocabulary of three styles, including one from the Global South, it opens up the field of research to continue to develop and compare harmonic corpora for triadic traditions in diverse contexts. Many more unique and shared harmonic devices and behaviors may be uncovered by comparative corpus research on diverse styles, with the potential to expand music theory's understanding of how musical harmony manifests in different parts across the world.

The potential connections observed here between choro and jazz suggest that comparative work including these styles might better reveal to what extent they share harmonic traits. Further work would also benefit from employing other cardinalities of *n*-grams, and from including repetitions of chord tokens, which were excluded in this study. Choro itself is not a monolithic genre, and is made up of several subgenres like waltz, polca, lundu, tango and baião. These subgenres are often explicitly notated in choro scores. Future research might attempt to better characterize the

choro subgenres based on quantitative data of their harmonic characteristics. This would require expanding the corpus of [19] to include larger samples of pieces from each subgenre, in particular from the early years of choro, an endeavor that is currently in its early stages.

Our results encourage, in particular, the expansion of data-driven research on the stylistic and harmonic traits of Brazilian music. Choro is reportedly a key influence on several later styles in the country (e.g. [14][20]), and quantitative investigations of these other styles can help reveal if and how they are indebted to, but also diverge from, choro harmony. Brazilian classical or concert music also shares significant connections with choro that deserve further exploration [14]. The area of computational analysis of Brazilian music as a whole is still in very incipient stages. Brazil is a country of continental dimensions that has historically been the destination of several large waves of migration from nearly all parts of the globe. Considering the immense diversity inherent in Brazilian culture and music, this constitutes an exciting field for future research.

References

[1] Aldwell, Edward; Schachter, Carl (2002). *Harmony and Voice Leading*. Wadsworth Group.

[2] Almada, Carlos (2006). *A Estrutura Do Choro*. Da Fonseca Comunicação: Rio de Janeiro.

[3] Barbosa, Eliana Rosa de Queiroz (2021). 'Being the culture' and 'playing the culture': Choro and the Brazilianness performed in Brussels. *Crossings: Journal of Migration & Culture*, v. 12, n. 2, pp. 413–428.

[4] Broze, Yuri; Shanahan, Daniel (2013). Diachronic Changes in Jazz Harmony: A Cognitive Perspective. *Music Perception: An Interdisciplinary Journal*, v. 31, n. 1, pp. 32–45.

[5] Burgoyne, John Ashley; Wild, Jonathan; Fujinaga, Ichiro (2011). An Expert Ground-Truth Set for Audio Chord Recognition and Music Analysis. *12th International Society for Music Information Retrieval Conference*, n. ISMIR, pp. 633–638.

[6] Caplin, William E. (2018). Beyond the Classical Cadence: Thematic Closure in Early Romantic Music. *Music Theory Spectrum*, v. 40, n. 1, pp. 1–26.

[7] Chuva, Maria Regina Romeiro (2024). Parecer da Relatora: 103a Reunião do Conselho Consultivo do Patrimônio Cultural, 29 de fevereiro de 2024. Instituto do Patrimônio Histórico e Artístico Nacional, Ministério de Cultura, Governo Federal.

[8] Clercq, Trevor de; Temperley, David (2011). A Corpus Analysis of Rock Harmony. *Popular Music*, v. 30, n. 1, pp. 47–70.

[9] Cover, Thomas M.; Thomas, Joy A. (2006). *Elements of Information Theory*. John Wiley & Sons: Hoboken, NJ.

[10] de Clercq, Trevor (2022). Tempo versus Average Rates of Harmonic Rhythm in Popular Music: A Study of Three Corpora. *Musicae Scientiae*, pp. 10298649221091483.

[11] Devaney, Johanna; *et al* (2015). Theme and Variation Encodings with Roman Numerals (TAVERN): A New Data Set for Symbolic Music Analysis. *Proceedings of the 16th International Society of Music Information Retrieval (ISMIR) conference.* pp. 728–734.

[12] Harrison, Peter M. C.; *et al* (2020). PPM-Decay: A Computational Model of Auditory Prediction with Memory Decay. *PLoS Computational Biology*, v. 16, n. 11, pp. e1008304.

[13] Harte, Christopher A.; *et al* (2005). Symbolic Representation of Musical Chords: A Proposed Syntax for Text Annotations. *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR)*, v. 56: London, pp. 66–71.

[14] Livingston-Isenhour, Tamara Elena; Garcia, Thomas George Caracas (2005). *Choro: A Social History Of A Brazilian Popular Music*. Indiana University Press: Bloomington.

[15] Manning, Christopher D.; Schütze, Hinrich (1999). *Learning to Listen, Listening to Learn: Music Perception and the Psychology of Enculturation*. MIT Press: Cambridge.

[16] Meyer, Leonard B. (1989). *Style and Music. Theory, History, and Ideology*. University of Chicago Press.

[17] Moss, Fabian C.; Fernandes de Souza, Willian; Rohrmeier, Martin (2023). Choro Songbook Corpus. Zenodo.

[18] Moss, Fabian C.; Nakamura, Eita (2024). Modeling the Evolution of Harmony in Popular Music from Different Cultural Contexts. *Proceedings of the Computational Humanities Research Conference*, v. 3834, pp. 137–152.

[19] Moss, Fabian C.; Souza, Willian Fernandes; Rohrmeier, Martin (2020). Harmony and Form in Brazilian Choro: A Corpus-Driven Approach to Musical Style Analysis. *Journal of New Music Research*, v. 49, n. 5, pp. 416–437.

[20] Neto, Lira (2017). *Uma história do samba*. Companhia das Letras: São Paulo.

[21] Neuwirth, Markus; *et al* (2018). The Annotated Beethoven Corpus (ABC): A Dataset of Harmonic Analyses of All Beethoven String Quartets. *Frontiers in Digital Humanities*, v. 5, n. July, pp. 1–5.

[22] Pearce, Marcus T. (2005). *The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition*. PhD thesis, City University, London.

[23] Pearce, Marcus T. (2025). *Learning to Listen, Listening to Learn: Music Perception and the Psychology of Enculturation*. Oxford University Press.

[24] Pearce, Marcus T.; Müllensiefen, Daniel; Wiggins, Geraint A. (2010). The Role of Expectation and Probabilistic Learning in Auditory Boundary Perception: A Model Comparison. *Perception*, v. 39, n. 10, pp. 1367–1391.

[25] R Core Team (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

[26] Sandroni, Carlos (2021). *A respectable spell: transformations of samba in Rio de Janeiro*. University of Illinois Press: Urbana.

[27] Sears, David R. W.; Forrest, David (2021). Triadic Patterns across Classical and Popular Music Corpora: Stylistic Conventions, or Characteristic Idioms? *Journal of Mathematics and Music*, v. 15, n. 2, pp. 140–153.

[28] Sears, David R. W.; Widmer, Gerhard (2020). Beneath (or beyond) the Surface: Discovering Voice-Leading Patterns with Skip-Grams. *Journal of Mathematics and Music*, v. 15, n. 3, pp. 209–34.

[29] Sears, David R. W.; *et al* (2018). Simulating Melodic and Harmonic Expectations for Tonal Cadences Using Probabilistic Models. *Journal of New Music Research*, v. 47, n. 1, pp. 29–52.

[30] Shea, Nicholas; *et al* (2024). Diversity in Music Corpus Studies. *Music Theory Online*, v. 30, n. 1.

[31] Temperley, David (2018). *The musical language of rock*. Oxford University Press: New York.

[32] Tinhorão, José Ramos (1990). *História social da música popular brasileira*. Editorial Caminho: Lisbon.