

Submitted exclusively to the *Journal of Mathematics and Music*  
 Last compiled on May 19, 2021

## A Computational Exploration of Melodic Patterns in Arab-Andalusian Music

Thomas Nuttall<sup>\*,a</sup>, Miguel G. Casado<sup>a</sup>, Andres Ferraro<sup>a</sup>, Darrell Conklin<sup>b,c</sup>, Rafael Caro Repetto<sup>a,d</sup>

<sup>a</sup>*Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain;*

<sup>b</sup>*Department of Computer Science and Artificial Intelligence,  
 University of the Basque Country UPV/EHU, San Sebastian, Spain;*

<sup>c</sup>*IKERBASQUE, Basque Foundation for Science, Bilbao, Spain;*

<sup>d</sup>*Institute of Ethnomusicology, Kunstuniversität Graz, Austria;*

()

Here we present a computational approach to identifying melodic patterns in a dataset of 145 MusicXML scores with the aim of contributing to centonization theory in the Moroccan tradition of Arab-Andalusian Music - a theory in development by expert performer and researcher of this tradition, Amin Chaachoo. Central to his work is the definition of a set of characteristic patterns, or centos, for each *ṭabʿ*, or melodic mode. We apply three methods: TF-IDF, Maximally General Distinctive Patterns (MGDP) and the Structure Induction Algorithm (SIA) to identify characteristic patterns at the level of *ṭabʿ*. A substantial number of the centos proposed by Chaachoo are identified and new melodic patterns are retrieved. A discussion with Chaachoo about the obtained results promoted the elicitation of other categories of recurrent patterns in the tradition different from the centos, contributing to a deeper musicological knowledge of the tradition.

**Keywords:** *computational musicology; pattern discovery; centonization; exploratory; Arab-Andalusian*

### 1. Introduction

#### 1.1. Characteristics of Arab-Andalusian music

Arab-Andalusian music formed in the medieval Islamic territories of the Iberian Peninsula, known as Al-Andalus, as a result of the combination of local musical traditions with Arab poetry and aesthetics from the Middle East. Its core element is the *ṣanʿa* (plural *ṣanāʿiʿ*), a poem sung by a choir accompanied by an instrumental ensemble. These *ṣanāʿiʿ* are performed in suites known as *nawabāt* (plural of *nawba*), which include orchestral pieces and both instrumental and vocal solo improvisations (Chaachoo 2016; Guettat 2000). The *nawba* is the essential form of Arab-Andalusian music. Traditionally, all *ṣanāʿiʿ* and other pieces in one *nawba* are composed in one single melodic mode, known as *ṭabʿ* (plural *ṭabūʿ*).

With the migration of the Andalusian population to North Africa following the defeat of Islamic kingdoms in the Iberian peninsula, Arab-Andalusian music was taken to North

---

\*Corresponding author. Email: thomas.nuttall@upf.edu

Africa, where it has been preserved until today. Nowadays, it is the classical music repertoire in countries such as Morocco, Algeria and Tunisia, in each of which it developed into a particular tradition (Poché 1997; Guettat 2000). In this paper, we focus on the Moroccan repertoire, known as *al-Āla* (Chaachoo 2016).

Over the last decade, the researcher and expert performer of *al-Āla* Amin Chaachoo has been developing a music theory for this tradition (Chaachoo 2011, 2016, 2019). One of the most relevant hypotheses of Chaachoo’s theory is the Iberian roots of the melodic dimension of Arab-Andalusian music, specifically arguing that the *ṭab‘* exhibits different characteristics to Middle Eastern *maqam* (Chaachoo 2019, 19). According to this claim, Chaachoo explains *ṭubū‘* in terms of the modal theory developed for plainchant. Thus, a *ṭab‘* is defined by a particular ascending and descending scale, a fundamental degree similar in function to the *finalis* of Gregorian modes, several principal degrees, one or two persistent degrees in the manner of “the reciting tone” and a series of characteristic melodic phrases.

As a theoretical framework for the function of these characteristic melodic phrases in Arab-Andalusian *ṭubū‘*, Chaachoo draws on the *centonization* model. Centonization, from Latin *cento* meaning patchwork, is defined as a plainchant compositional technique consisting of the combination of pre-existing melodic units called *centos* (Ferretti and Agaësse 1938; Chewand and McKinnon 2001; Apel 1958). Chaachoo argues that melodies in Arab-Andalusian music are also created by the combination of *centos*, thus strengthening the connection between this tradition and Iberian local practices. Consequently, the identity of a particular *ṭab‘* is also defined by a corresponding set of such *centos*, see Fig.1. In each new publication of his theory, Chaachoo presents a slightly modified list of *centos* per *ṭab‘*.

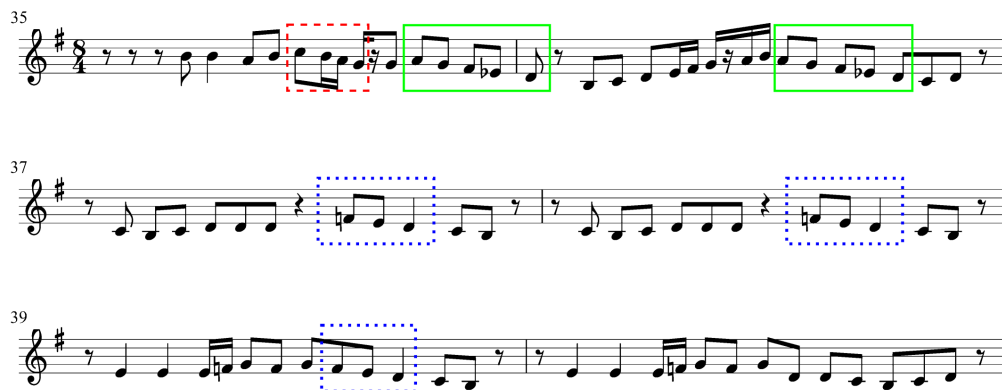


Figure 1. First measures of the first *san‘a* in the score *Btāyḥī al-ḥiḡāz al-kabīr* (with the MBID 12ce112f-38ed-4700-94ec-a329d06f6196). Boxes with the same border type indicate occurrences of the same *cento*.<sup>1</sup>

## 1.2. Motivation and objectives

Chaachoo’s centonization theory for *al-Āla* is in continuous development and we have available to us a transcribed symbolic *ṭab‘*-annotated collection of many hours of performance in this tradition. We attempt to contribute to this developing theory with an empirical investigation into the extent to which Chaachoo’s *centos* characterize the melodic content of each *ṭab‘*, further suggesting other salient melodic patterns identified

<sup>1</sup>The full score is available in *MuseScore* ([https://musescore.com/mtg/brihiorchestra\\_rtm1960s\\_btayhihijazkibir](https://musescore.com/mtg/brihiorchestra_rtm1960s_btayhihijazkibir)).

by our approach as being characteristic.

It is important to note that the *centos* proposed by Chaachoo can not be considered a strict ground truth and as such the main objective of this paper is not to validate our pattern discovery approaches on this collection but instead to explore whether we can validate parts of Chaachoo's theory empirically, suggesting new and interesting areas of investigation for its continual development.

### 1.3. Previous work

Automatic symbolic pattern discovery is an active research area in Music Information Retrieval that aims to automatically identify *musically meaningful* patterns in symbolic representations of music such as scores. An example of this is the MIREX challenge, *Discovery of Repeated Themes Sections*, proposed until 2017.<sup>1</sup>

There exists many studies into melodic pattern discovery in symbolic scores, summaries of which have been made by Janssen et al. (2013) and more recently Ren et al. (2017). Lack of agreement on the current state-of-the-art stems from the difficulty in evaluating approaches, with expertly annotated ground truth often required for performance measurement, more often than not on a study-by-study basis.

The task of pattern discovery within the context of Arab-Andalusian music is a much more recent development, with Nuttall et al. (2019) taking advantage of the Arab-Andalusian corpus gathered as part of the CompMusic project to identify patterns characteristic of *nawabāt*. Since then, Chaachoo has published a new and revised version of his centonization theory (Chaachoo 2019), offering us the opportunity to explore this tradition from a new perspective with a wider range of methods.

## 2. Dataset

Our dataset is a subset of the Music Scores Collection of the Arab-Andalusian Music Corpus gathered in the context of the CompMusic project (Serra 2014).<sup>2</sup> This corpus is the largest source of machine readable data for the computational study of this music tradition.

The Arab-Andalusian Music Corpus comprises three different but related collections: (1) The Audio Recordings Collection - containing 156 audio recordings from the personal collection of Amin Chaachoo, donated to the CompMusic project by him and which are now available under Creative Commons license. These recordings have a mean duration of one hour and contain performances on radio and at private events from the 1960s and 1970s. They are selected for the performance quality of the featured orchestras, namely the Tetouan Orchestra, Orchestra of the Conservatory of Tetouan, Brihi Orchestra and RTM Orchestra, which include celebrated maestros of *al-Āla*, (2) The Music Scores Collection consists of 158 manual transcriptions of audio recordings done by Amin Chaachoo. Since Arab-Andalusian music is heterophonic in texture, only the predominant melody is transcribed, which underlies the specific renditions by the different instruments in the orchestra and by the choir (see Figure 1). Solo improvisations have not been transcribed. The scores are stored as MusicXML files and are also available under Creative Commons license, (3) The Lyrics Collection - containing non-aligned lyrics of all recordings, both in their original Arabic script and in an automatically generated romanization (Sordo,

---

<sup>1</sup>[https://www.music-ir.org/mirex/wiki/2017:Discovery\\_of\\_Repeated\\_Themes\\_%26\\_Sections](https://www.music-ir.org/mirex/wiki/2017:Discovery_of_Repeated_Themes_%26_Sections)

<sup>2</sup><https://compmusic.upf.edu/>

Table 1. Distribution of scores across *ṭabūʿ*

<i>ṭabūʿ</i> name	Number of Scores
<i>al-istihlāl</i>	23
<i>al-iṣbahān</i>	13
<i>al-mašriqī</i>	10
<i>al-māya</i>	12
<i>al-raṣd</i>	10
<i>al-ḥiṣṣāz al-kabīr</i>	10
<i>al-ḥiṣṣāz al-mašriqī</i>	5
<i>al-ṣīka</i>	1
<i>al-ʿuṣṣāq</i>	7
<i>garībat al-ḥusayn</i>	12
<i>raml al-māya</i>	19
<i>raṣd al-dāyil</i>	16
<i>ʿirāq al-ʿaḡam</i>	7

Chaachoo, and Serra 2014). The metadata of the recordings is stored in MusicBrainz - the scores, audio and lyrics of an individual performance are connected by and stored under a common MusicBrainz ID (MBID).<sup>3</sup>

The corpus is integrated into Dunya (Porter, Sordo, and Serra 2013), from where it can be retrieved using a series of Jupyter Notebooks.<sup>4</sup> It can be also downloaded in bulk from a Zenodo repository.<sup>5</sup>

Our dataset consists of 145 scores from the Music Scores Collection of the Arab-Andalusian music corpus, having discarded the scores from the collection that do not belong to any of the classical *ṭabūʿ*. The metadata of the scores - inherited from the recordings - specify the *nawba* and rhythmic mode of each piece. The scores have been manually segmented in terms of structural sections according to the manual segmentation of the recordings done by Chaachoo and also inherit his annotations of form and *ṭabūʿ*. According to these annotations, the corpus covers 13 of the 26 classical *ṭabūʿ* theoretically established for the *al-Āla* tradition. The distribution of scores across *al-Āla* is shown in Table 1.

### 3. Methodology

We apply three pattern discovery techniques to our symbolic dataset in an attempt to identify which melodic patterns are most characteristic of each *ṭabūʿ*, cross-referencing our findings with Chaachoo’s (2019) latest publication on centonization theory.

#### 3.1. TF-IDF

Our first method is an approach first introduced by Nuttall et al. (2019) on the same dataset in which characteristic patterns are investigated at the level of *nawba*. Here we use the same methodology to identify representative patterns on a *ṭabūʿ* level.

##### 3.1.1. Data representation

Each score is represented by a “bag of patterns” (analogous to a bag-of-words representation in natural language processing) in which *n-grams* - in this instance concatenated

<sup>3</sup><https://musicbrainz.org/collection/142ea0d7-7fdf-4ea5-9b04-219f68023d01>

<sup>4</sup><https://github.com/MTG/andalusian-corpus-notebooks>

<sup>5</sup><https://doi.org/10.5281/zenodo.1291775>

sequences of consecutive notes - up to a pre-specified length,  $N$ , are extracted from each score to build an unordered multi-set of patterns that exist within it. Any n-gram that contains a rest,  $R$ , is discarded, as is any temporal information. For a score of  $[G, E, F, F, R, E, G, E]$ , the bag-of-patterns representation for  $N = 2, 3$  and  $4$  is as follows:

$$\begin{aligned} N = 2: & [GE, EF, FF, EG, GE]; \\ N = 3: & [GE, GEF, EF, EFF, FF, EG, EGE, GE]; \\ N \geq 4: & [GE, GEF, GEFF, EF, EFF, FF, EG, EGE, GE] \end{aligned}$$

### 3.1.2. TF-IDF statistic

For each pattern,  $p$ , and each score,  $s$ , a TF-IDF statistic with sub-linear tf scaling is computed to measure the extent to which each pattern is *over/under-represented* in the score. The *term-frequency*, *inverse document frequency* and TF-IDF statistic are presented in equations 1, 2 and 3 respectively.  $df(p)$  is the raw count of scores in which pattern  $p$  occurs;  $f_{p,s}$  is the raw count of occurrences of pattern  $p$  in score  $s$ ;  $n$  is the total number of scores.

$$\text{tf}(p, s) = 1 + \log(f_{p,s}) \quad (1)$$

$$\text{idf}(p) = \log\left(\frac{1+n}{1+df(p)}\right) + 1 \quad (2)$$

$$\text{TF-IDF}(p, s) = \text{tf}(p, s) \cdot \text{idf}(p) \quad (3)$$

Patterns that occur more frequently than would be expected for a given score will return a higher TF-IDF statistic and are considered *characteristic* of the score.

With the TF-IDF statistic as a measure of how characteristic each pattern is of each score we group scores by our *tab*' annotations and average the statistic for each pattern (as in Nuttall et al. 2019). The result is a measure of how characteristic each pattern is of each *tab*'.

## 3.2. Structure Induction Algorithm

Structure Induction Algorithm (SIA) (Meredith, Lemström, and Wiggins 2002; Meredith and Wiggins 2001) is one of the most popular methods for symbolic pattern discovery, basing its approach on geometric methods. The algorithm has recently been applied to Western music with successful results on a variety of tasks such as genre classification (Ferraro and Lemström 2018) and melodic pattern detection in Dutch folk music (Janssen, van Kranenburg, and Volk 2015). We are not interested in translations of patterns i.e. we do not consider two instances of the same progression in different keys as equal. For this reason we choose to use the SIA algorithm rather than the more specialised SIATEC.

### 3.2.1. Discovering maximal repeated patterns in multidimensional datasets

SIA represents scores as data collections,  $D$ , of two dimensions: onset time and pitch. Each score is made up of notes  $d_1, d_2, \dots, d_n$ . A pattern  $P$  is translatable by a vector  $v$  in a data collection  $D$  if and only if  $P$  can be translated by  $v$  to give a pattern that is a subset of  $D$ . Fundamental to SIA is the concept of the *maximal translatable pattern* (*MTP*). Formally, the *MTP* for a vector  $v$  in a data collection  $D$ , denoted by  $MTP(v, D)$ , is the largest pattern translatable by  $v$  in  $D$ . Mathematically:

$$MTP(v, D) = \{d \mid d \in D \wedge d + v \in D\}. \quad (4)$$

In music, *MTPs* often correspond to the patterns involved in perceptually significant repetitions. The main goal of SIA is to compute all non-empty *MTPs* in a data collection. The *MTP* for a vector,  $v$ , in a data collection,  $D$ , is defined as non-empty if and only if there exists at least two data points  $d_1$  and  $d_2$  in  $D$  such that  $v = d_2 - d_1$ . To sum up, SIA finds for every possible vector the largest pattern in the data collection that can be translated by that vector to give another pattern in the data collection. (Meredith, Lemström, and Wiggins 2002).

SIA is applied independently to each score of the dataset and the results grouped by *tab'* to generate a list of relevant patterns on a *tab'* level. Patterns that consist of notes that are not consecutive in the score are discarded so as to make the results comparable with the other algorithms used in this paper that do not have the ability of finding patterns of non-consecutive notes.

### 3.3. Maximally General Distinctive Patterns (MGDP)

Maximally general distinctive pattern mining (MGDP) (Conklin 2010) is a method designed to find general patterns that are also distinctive. A *distinctive pattern* is one that is frequent and over-represented in a positive corpus (here, scores in a particular *tab'*) as compared to an anticorpus (all other scores). The degree of over-representation of a pattern  $\Phi$  is measured by its relative frequency between the positive corpus (denoted  $\oplus$ ) and anticorpus (denoted  $\ominus$ ):

$$\Delta = \frac{P(\Phi \mid \oplus)}{P(\Phi \mid \ominus)}$$

In the above,  $P(\Phi \mid \oplus)$  and  $P(\Phi \mid \ominus)$  are the relative frequencies of the pattern  $\Phi$  in the positive corpus and anticorpus, respectively.

A *maximally general* distinctive pattern is a pattern with  $\Delta \geq \epsilon$  and for which no containing pattern is also distinctive. In this paper  $\epsilon = 3$  for distinctiveness is used, signifying a three-fold or higher over-representation in the positive corpus. To find maximally general distinctive patterns, a containment hierarchy of patterns is traversed from general to specific, terminating the search at a branch when a distinctive pattern is encountered.

### 3.4. Pattern selection

To take all patterns returned by any of our three methods would result in an unmanageable quantity of characteristic patterns. Two pattern selection processes are applied in each approach.

### 3.4.1. Minimum Frequency of Occurrence Threshold (MFO)

As noted by Nuttall et al. (2019), many longer, rarer patterns (that occur infrequently across the whole dataset) have a disproportionately high *importance* as determined by each of our models, this in reality is a reflection of infrequency of occurrence, rather than correlation between occurrence and *tab*<sup>‘</sup>. We consider the infrequency of such patterns as not providing us with enough *signal* to have confidence in this *importance* and like Nuttall et al. (2019), impose a *minimum frequency of occurrence* (MFO) for each pattern per score per *tab*<sup>‘</sup>.

Our *minimum frequency of occurrence threshold* of 59 per *tab*<sup>‘</sup> dictates that for a particular *tab*<sup>‘</sup>, a pattern is discarded if the number of occurrences of that pattern in scores annotated with that *tab*<sup>‘</sup> divided by the number of scores annotated with that *tab*<sup>‘</sup> is less than the minimum frequency of occurrence, 59. This threshold is applied to all three methods and determined by maximizing the F<sub>1</sub> score (harmonic mean of recall and precision) when evaluated on the *centos* provided by Chaachoo (2019). It is important to note that although the *centos* provided by Chaachoo are not necessarily a canonical ground truth, they are the most useful guide we have in parameter selection and that many of Chaachoo’s patterns themselves do not reach this minimum frequency. Indeed some patterns proposed by him do not exist at all in the dataset (more on this in Section 5). It is also true that by imposing such a threshold we limit our output to patterns of shorter lengths (since longer ones are less common), this is understood and considered a necessary limitation of the pattern-space we want to explore.

### 3.4.2. Pattern character

The *centos* as outlined by Chaachoo (2019) are monophonic sequences of notes, each with at least two unique pitches, without duration or octave and length ranging between and including 3 and 7 notes. As such we limit our search to operate within this space by

- Considering only patterns between and including lengths of 3 and 7 notes
- Removing duration and octave from the data representation
- Not considering as a pattern any sequence of notes that include a rest
- Discarding “patterns” that consist of only one unique pitch

The code to explore the dataset using some of our approaches can be found in the project repository on Github.<sup>6</sup>

## 4. Results

Each model described in Section 3 is applied to our dataset. The full output of each can be found in the results directory of the accompanying Github repository.

Table 2 displays the recall and precision of our three methods combined on a *tab*<sup>‘</sup> level, **# Centos** indicate how many of Chaachoo’s *centos* were available to find and **# Retrieved** is the number of unique patterns returned by all models. We consider Chaachoo’s *centos* that occur more frequently than our *minimum frequency of occurrence* as ground truth. Also displayed are the number of Chaachoo’s *centos* above this threshold for each *tab*<sup>‘</sup> and how many in total were discovered by our methods.

In total we find 186 significant patterns across all *tubū*<sup>‘</sup>, 42 of which correspond to the 45 of Chaachoo’s above the MFO threshold. Resulting in a total recall of 0.93 and total

<sup>6</sup><https://github.com/centonization/centonizationtheory/>

Table 2. Evaluation on Chaachoo’s Patterns above Minimum Frequency Occurrence.

<i>ṭab‘</i>	Recall	Precision	# Centos	# Retrieved
<i>al-istihlāl</i>	1.00	0.26	6	23
<i>al-iṣbahān</i>	0.83	0.33	6	15
<i>al-mašriqī</i>	0.75	0.30	4	10
<i>al-ḥiḡāz al-mašriqī</i>	n/a	0.00	0	15
<i>al-māya</i>	1.00	0.21	3	14
<i>al-raṣd</i>	1.00	0.25	4	16
<i>al-ḥiḡāz al-kabīr</i>	1.00	0.19	3	16
<i>al-ṣīka</i>	1.00	0.25	1	4
<i>al-‘uṣṣāq</i>	1.00	0.22	2	9
<i>garībat al-ḡusayn</i>	1.00	0.25	4	16
<i>raml al-māya</i>	0.75	0.23	4	13
<i>raṣd al-dāyl</i>	1.00	0.17	3	18
<i>‘irāq al-‘aḡam</i>	1.00	0.29	5	17
<b>total</b>	<b>0.93</b>	<b>0.25</b>	<b>45</b>	<b>186</b>

precision of 0.25. A full breakdown of results can be found in the Github repository.

#### 4.1. Distinctive patterns

One of the interesting features of the MGDGP approach (Section 3.3) is that it offers us an intuitive comparison of a pattern’s relative probability within a *ṭab‘* and outside of a *ṭab‘* (distinctiveness  $\Delta$ , Section 3.3). Table 3 presents these deltas for Chaachoo’s patterns to reflect their ability to be recalled when mining for all patterns subject to our selection processes in Section 3.4. The table is sorted from high to low distinctiveness,  $\Delta$ . For every pattern, its global rank (the position when all patterns for all classes are concatenated together) and its local rank (position within the patterns for the particular target) are indicated. These results indicate that of the 45 *centos* occurring above the minimum frequency threshold, 8 (at the bottom of Table 3) have a  $\Delta$  below 1.00 which indicates that they are in fact under-represented in the particular *ṭab‘*.

## 5. Discussion

The exploratory study described in this paper and particularly the obtained results presented in the previous section produce valuable contributions to the deeper understanding of the analysed dataset, the selected methods and the very task itself from both a musicological and technical perspective. This work has also given us the opportunity to contact Chaachoo, who shared with us his thoughts on the research carried out and its output. It is beyond the scope of this paper to engage in a detailed discussion of the obtained results for each *ṭab‘*. However, we will highlight in the current section the major outcomes of this process.

In previous studies (Sordo, Chaachoo, and Serra 2014; Caro Repetto et al. 2018), the Music Scores Collection of the Arab-Andalusian Music Corpus has been analysed according to the five criteria proposed by Serra (2014). Here we focus on its coverage for the specific task of melodic pattern analysis. Table 1 shows that the collection covers 13 of the 26 classical *ṭabū‘* and not all of them equally (since the distribution of scores across *ṭabū‘* is uneven). Most remarkable is the case of the *ṭab‘ al-ṣīka*, for which only one score is available, an exceptionally short one, corresponding to a recording of 7’13”. Consequently, only one of the 6 *centos* proposed by Chaachoo for this *ṭab‘* is found above

Table 3. Distinctiveness of Chaachoo's *centos*.

<i>ṭab'</i>	pattern	$\Delta$	rank	
			global	local
<i>al-ḥijāz al-kabīr</i>	AGF#E-D	62.58	5	5
<i>'irāq al-'aḡam</i>	GF#ED	13.14	11	3
<i>al-ḥijāz al-kabīr</i>	F#GA	7.25	20	11
<i>'irāq al-'aḡam</i>	EF#G	6.75	21	5
<i>al-rasd</i>	F#GA	5.52	25	7
<i>al-māya</i>	B-AG	4.01	35	3
<i>al-rasd</i>	AGE	3.99	36	9
<i>'irāq al-'aḡam</i>	BAG	2.71	46	9
<i>al-isbahān</i>	BCD	2.57	51	4
<i>rasd al-dāyl</i>	CDE	2.31	55	5
<i>al-ḥijāz al-kabīr</i>	CBAG	2.31	57	13
<i>al-istihlāl</i>	FAG	2.18	66	2
<i>al-istihlāl</i>	ABC	2.09	71	3
<i>al-'uṣṣāq</i>	BAG	2.02	75	1
<i>al-ṣīka</i>	AGFE	1.89	82	2
<i>al-māšriqī</i>	FGA	1.88	84	2
<i>raml al-māya</i>	FGA	1.84	88	1
<i>al-māya</i>	EFG	1.78	96	6
<i>rasd al-dāyl</i>	EDC	1.57	111	16
<i>garībat al-ḥusayn</i>	CDE	1.57	112	3
<i>al-istihlāl</i>	FEDC	1.52	118	8
<i>al-isbahān</i>	FEDC	1.48	125	12
<i>al-istihlāl</i>	GFE	1.46	127	10
<i>al-isbahān</i>	FED	1.46	128	13
<i>rasd al-dāyl</i>	AGFE	1.45	131	20
<i>al-rasd</i>	CDE	1.44	134	18
<i>al-isbahān</i>	EFG	1.40	141	16
<i>al-māya</i>	AGFE	1.36	150	8
<i>garībat al-ḥusayn</i>	AGFE	1.34	154	7
<i>al-istihlāl</i>	CBAG	1.29	163	20
<i>al-istihlāl</i>	EFG	1.20	181	22
<i>al-māšriqī</i>	AGF	1.19	182	4
<i>garībat al-ḥusayn</i>	FED	1.16	190	10
<i>al-isbahān</i>	GFE	1.13	194	20
<i>garībat al-ḥusayn</i>	AGF	1.12	198	12
<i>raml al-māya</i>	EFG	1.03	215	13
<i>raml al-māya</i>	FED	1.00	218	14
<i>al-māšriqī</i>	FED	0.98	223	7
<i>al-rasd</i>	EDC	0.96	225	19
<i>al-isbahān</i>	AGFE	0.93	230	22
<i>al-'uṣṣāq</i>	FED	0.87	236	9
<i>raml al-māya</i>	FEDC	0.84	239	15
<i>al-māšriqī</i>	FEDC	0.80	241	11
<i>'irāq al-'aḡam</i>	EDC	0.58	254	17
<i>'irāq al-'aḡam</i>	FED	0.48	258	19

the MFO (see Section 3.4), and two of them are not present at all in the available score.<sup>7</sup> Also worth noting is the case of *al-ḥijāz al-māšriqī*, for which 5 scores are available (audio duration of 3h2'25"), but none of the 3 *centos* proposed by Chaachoo are present above the MFO.<sup>7</sup> Despite the direct implications that this has for our applied methods, there are still interesting outcomes. In the case of *al-ṣīka* both TF-IDF and SIA found the only *cento* which occurs above the MFO, AGFE. This pattern has a local rank in terms of delta of 2 (Table 3), and it ranks 3rd in TF-IDF and 2nd in SIA. This result hence contributes to consolidate its significance for this *ṭab'*, as confirmed by Chaachoo.

A fundamental contribution of using computational methods for musicological research is the opportunity to make implicit knowledge from musicians or expert *connoisseurs* of the studied music tradition explicit. This is precisely the case for our study, whose

<sup>7</sup> See [https://github.com/centonization/centonizationtheory/blob/main/results/representative\\_patterns.pdf](https://github.com/centonization/centonizationtheory/blob/main/results/representative_patterns.pdf)

discussion with Chaachoo highlighted knowledge that was only briefly mentioned in his publications (if at all) but that is essential for the goals addressed in our research.

Although Chaachoo acknowledges the usefulness of computational analysis in the revision of his theory, it was apparent from our discussion that his main source for evaluation is his own experiential knowledge and, as he stated, “intuition”. Therefore, in order to evaluate the importance of the patterns retrieved from the algorithms, his only criterion was his experience with the music - the associated metrics reported by the used methods did not contribute to his evaluation. The inspection of the retrieved patterns gave Chaachoo the opportunity to specify the existence of melodic patterns in the *al-Āla* tradition that are not considered as defining of the *tubū‘*, that is, that can not be considered *centos*, and therefore are not theorized in his publications. If *centos* are “minimal recurrent units” used as “building blocks” of Arab-Andalusian melodies, *melodic formulas* are recurrent combinations of *centos*. These are mentioned in Chaachoo’s publications only briefly and not studied in detail (2016, 221-22). The melodies of the specific pieces, besides *centos* and *melodic formulas* also contain piece specific material (see for example the melody not contained in boxes in Figure 1). Chaachoo observed that some of the retrieved patterns correspond to joining points of the two *centos* of a *melodic formula*, that is, the last note(s) and the first one(s) of two frequently joined *centos*, or the beginning or ending of a *cento* together with a note that commonly precedes or follows that *cento*. Besides, some of the retrieved patterns correspond, according to Chaachoo, to ornamentation preferences of the particular performing orchestra, a type of pattern not specified in his publications. And finally, Chaachoo also recognized some patterns that he would describe as “expressive trends” related to the tradition as a whole. The existence of such *non-cento* patterns was unknown by the authors, and their elicitation and definition is an outcome of the discussion precipitated by the study presented here.

In order to gain a more concrete understanding of Chaachoo’s evaluation of our results, we discussed the case of *al-ḥiḡāz al-kabīr* (Figure 1 offers a short example). Although 10 scores are available for *al-ḥiḡāz al-kabīr* (audio duration of 5h16’33”), only three out of the five *centos* from Chaachoo occur over the MFO. Among them, the pattern AGF#E-D is the most distinctive of Chaachoo’s *centos* as shown in Table 3. The distinctiveness of this pattern is explained by the fact that it is one of the few that contains what Chaachoo defines as the *al-Zayadan* genre, that is, an intervallic sequence of minor second, augmented second and minor second. The very name of the *ṭab‘*, literally “the great *ḥiḡāz*,” suggests the relevance of this sequence, since *ḥiḡāz* is the name of a relevant Arabic *maqam*, also characterised by this sequence. This *al-Zayadan* genre confers to these *tubū‘* their very characteristic “oriental flavour”. However, Chaachoo also discusses how the importance of this sequence has been overemphasized recently and extensively used in improvised solos, but in fact does not appear in traditional instrumental preludes and many *ṣanā‘i‘* (2019, 257). Considering this last observation, we asked Chaachoo if the fact that only one of the tested methods retrieved the *cento* as it is proposed by him, but that many substrings of it are retrieved by the three methods around the characteristic augmented second might invite him to reconsider the contour of this *cento*. He maintains that he is certain of the relevance of this *cento* in its current contour and proposes possible reasons for it not being retrieved with higher relevance by our methods.

The discussion of the results related to *al-ḥiḡāz al-kabīr* resulted in a broader discussion on the concept and function of *centos* themselves. None of the results obtained in our study, even when a particular *cento* proposed by Chaachoo is not retrieved by any of the methods, is sufficient to disprove their significance. First of all, our dataset does not exhaustively cover the reality of the Arab-Andalusian music tradition, as does the knowledge upon which Chaachoo’s theory is built. Beyond the coverage of our dataset,

Table 4. *Centos* and retrieved patterns for *al-ḥiḡāz al-kabīr*, those of Chaachoo’s centones that do not occur above the MFO are in parenthesis. Retrieved patterns in bold are exact matches to Chaachoo’s, those underlined are substrings and those italicized are superstrings.

Patterns for <i>al-ḥiḡāz al-kabīr</i>	
Chaachoo’s <i>centos</i>	AGF#E-D, F#GA, CBAG, (CED, FED)
TF-IDF	<u>GF#E-</u> , F#E-D, GF#E-D, <u>AGF#E-</u> , <u>AGF#</u> , <b>AGF#E-D</b> , <u>GAGF#</u> , <u>F#GAG</u> , <b>F#GA</b> , GAG, EF#G
SIA	<u>AGF#</u> , GAG, <b>F#GA</b> , <u>BAG</u> , <u>CBA</u> , EF#G, <b>CBAG</b> , DCB
MGDP	<u>GF#E-</u> , F#E-D, <u>AGF#</u> , DEF#, <b>F#GA</b> , EF#G

the contained music scores represent, as confirmed by Chaachoo, an intermediate level between the performative, sonic surface and the compositional, theoretical melodic skeleton, which is the deep level on which *centos* operate. In the intermediate level represented in the scores, *centos* might be “hidden”, as Chaachoo put it, behind an ornament produced by the aesthetic preferences of an orchestra or a more general expressive trend. This might point to a discrepancy between practice and theory, a fact that is known and commonly studied in other music traditions, such as in Turkish *makam* music (Bozkurt, Ayangil, and Holzapfel 2014) or in South Indian *rāga* music (Pearson 2016). Secondly, it should be considered that *centos* are part of a broader system, the *ṭabʿ*, and therefore, their relevance might arise from interaction with other elements of this system. In our discussion, Chaachoo emphasized how *centos* are directed to stress one of the relevant degrees of the *ṭabʿ*. For example, the above discussed *cento* for *al-ḥiḡāz al-kabīr*, AGF#E-D, not only includes the characteristic *al-Zayadan* genre, but also represents a movement from the persistent degree of the *ṭabʿ*, A, to its fundamental degree, D. Equally, the two *centos* below the minimum frequency threshold, CED and FED (see Figure 1 for the latter), are melodic movements towards the fundamental degree, a fact that reinforces their significance. Furthermore, looking at the context even more broadly, a melodic pattern might not only be relevant because of its statistical occurrence, but because of its structural location (as pointed out by Volk and van Kranenburg 2001). Asked about this issue in Arab-Andalusian music, Chaachoo confirmed that *centos* tend to be used to conclude poetic/melodic phrases, and some of them are very characteristic for their use in concluding *ṣanāʿiʿ*, as is the case of the discussed *cento* AGF#E-D of *al-ḥiḡāz al-kabīr*. All these reflections offer very important contributions for improving the automatic analysis of Arab-Andalusian *centos* in future work.

It is also worth reflecting on the three methods used in our study, inspecting the differences in patterns retrieved by each of them for *al-ḥiḡāz al-kabīr* (Table 4). Regarding Chaachoo’s *cento* AGF#E-D, the full pentagram is only retrieved as such by TF-IDF (ranked 5th). The TF-IDF method also retrieves the 3 trigram and 2 tetragram substrings of this pattern, leading to redundancy in the output. The MGDP retrieves the 3 trigram substrings and SIA one. This recalls the fact that TF-IDF does not necessarily find minimal patterns, whereas MGDP is explicitly instructed not to consider longer patterns if a shorter one is already distinctive, and SIA finds minimal patterns if they are not part of a longer pattern with the same frequency. However, SIA finds only one of the trigram substrings due to an artefact of the post-filtering process that has been applied (see Section 3.2), removing all non-contiguous MTP from the result set and removing patterns of length greater than 7. It turns out that the substrings GF#E- and F#E-D always appear, in this particular *ṭabʿ*, within a longer frequent pattern that is non-contiguous, and/or one that is longer than the maximum length limit.

The *centos* CED and FED (bracketed in Table 4) are missed by all methods as they

occur below the minimum frequency. The SIA method alone finds the *cento* CBAG, split into the two overlapping trigrams CBA and BAG, but these are found by neither TF-IDF nor MGDP. For MGDP none of its substrings are distinctive above  $\epsilon = 3$  (CBA:  $\Delta = 2.10$ ; BAG: 2.78; CBAG: 2.31). These patterns can thus be found with a slight reduction of the distinctiveness threshold though with a larger set returned overall. The three patterns are discarded by TF-IDF, also due to low scores, all effectively zero.

Regarding some of the patterns which are not *centos* but are reported by one or more methods, SIA and TF-IDF report the pattern GAG which is frequent but not reported by MGDP, due to its  $\Delta$  value of 1.74, below the minimum  $\epsilon = 3.0$ . The TF-IDF method also finds the extended GAGF# while MGDP and SIA find its suffix AGF#. As discussed above, all methods find the novel pattern EF#G. MGDP returns uniquely the pattern DEF#, and SIA uniquely the pattern DCB. These findings might correspond to recurrent *non-cento* patterns, such as orchestra performance preferences or expressive trends, and therefore open a path of further exploration.

## 6. Conclusions and future work

In this paper we explore how computational methods can contribute to musicological research by focusing on the automatic analysis of melodic patterns in a music tradition that is underrepresented in such studies, as is the case of Arab-Andalusian music. Drawing on theoretical formulations that are still in development by an expert performer and researcher of this tradition, Amin Chaachoo, we have tested three algorithms with the aim of learning how they perform in this specific music tradition and how the obtained results can contribute to the development of Chaachoo's theory. To implement this study, we have created a dataset of 145 machine readable scores gathered from the Music Scores Collection of the CompMusic Arab-Andalusian Music Corpus, the largest source of machine readable data for the computational study of this music tradition.

From a musicological point of view, our research has helped deepen our understanding of the dataset, both in terms of its coverage and what the scores represent. Furthermore, through a discussion of the results with Chaachoo, we have been able to elicit theoretical information that was not completely developed in state of the art literature to inform the research presented here. Of special relevance amongst these outcomes is the identification of recurrent patterns that are not considered *centos*, such as orchestra preference and expressive trends. According to this finding, in future research a search for recurrent patterns per orchestra should be performed, as well as on the whole dataset and per fundamental degree of the *tubū'*, in order to find patterns related to expressive trends.

From a technical point of view, our methods are able to identify many of the frequent *centos* proposed by Chaachoo. It should be highlighted that in our research the list of these *centos* can not be taken as a strict ground truth, but as a guide for studying the performance of the implemented technologies. This is in fact an interesting opportunity for developing strategies in which computational methods are used for exploratory analyses and discovery tasks such as the ones commonly addressed by musicologists, and is an element which we will need to emphasize in future work. Equally, from the discussions with Chaachoo, it seems that methods based on frequency are not sufficient for retrieving musically meaningful patterns, and that some domain knowledge should be integrated in the task. Achieving this integration is an important target for future research.

## Acknowledgements

The authors would like to thank Amin Chaachoo for his valuable work in creating and annotating the Arab-Andalusian Music Corpus and especially for his discussion of our results and the contributions to this paper that arose from that. We would also like to thank the CompMusic project for making the Arab-Andalusian Music Corpus publicly available, and Antoni Abelló Sanz for his work in manually segmenting the Music Scores Collection. Finally we express gratitude to the reviewers for their well-considered and constructive feedback, crucial to the improving of this work.

## ORCID

Do not change this. Production will take care of it if the paper is accepted.

## References

- Apel, W. 1958. *Gregorian Chant*. A Midland book. Indiana University Press.
- Bozkurt, Barış, Ruhi Ayangil, and Andre Holzapfel. 2014. “Computational analysis of Turkish makam music: review of state-of-the-art and challenges.” *Journal of New Music Research* 43 (1): 3–23. <http://hdl.handle.net/10230/25935>.
- Caro Repetto, Rafael, Niccolo Pretto, Amin Chaachoo, Barış Bozkurt, and Xavier Serra. 2018. “An open corpus for the computational research of Arab-Andalusian music.” In *5th International Conference on Digital Libraries for Musicology (DLfM 2018)*, Paris, France, 28/09/2018, 78–86. <http://hdl.handle.net/10230/35470>.
- Chaachoo, A. 2011. *La música andalusí Al-Ála: Historia, conceptos y teoría musical*. Córdoba: Editorial Almuzara.
- Chaachoo, A. 2016. *La musique hispano-arabe, al-Ála*. Univers musical. Editions L’Harmattan.
- Chaachoo, A. 2019. *Al-qawā'id al-naẓariyya lil-mūsīqā al-'andalusiyya al-maġribiyya, al-'Ála (Theoretical principles of the Andalusian music from Morocco, al-Ála)*. Maṭḥaa al-Kalīj al-'Arabiyy.
- Chewand, G., and J. W. McKinnon. 2001. “Centonization.” *Oxford Music Online*.
- Conklin, D. 2010. “Discovery of distinctive patterns in music.” *Intelligent Data Analysis* 14 (5): 547–554.
- Ferraro, Andres, and Kjell Lemström. 2018. “On Large-Scale Genre Classification in Symbolically Encoded Music by Automatic Identification of Repeating Patterns.” In *Proceedings of the 5th International Conference on Digital Libraries for Musicology, DLfM '18*, New York, NY, USA, 34–37. Association for Computing Machinery. <https://doi.org/10.1145/3273024.3273035>.
- Ferretti, P., and A. Agaësse. 1938. *Esthétique grégorienne ou Traité des formes musicales du chant grégorien. Volume I. Traduit de l'italien par Dom A. Agaësse*. Desclée.
- Guettat, M. 2000. “La musique arabo-andalouse, l’empreinte du Maghreb.” 560.
- Janssen, B., W. De Haas, A. Volk, and P. Van Kranenburg. 2013. “Discovering repeated patterns in music: state of knowledge, challenges, perspectives.” In *Proc. of the 10th International Symposium on Computer Music Multidisciplinary Research, Marseille, France*, Vol. 20Vol. 20, 74.
- Janssen, Berit, Peter van Kranenburg, and Anja Volk. 2015. “A comparison of symbolic similarity measures for finding occurrences of melodic segments.” In *Proceedings of the 16th ISMIR Conference, Málaga, Spain, October 26-30, 2015*, 659–665. ISMIR press.
- Meredith, David, Kjell Lemström, and Geraint A. Wiggins. 2002. “Algorithms for discovering repeated patterns in multidimensional representations of polyphonic music.” *Journal of New Music Research* 31 (4): 321–345. <https://doi.org/10.1076/jnmr.31.4.321.14162>.
- Meredith, Dave, and Geraint A. Wiggins. 2001. “Pattern induction and matching in polyphonic music and other multi-dimensional datasets.” In *In Callaos*, 61–66.
- Nuttall, Thomas, Miguel García-Casado, Víctor Núñez-Tarifa, Rafael Caro Repetto, and Xavier Serra. 2019. “Contributing to new musicological theories with computational methods: The case of centonization in Arab-Andalusian music.” In *20th Conference of the International Society for Music Information Retrieval*, Delft, The Netherlands, 04/11/2019, 223–228. <https://repositori.upf.edu/handle/10230/42789>.

- Pearson, Lara. 2016. “Coarticulation and Gesture: an Analysis of Melodic Movement in South Indian Raga Performance.” *Music Analysis* 35 (3): 280–313. <https://doi.org/10.1111/musa.12071>.
- Poché, C. 1997. *La música arábigo-andaluza (con CD)*. Músicas del mundo. Ediciones Akal.
- Porter, A., M. Sordo, and X. Serra. 2013. “Dunya: A System for Browsing Audio Music Collections Exploiting Cultural Context.” In *14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, Curitiba, Brazil, 04/11/2013, 101–106. <http://hdl.handle.net/10230/32251>.
- Ren, IY, HV Koops, A Volk, and W Swierstra. 2017. “In Search of the Consensus Among Musical Pattern Discovery Algorithms.” In *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China, October 23-27, 2017*, edited by Sally Jo Cunningham, Zhiyao Duan, Xiao Hu, and Douglas Turnbull, 671–678. [https://ismir2017.smcnus.org/wp-content/uploads/2017/10/120\\_Paper.pdf](https://ismir2017.smcnus.org/wp-content/uploads/2017/10/120_Paper.pdf).
- Serra, Xavier. 2014. “Creating Research Corpora for the Computational Study of Music: the case of the CompMusic Project.” In *AES 53rd International Conference on Semantic Audio*, 26/01/2014, 1–9. AES, AES. <http://hdl.handle.net/10230/44221>.
- Sordo, Mohamed, Amin Chaachoo, and Xavier Serra. 2014. “Creating Corpora for Computational Research in Arab-Andalusian Music.” In *Proceedings of the 1st International Workshop on Digital Libraries for Musicology, DLFM '14*, New York, NY, USA, 1–3. Association for Computing Machinery. <https://doi.org/10.1145/2660168.2660182>.
- Volk, Anja, and Peter van Kranenburg. 2001. “Melodic similarity among folk songs: An annotation study on similarity-based categorization in music.” *Musicae Scientiae* 16 (3): 317–339.