An open corpus for the computational research of Arab-Andalusian music

Rafael Caro Repetto
Music Technology Group,
Universitat Pompeu Fabra
Barcelona, Spain
rafael.caro@upf.edu

Niccolò Pretto
Department of Information
Engineering, University of Padova
Padova, Italy
niccolo.pretto@dei.unipd.it

Amin Chaachoo
Asmir Center for
Musicological Research
Tetouan, Morocco
chaachooamin@gmail.com

Barış Bozkurt
Music Technology Group,
Universitat Pompeu Fabra
Barcelona, Spain
baris.bozkurt@upf.edu

Xavier Serra
Music Technology Group,
Universitat Pompeu Fabra
Barcelona, Spain
xavier.serra@upf.edu

ABSTRACT

In the medieval Islamic territories of the Iberian Peninsula known as Al-Andalus a unique style of music was formed combining local practices with Arab sensibilities. After the fall of the last Andalusian kingdom, this classical repertoire has been preserved to the present in North African countries. The idiosyncrasies of this repertoire, which combines musical traits from Western and Eastern Mediterranean traditions in orchestral and choral settings, as well as instrumental and vocal solos, deserves an in depth musicological study, that can benefit from computational tools for corpus-driven research. On the other hand, the characteristics of this music poses interesting challenges to MIR methods and therefore offer new research opportunities to this field. To address these topics, we present here the first complete release of the corpus for the research of the Moroccan tradition of Arab-Andalusian music built in the framework of the CompMusic project. The corpus comprises three data collections, namely audio recordings, music scores and lyrics, as well as related annotations and metadata. We also present a series of Jupyter Notebooks for browsing and retrieving data from the corpus. Both the corpus and notebooks are completely open to the research community.

CCS CONCEPTS

• **Information systems → Digital libraries and archives**;

KEYWORDS

Research Corpus, Arab-Andalusian Music

ACM Reference Format:
Rafael Caro Repetto, Niccolò Pretto, Amin Chaachoo, Barış Bozkurt, and Xavier Serra. 2018. An open corpus for the computational research of Arab-Andalusian music. In *Proceedings of 5th International Conference on Digital Libraries for Musicology (DLfM 2018).* ACM, New York, NY, USA, 9 pages. https: //doi.org/10.1145/3273024.3273025

DLfM '18, September 28, 2018, Paris, France

ACM ISBN 978-1-4503-6522-2/18/09…$15.00
https://doi.org/10.1145/3273024.3273025

1 INTRODUCTION

The creation of a solid corpus is essential for the implementation of data-driven research tasks. In the CompMusic project [12, 14], a large amount of effort has been invested in gathering, curating, organizing and making accessible research corpora for each of the music traditions studied [3, 16, 17]. During this process, we have realized that corpus creation is a research task in itself. To address this task hence, we designed a methodology within the project, presented in [13], which provides a global framework at the time that allows for culture specificity. One of the most important outcomes of this reflection is the distinction between research corpus and test corpus. The latter, which we refer to as dataset, is a fixed collection of data and related annotations put together for the implementation—and future reproduction—of a specific research task or experiment. Contrastingly, a research corpus is a collection of data, metadata and annotations gathered for a broader research goal and that is ever growing and perfected. The creation of the Arab-Andalusian music research corpus started in 2013. The challenges this music tradition poses for the creation of a research corpus suitable for computational research was published previously [15], together with a description of its initial status. Since that time, the corpus has been growing, specially in the machine readable data related to music scores and lyrics. But also, benefiting from the experience from other corpora within the CompMusic project, an important effort has been dedicated to its coherent organization, the interrelation between the data, and specially to the development of tools that ease the access to the data in a customizable and culturally informed way.

Consequently, in this paper, we present the current status of the corpus, including the full collection of machine readable music scores (Section 3), a derived dataset (Section 4), and Jupyter Notebooks created for browsing and retrieving data from the corpus (Section 5). To better understand its content, we first describe the characteristics of Arab-Andalusian music (Section 2). The paper closes with some concluding remarks and hints to future work (Section 6).

Table 1: List of *nawabāt* present in the corpus. The non-canonical one is marked with *.

Name	Transliterated name	Recordings	Duration
الاستهلال	*al-istihlāl*	22	17h 00' 45"
الاصبهان	*al-iṣbahān*	13	9h 27' 14"
الحجاز الكبير	*al-ḥiŷāz al-kabīr*	7	5h 21' 57"
الحجاز المشرقي	*al-ḥiŷāz al-mašriqī*	14	7h 12' 47"
الرصد	*al-raṣd*	10	7h 01' 57"
العشاق	*al-ʿuššāq*	7	4h 35' 26"
الماية	*al-māya*	12	9h 16' 06"
رصد الذيل	*raṣd al-ḏāyl*	17	12h 08' 18"
رمل الماية	*raml al-māya*	16	12h 43' 54"
عراق العجم	*ʿirāq al-ʿaŷam*	7	4h 08' 54"
غريبة الحسين	*garībat al-ḥusayn*	11	6h 54' 20"
بواكر الماية	*bawākir al-māya* *	2	1h 23' 16"
بدون	none	21	2h 56' 07"

2 ARAB-ANDALUSIAN MUSIC

Arab-Andalusian music is one of the terms[1] given to the music tradition that was formed around the 12th century in Al-Andalus, the territories of the Iberian Peninsula under Islamic rule during the Middle Ages. Even though this term includes the word "Arab", added to avoid confusion with the music from the contemporary Spanish region of Andalusia, this tradition is essentially an indigenous Iberian music, as shown by the fact that the characteristic microtones and *glissandi* of Arab music traditions are not proper of Arab-Andalusian music [4]. Notwithstanding this, due to political, religious and social reasons it received the influence from Arabic poetry and music, in the language and poetic forms used and the adoption of some melodic modes. With the migration of the Islamic Andalusian population to North Africa, this music has been preserved there to the present, and has become indigenized as classical traditions in Morocco, Tunisia and Algeria, developing local characteristics. The corpus here presented is formed by the Arab-Andalusian music tradition surviving today in Morocco, where it is known as al-Āla.[2]

Arab-Andalusian music is performed through *nawabāt* (plural of *nawba*), suites of instrumental and vocal compositions ordered according to their metrical mode in a sequence of increasing tempo. Originally, each *nawba* consisted of pieces composed in the same *ṭāb'* (plural *ṭubū'*), or melodic mode. A *ṭāb'* is characterized by a diatonic scale built upon a fundamental note, within a certain range and including two or three predominant degrees and one or two sustained degrees (similar in concept to the Gregorian reciting tone), as well as a specific collection of characteristic melodic motives [5]. Each *ṭāb'* is also associated to certain emotional and spiritual content [4]. During its historical evolution in the Maghreb, this tradition developed different local practices and styles in each of the countries where it was preserved, at the same time it was receiving the influence of folk and neighboring Arab music traditions. In the 18th century, the Moroccan al-Āla tradition fixed its repertoire in 26 *ṭubū'*, but only 11 *nawabāt*. Since the *nawabāt* for the reminding *ṭubū'* were preserved only in fragmentary status, their surviving pieces were added to the other ones, according to similarities in their *ṭubū'* [8]. Therefore, each *nawba* has a primary *ṭāb'*, and most of them have also secondary *ṭubū'*. Nowadays, orchestras of the al-Āla tradition have incorporated *ṭubū'* of folk origin into their repertoire, one of them even forming a new *nawba*, as well as *maqam* music from Eastern traditions. Table 1 shows the *nawabāt* contained in our corpus, and Table 2 the *ṭubū'* and related modal entities.

Table 2: List of *ṭubū'* present in the corpus. Noncanonical ones are marked with *.

Name	Transliterated name	Secs.	Duration
الاستهلال	*al-istihlāl*	146	17h 10' 34"
الاصبهان	*al-iṣbahān*	72	9h 38' 37"
الحجاز الكبير	*al-ḥiŷāz al-kabīr*	49	5h 41' 10"
الحجاز المشرقي	*al-ḥiŷāz al-mašriqī*	25	3h 09' 51"
الرصد	*al-raṣd*	55	7h 01' 57"
العشاق	*al-ʿuššāq*	40	4h 35' 26"
الماية	*al-māya*	73	9h 16' 06"
رصد الذيل	*raṣd al-ḏāyl*	89	12h 08' 18"
رمل الماية	*raml al-māya*	95	13h 11' 37"
عراق العجم	*ʿirāq al-ʿaŷam*	41	4h 08' 54"
غريبة الحسين	*garībat al-ḥusayn*	77	7h 07' 09"
بواكر الماية	*bawākir al-māya* *	11	1h 23' 16"
الصيكة	*al-ṣīka*	1	7' 11"
المشرقي	*al-mašriquī*	57	4h 08' 40"
الزركة	*al-zerga* *	6	31' 59"
قريب إلى الاستهلال	similar to *al-istihlāl* *	2	9' 41"
قريب إلى الماية	similar to *al-māya* *	2	11' 05"
رصد الذيل مغربي	Moroccan *raṣd al-ḏāyl* *	2	8' 29"
خليط الطبوع	mixed *ṭubū'*	1	7' 03"
النهاوند	*Nahawand* (*maqam*) *	3	13' 58"

Regarding the structure of the *nawba*, it can be understood as a sequence of sung poems known as *ṣanā'i'* (plural of *ṣan'a*), performed by mixed choir and orchestra in heterophonic structure. The instrumental ensemble differs in each orchestra, but is generally formed by string instruments such as ʿūd, rabāb, qānūn, violin, viola, cello, double bass and piano, percussion instruments such as tar and darbuka, and occasionally also a clarinet as a wind instrument. These *ṣanā'i'* can be preceded or interpolated by other pieces, either instrumental, both solo and orchestral, or vocal, both solo and choral. These pieces are grouped in performance according to each of the five rhythmic modes or *mawāzīn* (plural of *mīzān*) established in the tradition, and ordered according to the three possible renditions of the *mīzān* in increasing tempo. Since the performance of a whole *nawba* can last several hours, nowadays orchestras usually perform one *mīzān* of one *nawba*. Therefore, the three tempi in which the *mīzān* is performed, known as *muassa'*, *mahzūz* and *inṣirāf*, become structural sections and consequently also considered as forms. Table 3 shows the list of forms included

[1] For a discussion about the terminology for referring to this music tradition, please refer to [9]

[2] All the Arab terms are transcribed according to the standard proposed by [6], unless otherwise stated.

Table 3: List of forms present in the corpus.

Name	Transliterated name	Sections	Duration
إنشاد	*inšād*	75	2h 43' 52"
توشية	*tawāšī*	105	3h 06' 10"
مشالية	*mišālia*	168	4h 44' 06"
تقسيم	*taqsīm*	22	43' 54"
موال	*mawwāl*	16	1h 27' 12"
موسع	*muassa'*	138	46h 08' 07"
مهزوز	*mahzūz*	144	14h 37' 25"
انصراف	*inṣirāf*	178	25h 34' 58"
صامت رصد الذيل	Instr. *raṣd al-ḏayl*	1	5' 17"

Table 4: List of *mawāzīn* contained in the corpus. Noncanonical meters are marked with *.

Name	Transliterated name	Sections	Duration
بسيط	*basīṭ*	105	22h 16' 08"
بطايحي	*bṭāyḥī*	115	23h 30' 45"
درج	*darŷ*	17	2h 42' 26"
قائم ونصف	*qā'im wa niṣf*	97	14h 56' 07"
قدام	*quddām*	174	23h 46' 39"
ثنائي	binary *	50	2h 10' 25"
ثلاثي	ternary *	2	6' 09"
حضّاري	*ḥaddārī* *	1	10' 24"
ثنائي وبطايحي	binary and *bṭāyḥī* *	1	5' 17"
زنداني وقدام	*zandānī* and *quddām* *	5	50' 42"
بدون	none	280	9h 35' 59"

in our corpus. As in the case of the *ṭubū'*, besides the five classical *mawāzīn*, the orchestras also incorporated rhythms from folk traditions to their repertoire. The rhythmic modes covered in our corpus are listed in Table 4.

3 THE ARAB-ANDALUSIAN MUSIC RESEARCH CORPUS

The Arab-Andalusian music research corpus is formed by three data collections, namely the audio recordings collection, the music scores collection and the lyrics collection. Each of them was created according to the methodology established in the CompMusic project and presented in [13]. This method proposes five criteria for the creation and evaluation of a corpus, namely purpose, coverage, completeness, quality and re-usability. Derived from the principles of the CompMusic project, the general purpose of the whole corpus, and therefore shared by the three collections, consists in pushing forward the state of the art in music information retrieval and computational musicology by addressing the research challenges that this tradition poses to these fields, with a special focus on its melodic and rhythmic aspects. The three collections are closely interrelated, having the audio recording as main reference, since the music scores and lyrics are transcriptions of respectively their predominant melody and their sung lyrics. Each audio recording is given a MusicBrainz identifier (MBID), with which the related music score and lyrics are also identified.

One of the goals of the CompMusic project is to carry out culturally aware research. Therefore, the corpus for each of the music traditions studied was designed according to their musical characteristics. This work implies the collaboration of musicologists, MIR researchers, developers and musicians. The challenges that this task poses in the specific case of the Arab-Andalusian music research corpus, as well as the origin of the recordings, the difficulties for transcribing the music and the source for the lyrics, are discussed in [15].

Since the common purpose for the three collections of the corpus has been already stated, in the remainder of this section, we describe each collection according to the criteria of coverage, completeness, quality and re-usability.

3.1 Audio Recordings Collection

The audio recordings collection of the Arab-Andalusian corpus consists of 164 long recordings, covering more than 125 hours of music. The length of the recordings is variable, but the average duration is almost 45 minutes.

3.1.1 Coverage. The audio collection covers all the main instances of the four musical entities that characterize this music as described in Section 2, namely *nawba*, *ṭāb'*, form and *mīzān*. As shown in Table 1, the collection covers the eleven *nawabāt* of the al-Āla tradition, plus one *nawba* created more recently. Besides, 21 recordings are considered as not belonging to any of these *nawabāt*. Of the 26 *ṭubū'* of this tradition, the collection covers 13 of them, plus other melodic modes from different origins, as shown in Table 2. Table 3 lists the forms contained in the collection, which are all the forms considered in the tradition with the exception of the *bugia*. Finally, the five traditional *mawāzīn* are covered, as listed in Table 4, plus other rhythmic modes from different origins. Besides, 134 recordings contain sections annotated as not related to any *mīzān*. There is only one case of a recording containing two *nawabāt*.

3.1.2 Completeness. In order to organize the collection and its related metadata, we used MusicBrainz for two main reasons. First, it provides a unique MBID to each recording, so that we can link all the related data, such as music score, lyrics, computed features and metadata, to it. Secondly, it offers a framework for storing all the metadata related to each recording, which can be retrieved via the MusicBrainz API. Besides the name of the recording and corresponding release, the metadata include the name of the orchestra, the names of its director and musicians, as well as the instrument each of them plays. The metadata are stored in its original Arabic script, but automatically generated transliterations[3] are also provided.

As mentioned previously, each recording covers the performance of one *mīzān* for one particular *nawba*. However, in order to allow more specific and musically relevant analysis, each recording has been manually segmented by the third author, who also annotated the *nawba*, *ṭāb'*, form and *mīzān* of each section. Tables 1, 2, 3 and 4

[3]The code for these transliterations are based in the American Library Association - Library of Congress (ALA-LC) standard for Arabic [15]

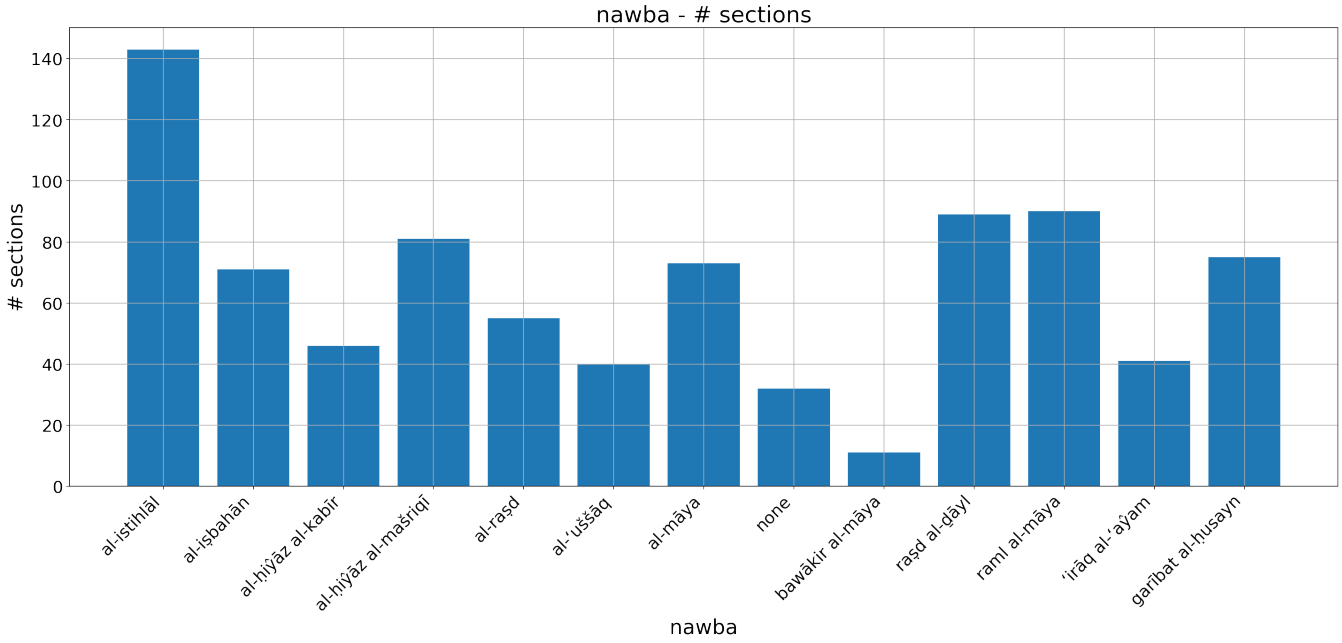

Figure 1: Bar chart showing the number of sections per *nawabāt* as plotted in the Metadata Jupyter Notebook.

show the set of annotations respectively used for each of these entities, the number of sections annotated with each of these labels—except for *nawabāt*, for which the number of recordings is shown, provided that each recording, with the one single exception previously mentioned, contains only one *nawba*—and the aggregated duration of the sections. Figure 1 shows the distribution of sections per *nawba*. Only six recordings in the corpus containing performances of folk repertoire have not been segmented and annotated.

From the audio collection, the authors extracted various features such as the fundamental frequency series and pitch distribution. In order to extract the first ones, we used an algorithm originally developed to analyze Turkish makam music [2]. The quality of the pitch estimation was confirmed by a visual inspection of several samples, by plotting the series together with the spectrograms. In general, the pitch estimation quality is high in most parts of the audio with occasional octave errors mainly during the low-pitch instrumental solos [11]. The pitch distributions are extracted from the pitch series using 7.5 cents resolution and smoothed using a Gaussian kernel with standard deviation of 7.5 cents. The latter operation and the long duration of the recordings cause a highly smooth distribution. All these features are also available together with the recordings collection.

3.1.3 Quality. The recordings gathered for the Arab-Andalusian corpus were chosen for their artistic quality and representativity of the masters recorded, so that, due to the historical circumstances of the tradition in the last decades, most of them were recorded in the 1960s and 1970s. Consequently the sound quality in occasions can be poor, with noisy and sporadically inaudible sections. In order to reduce storage capacity, the recordings are converted to mp3 audio format, sampled at 44.1 or 48 kHz. There is not a unique bit rate, however the recordings usually have at least 128 Kbps (with only six exceptions). The recordings can be either mono or stereo.

3.1.4 Re-Usability. All the recordings of the collection are copyright free and therefore can be freely distributed. The whole collection is available in the Internet Archive[4] and the metadata are accessible in MusicBrainz.[5]. To facilitate browsing the collection according to musically relevant concepts such as artist, *nawba* and *mīzān*, the CompMusic project developed the online tool Dunya[6] [10], which also offers an API for retrieving metadata and downloading data. In order to offer an even more customizable tool for accessing the collection for research purposes, we developed Jupyter Notebooks that will be described in the following section. Finally, all the data can be directly downloaded as a zip file from a Zenodo repository[7]. In this repository, each recording is stored in a folder named with its own MBID. Besides the mp3 file, each folder contains a JSON file with the editorial metadata from MusicBrainz, the transcribed score in XML format and two derived files, a TXT file, importable into the spectrogram tool of Sonic Visualizer[8], with the fundamental frequency series extracted from the audio recording, and the pitch distribution in JSON format. The overall size of the zip file is approximately 9 GB.

3.2 Music Scores Collection

We present here the complete collection of machine readable music scores for all the recordings in the audio collection. The scores

[4]https://archive.org/details/arabandalusian
[5]https://musicbrainz.org/collection/142ea0d7-7fdf-4ea5-9b04-219f68023d01
[6]https://dunya.compmusic.upf.edu/andalusian/
[7]https://doi.org/10.5281/zenodo.1291775
[8]https://www.sonicvisualiser.org/

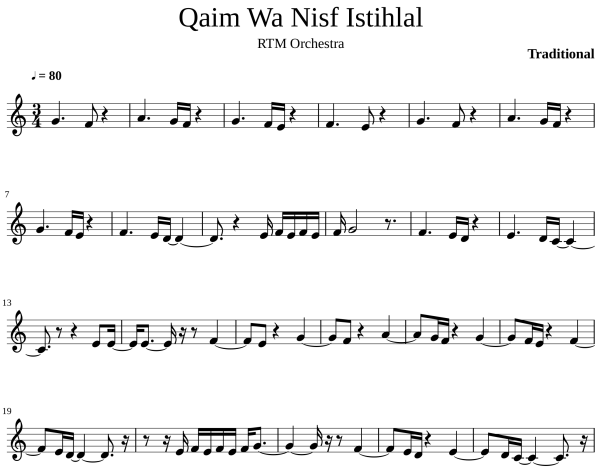

Figure 2: Excerpt of the music score for the recording with MBID e22549ae-4a0c-43ef-87f4-e0f81ed49d58, notating the performance of the *mīzān qā'im wa niṣf* of *nawba al-istihlāl*.

are manual transcriptions made by the third author, an expert musician and musicologist of this tradition. In doing these transcriptions, the focus is set on the stock of melodies transmitted orally that form the core of the al-Āla tradition and that, even though divergences across performances are common, are shared by all the orchestras. Therefore, improvisatory sections of instrumental and vocal solos have not been transcribed. The scores were created using MuseScore[9], and then exported in MusicXML format. The scores are linked with their corresponding recordings via the MBID.

3.2.1 Coverage. Since the music scores consist in transcriptions of the recordings contained in the audio collection, the coverage is the same (see Section 3.1.1). The only exceptions are the sections corresponding to the forms *inšād*, *taqsīm* and *mawwāl*, since their melody is improvised, as well as the six not segmented recordings for which the scores are not available.

3.2.2 Completeness. All the metadata regarding title, orchestra, director and musicians from the audio collection are applicable to the scores collection via the MBID. Due to the fact that Arabic is a right-to-left writing system, we have not found a satisfactory system to include lyrics in their original script in the scores. Therefore, the first line of each *ṣan'a* is annotated inside the score with a textual label. The scores are not annotated yet in terms of sections, a task which is in process at the time of writing this paper. Consequently, annotations regarding *nawba* and *mīzān*, which are unique per recording, can be directly applicable to the scores, and those about *ṭāb'* and form will be also applicable once the segmentation process is finished.

3.2.3 Quality. Since Arab-Andalusian music is performed in a heterophonic texture, the scores contain the predominant melody underlaying the different performance layers (see Figure 2), disregarding occasional octave changes by specific vocalists and instrumentalists, and changes between vocal and instrumental sections. They include tempo markings at the beginning and for each tempo change throughout the piece related to the performance in the transcribed audio recording.

3.2.4 Re-Usability. As in the case of the audio recordings collection, the music scores collection is completely open for research purposes, and accessible through different means. Firstly, they can be viewed and downloaded from the MuseScore website[10]. The website offers a player tool that allows playing the score and other interactions online. The description of each score contains a link to the corresponding recording in the Internet Archive and its entry in MusicBrainz. Equally, the scores can be searched by artist, *nawba* and *mīzān* in Dunya (see Section 3.1.4), where the score is displayed together with a player of the related recording. Then, the Jupyter Notebooks presented in the Section 5 allow to retrieve and download the scores according to more customizable research needs, as well as to visualize general statistics about the collection. Finally, they are included in the same Zenodo repository as the audio collection for downloading in a zip file (see Section 3.1.4).

3.3 Lyrics Collection

As explained in Section 2, the *ṣan'a* or sung poem is the fundamental form in *nawba* performance and the main structuring element. Indeed, these poems are considered as one of most valuable repertoires of Arab-Andalusian poetry. The relationship between lyrics, melody and emotional content associated with the corresponding *ṭāb'* and *mīzān* confers them a high value for musicological research.

3.3.1 Coverage. Although the lyrics for the *ṣanā'i'* of all the recordings of nine *nawabāt* in the corpus have been identified, at the moment of writing this paper the lyrics collection contains machine readable lyrics for the audio recordings of the *nawabāt al-iṣbahān*, *al-māya* and *raml al-māya*.

3.3.2 Completeness. The collection contains the lyrics both in its original Arabic script and in automatically computed transliterations according to the American Library Association - Library of Congress (ALA-LC) standard [15]. Although the lyrics are not aligned with the recordings, they are related to the corresponding recording section for better usability. For the same reason, each *ṣan'a* has been divided into lines and line sections, and stored both in TSV and JSON formats.

3.3.3 Quality. The lyrics sung in a particular performance might change due to several reasons, such as the school to which the orchestra belongs to, the social context of the performance or the personal criterion of the director of the orchestra. The lyrics contained in this collection have been manually typed by the third author, who listened to each recording and selected the corresponding *ṣan'a* as compiled by the musician and scholar Mehdi Chaachoo, whose compendium takes into account the current differences across schools within the al-Āla tradition.

[9]https://musescore.org

[10]https://musescore.com/mtg

3.3.4 Re-Usability. The lyrics collection, both in their original Arabic script and their transliterations, both in TSV and JSON formats, are available for download in Zenodo.[11]

3.4 Potential of the corpus

To the best of our knowledge, the Arab-Andalusian music research corpus here presented is to date the largest completely open resource for the study of this tradition by means of MIR methods. According to the general purpose for which the corpus was gathered, these data and their related annotations and extracted features offer a great potential for the analysis of the melodic and rhythmic elements specific to this music tradition, such as *nawba*, *ṭāb'* and *mīzān*. From a musicological point of view, the availability of all these data in machine readable format allows the computation of quantitative and statistical information, regarding pitch, intervals, predominant scales degrees, intonation profiles, note durations, meters, tempograms, etc. The fact that the corpus includes audio and symbolic representations of the same performances enhances the possibilities for such research.

From an MIR point of view, it also suggests interesting research tasks, such as predominant melodic extraction from a heterophonic source, fundamental degree detection, *nawba*, *ṭāb'* or *mīzān* automatic identification and classification, structural segmentation, tempo and meter estimation, etc. The manual annotations provide a valuable ground truth for these tasks. Audio to score alignment has been one of the research topics in the CompMusic project, and it is our plan to implement and adapt existing results [19] to this corpus. Benefiting from this availability of audio and scores, we already started working on the automatic identification of *nawabāt*, by matching templates computed from the music scores with the pitch distributions extracted from the audio recordings. In order to carry out this work, we created the dataset presented in the following section. The first results of this research were recently published in [11]. This music tradition is also especially interesting for the study of melodic motives, since many *ṭubū'* share similar fundamental and predominant degrees, scale profiles and pitch ranges, but differ in their characteristic melodic motives, which has been described in detail in [5], thus offering a valuable ground truth for this task. On the other hand, this corpus poses also challenges for the implementation of state of the art methods for such topics, such as working with very long recordings, which contain very different textures, from choir plus orchestra to instrumental or vocal soloists, and different audio qualities.

The lyrics collection offers a new window of possibilities. Lyrics to audio alignment has also been a research task in the CompMusic project, and the developed methods [7] will be also applied to this repertoire. The relationship between lyrics content and melodic features is also a promising study. We have previously applied Natural Language Processing methods to the analysis of such relationship [18], which could be applied to Arab-Andalusian music for the study of how melodic and rhythmic modes are selected according to the lyrics contents.

[11] https://doi.org/10.5281/zenodo.1291903

Table 5: Overall duration of the recordings for each *nawba* in the *nawabāt* detection dataset.

Nawba name	Duration
al-istihlāl	5h 41'
al-iṣbahān	6h 20'
al-ḥiŷāz al-kabīr	5h 22'
al-ḥiŷāz al-mašriqī	3h 37'
al-raṣd	5h 18'
al-'uššāq	4h 35'
al-māya	6h 01'
raṣd al-ḏāyl	5h 49'
raml al-māya	6h 23'
'irāq al-'aŷam	4h 13'
garībat al-ḥusayn	5h 26'

4 *NAWABĀT* DETECTION DATASET

In order to start exploiting the potential of the Arab-Andalusian music research corpus, among the many tasks that it allows as described in the previous section, the first one we undertook is the automatic recognition of *nawabāt*, whose methodology and results are published in [11]. In order to carry out this task, we extracted a dataset from the overall corpus containing 77 audio recordings. Those ones not belonging to a canonical *nawba* were excluded and the other ones were selected according to the criteria of audio quality and length of the recording. The only recording with two *nawabāt* in the the same audio file was excluded *a priori* . The audio quality criterion involved an auditory evaluation of several samples of each recordings. The second criterion—the length of the recording—is derived from music theory, in the sense that, if the performance of the *mīzān* of a *nawba* requires almost one hour, a track of few minutes can't adequately represent such a performance. Since *nawabāt al-ḥiŷāz al-kabīr*, *al-'uššāq* and *'irāq al-'aŷam* have only seven recordings each, and considering the necessity of a weighted dataset, these cases imposed the maximum number of recordings for each *nawba*. Therefore, the resulting dataset includes seven recordings for each of the eleven *nawabāt*. The overall duration of the recordings for each *nawba* is presented in Table 5.

The approach proposed in [11] relies on template matching applied to pitch distributions computed from audio recordings. The templates were synthesized from average pitch class distribution of several scores. The proposed method achieved an accuracy of 75% on the addressed task. The dataset is openly accessible and the experiment completely reproducible through the Jupyter Notebooks described in the following section.

5 NOTEBOOKS

In order to easily use and exploit the Arab-Andalusian corpus, a set of Jupyter Notebooks, written in Python, has been developed and shared in a public GitHub repository[12]. Furthermore, several visual tools composed by widgets[13] are provided allowing to visualize

[12] https://github.com/MTG/andalusian-corpus-notebooks
[13] https://ipywidgets.readthedocs.io/en/latest/

mizan / nawba	none	basīṭ	bṭāyḥī	ternary	binary	binary and bṭāyḥī	ḥaddārī	darǰ	zandānī and quddām	qā´im wa niṣf quddām
al-istihlāl	44	8	26	0	15	0	0	9	0	27	14
al-iṣbahān	26	5	12	0	1	0	0	0	0	15	12
al-ḥiǰāz al-kabīr	17	7	9	0	2	0	0	1	0	5	5
al-ḥiǰāz al-mašriqī	32	6	11	2	15	0	0	3	0	0	12
al-raṣd	19	5	9	0	0	0	0	0	0	0	22
al-ʿuššāq	13	5	2	0	2	0	0	0	0	9	9
al-māya	24	19	10	0	2	0	0	2	0	3	13
none	4	0	1	0	1	0	1	0	5	0	20
bawākir al-māya	2	0	1	0	0	0	0	0	0	0	8
raṣd al-ḏayl	31	19	8	0	1	1	0	0	0	14	15
raml al-māya	31	15	15	0	0	0	0	0	0	12	17
ʿirāq al-ʿaǰam	11	2	7	0	4	0	0	0	0	0	17
garībat al-ḥusayn	26	14	4	0	7	0	0	2	0	12	10

Figure 3: Table showing the number of sections per *nawba* and *mīzān* as plotted in the Metadata Jupyter Notebook.

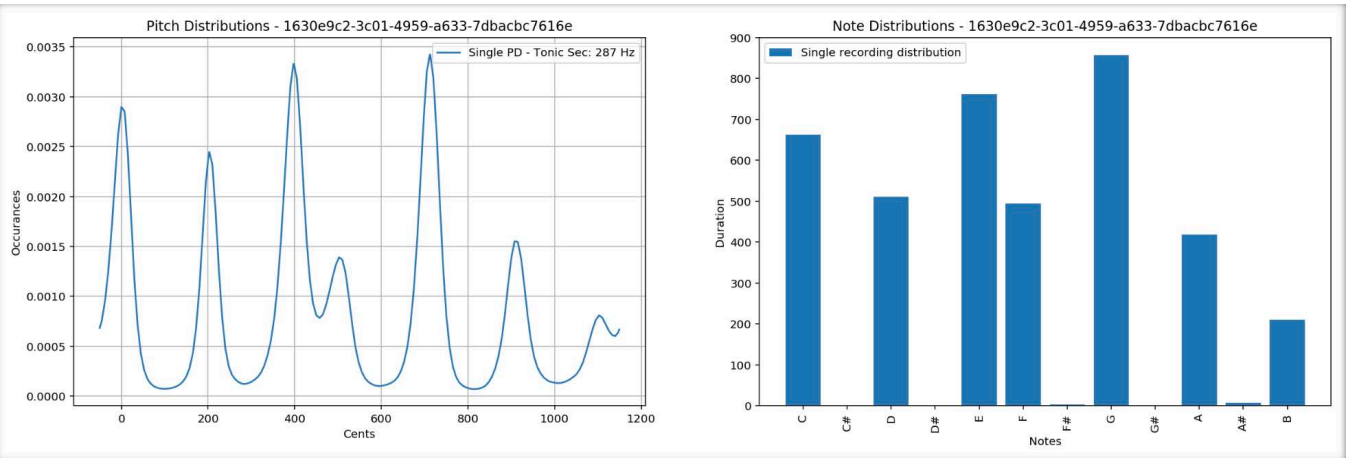

Figure 4: Visualization of the pitch distribution computed from audio and of the pitch class distribution from the scores of a single recording as plotted in the NawbaPitchAnalysis Notebook.

data and metadata as well as performing repetitive tasks in a safe way.

This repository contains four notebooks:

- **Corpus**: to select, download and compute data and metadata of the corpus;
- **Metadata**: to group, visualize and analyze metadata;
- **NawbaPitchAnalysis**: to visualize pitch distribution and note/class distribution of a single recording or of a group of them;
- **NawbaRecognition**: to compute several experiments to evaluate the performance on *nawba* recognition of algorithms based on templates derived from scores.

All the notebooks have a common section to download (if necessary) all the metadata from Dunya and to create a Python object to manage them. Several methods are available in the class that can help the developers that firstly approach this corpus.

The first notebook provides three main functionalities:

(1) a widget to select a list of recordings using filters and checkboxes;

(2) a widget to download mp3, scores in XML and metadata from MusicBrainz (in JSON format);
(3) a widget to compute the pitch-related characteristics, namely fundamental frequency series (unfiltered or filtered with the algorithm proposed in [2]), pitch distribution and several estimations of the tonic frequency.

All the files computed in the latter widget are stored in JSON format, but for the fundamental frequency series it is possible to create a text file with all the values importable in the spectrogram tool of Sonic Visualizer. Furthermore, for each recording the tonic frequency is estimated with three algorithms, two of them using the algorithm proposed in [1] to the filtered and unfiltered version of the fundamental frequency series and the last one applying the algorithm to each section of the recordings and evaluating the most frequent value. As explained in Section 3.1.4, a TXT file with the filtered fundamental frequency series and a JSON file with the pitch distribution are downloadable from Zenodo (see Section 3.1.4), whereas the other derived files are computable only with the notebook.

The second notebook has got three other widgets. The first one allows to group the recordings or their sections by the main musical entities of Arab-Andalusian music (*nawba*, *ṭāb'*, form and *mīzān*) and to show the main statistics concerning the number of sections and recordings and the overall time. This information can be also visualized in bar charts (see Figure 1). The second widget provides a tool to cross information for two different characteristics and to visualize the results in an interactive table (see Figure 3). The last one provides a simple function to visualize the characteristics of the sections in a selected recording.

The third notebook provides a simple tool to visualize both pitch distributions computed from audio recordings and pitch class distributions from the scores, such as in Figure 4. This widget offers the options of (1) comparing distributions of a single recording with the average distribution of a single *nawba*, (2) centering the distribution in three octaves and (3) folding the distributions in a single octave.

With the last notebook, the *nawabāt* recognition task described in [11] can be reproduced.

6 CONCLUSION

Arab-Andalusian music is a unique music tradition, which reflects the rich culture developed in the encounter of Western and Eastern Mediterranean cultures in the Iberian Peninsula. It also reflects the history of that people and their perseverance for preserving such a valuable cultural legacy, often through very adverse conditions. Musicological research on this music started during the colonial occupation of the Maghreb, and even though it is still active in Spanish and Arab publications, is underrepresented in English speaking academia. The Arab-Andalusian music research corpus here presented offers a large collection of openly accessible, well curated and annotated data of different types with the aim of fostering the research of the Moroccan tradition of this repertoire, both from the fields of musicology and MIR.

In order to exploit the potential of the corpus, we plan to work on some of the research tasks it offers (see Section 3.4). We are currently in the process of sectioning the music scores in order to study the characteristics of *ṭāb'* and implement methods for fundamental degree and *ṭāb'* detection using symbolic data. We also intend to study melodic motives from music scores with the aim of characterizing *ṭubū'*. In the future, we will apply existing methods and, if needed, adapt them for score and lyrics to audio alignment. Finally, according to the definition of research corpus described in Section 1, we will continue expanding and improving the corpus, especially with the addition of the lyrics for all the *nawabāt*. Related to this, we argue that one of the important contributions of this corpus is its data management system, which combines the Dunya infrastructure for storing and organizing, and the Jupyter Notebooks for browsing and retrieving data. Consequently, future additions or changes to the corpus will be easily and straightforwardly accessible to the user from the very same interface.

It is also our goal to contribute to the dissemination of this music to new audiences. Therefore, within the framework of a recently started project, Musical Bridges[14], we aim at using these data and the related analysis methods for the development of didactic tools to ease the understanding and appreciation of this repertoire.

ACKNOWLEDGMENTS

The authors would like to thank Alastair Porter, the developer of Dunya, and Andrés Ferraro, in charge of uploading Arab-Andalusian data to Dunya, for their collaboration and technical support. This work was carried out with the support of the Musical Bridges project, funded by RecerCaixa. The Andalusian Corpus was created by the CompMusic project, funded by the European Research Council under the European Union's Seventh Framework Program (ERC grant agreement 267583).

REFERENCES

[1] Hasan Sercan Atlı, Barıs Bozkurt, and Sertan Şentürk. 2015. A method for tonic frequency identification of Turkish makam music recordings. In *5th International Workshop on Folk Music Analysis (FMA 2015)*. Paris, France, 119–122.
[2] Hasan Sercan Atlı, Burak Uyar, Sertan Sentürk, Barıs Bozkurt, and Xavier Serra. 2014. Audio feature extraction for exploring Turkish makam music. In *3rd International Conference on Audio Technologies for Music and Media*. 142–153. http://hdl.handle.net/10230/35018
[3] Rafael Caro Repetto and Xavier Serra. 2014. Creating a Corpus of Jingju (Beijing Opera) Music and Possibilities for Melodic Analysis. In *15th International Society for Music Information Retrieval (ISMIR '14)*. Taipei, Taiwan, 313–318. https://dblp.org/rec/html/conf/ismir/RepettoS14
[4] Amin Chaachoo. 2011. *La música andalusí al-Ála.* Almuzara, Córdoba.
[5] Amin Chaachoo. 2016. *La musique hispano-arabe, al-Ala.* L'Harmattan, Paris.
[6] Mercedes del Amo. 2002. Sistema de transliteración de estudios Árabes contemporáneos. Universidad de Granada. *Miscelánea de Estudios Árabes y Hebraicos. Sección Árabe-Islam* 51 (2002), 355–359.
[7] Georgi Dzhambazov. 2017. *Knowledge-based Probabilistic Modeling for Tracking Lyrics in Music Audio Signals.* Ph.D. Dissertation. Universitat Pompeu Fabra, Barcelona, Spain. https://doi.org/10.5281/zenodo.841980
[8] Mahmoud Guettat. 2000. *La musique arabo-andalouse.* Éditions al-Ouns, Paris.
[9] Christian Poché. 2005. *La música arábigo-andaluza.* Akal, Móstoles.
[10] Alastair Porter, Mohamed Sordo, and Xavier 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, 101–106. http://hdl.handle.net/10230/32251
[11] Niccolò Pretto, Barış Bozkurt, Rafael Caro Repetto, and Xavier Serra. 2018. Nawba Recognition for Arab-Andalusian Music using Templates from Music Scores. In *15th Sound and Music Computing Conference (SMC 2018)*. Limassol, Cyprus, 405–410. https://doi.org/10.5281/zenodo.1257388
[12] Xavier Serra. 2011. A Multicultural Approach in Music Information Research. In *12th International Society for Music Information Retrieval Conference (ISMIR 2011)*. Miami, United States of America, 151–156. http://hdl.handle.net/10230/22723

[14] https://www.upf.edu/web/musicalbridges

[13] Xavier Serra. 2014. Creating Research Corpora for the Computational Study of Music: the case of the CompMusic Project. In *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*. 1–9.

[14] Xavier Serra. 2017. The computational study of a musical culture through its digital traces. *Acta Musicologica* 89, 1 (2017), 24–44. http://hdl.handle.net/10230/32302

[15] Mohamed Sordo, Amin Chaachoo, and Xavier Serra. 2014. Creating corpora for computational research in Arab-Andalusian music. In *1st International Workshop on Digital Libraries for Musicology (DLfM 2014)*. ACM, London, United Kingdom, 1–3. https://doi.org/10.1145/2660168.2660182

[16] Ajay Srinivasamurthy, Gopala Krishna Koduri, Sankalp Gulati, Vignesh Ishwar, and Xavier Serra. 2014. Corpora for music information research in Indian art music. In *International Computer Music Conference/Sound and Music Computing Conference (ICMC/SMC 2014)*. Athens, Greece, 1029–1036.

[17] Burak Uyar, Hasan Sercan Atli, Sertan Şentürk, Barış Bozkurt, and Xavier Serra. 2014. A Corpus for Computational Research of Turkish Makam Music. In *1st International Workshop on Digital Libraries for Musicology (DLfM 2014)*. London, United Kingdom, 1–7. https://doi.org/10.1145/2660168.2660174

[18] Shuo Zhang, Rafael Caro Repetto, and Xavier Serra. 2017. Understanding the expressive functions of jingju metrical patterns through lyrics text mining. In *18th International Society for Music Information Retrieval Conference*. Suzhou, China, 397–403. http://hdl.handle.net/10230/32652

[19] Sertan Şentürk. 2016. *Computational Analysis of Audio Recordings and Music Scores for the Description and Discovery of Ottoman-Turkish Makam Music*. Ph.D. Dissertation. Universitat Pompeu Fabra.