

RELATIONSHIP BETWEEN LYRICS’ SEMANTIC FIELDS AND MELODIC INTERVALS IN ARAB-ANDALUSIAN MUSIC: AN EXPLORATORY ANALYSIS

Gonzalo Nieto Morgan Buisson

Music Technology Group, Pompeu Fabra University

gonzalo.nieto01@estudiant.upf.edu, morgan.buisson01@estudiant.upf.edu

ABSTRACT

Origins of the Arab-Andalusian music tradition can be traced back to Middle-Age, when the Iberian peninsula was being populated by Islamic communities. Although *sana’i* (poems) and vocals play a fundamental role in this music tradition, there is little research on the relationship between lyrics and other elements of this music. In this paper, we investigate the relationship between lyrics’ semantic fields and melodic intervals from an exploratory approach. The dataset used is based on the Arab-Andalusian corpus, taking only poems in the *qasidah* form, resulting in a total of 147 poems. Semantic similarity of poems is calculated using a Bag-Of-Words based approach, taking n-grams into account, while melodic similarity is measured with a basic distance metric on the interval histograms. Results show that there seems to be a link between semantic similarity among poems and their melodic content in Arab-Andalusian music. However, the employed methodology has various limitations that need to be considered.

1. INTRODUCTION

Arab-Andalusian music is usually the term given to the music tradition formed near the 12th Century in the Islamic territories of the medieval Iberian Peninsula, Al-Andalus. It has been preserved to this day as a classical repertoire in several North African countries, acquiring different characters in each of them. This paper centers on the Moroccan tradition, as collected in [8].

Arab-Andalusian music is performed through *nawabat* (plural of *nawba*), musical compositions of instrumental and vocal pieces ordered according to their metrical mode with increasing tempo. Each *nawba* consists originally of pieces composed in the same melodic mode (*tab*) which can be seen as a diatonic scale built upon a fundamental note, as well as a specific collection of characteristic melodic motives. *Nawabat* are structured in five *mawazin* (plural of *mizan*), which can be understood as a sequence of sung poems known as *sana’i* (plural of *san’a*) with some instrumental interludes, performed in heterophony. Originally, a *nawba* per part of the day was performed, telling different stories throughout the daytime [3].

Given the central role of poems and vocals in Arab-Andalusian music, it seems like a question of musicological interest to know more about the relationship between lyrics and other elements of the music. It is known that

tubu’ are associated with certain emotional and spiritual content [4]. However, to our knowledge, there is no research on lyrics’ semantic content related to melodic variations.

Therefore, the aim of the present paper is to gain insights on the mutual influence between lyrics’ semantics and melody variations in Arab-Andalusian music. For this purpose, the data available needs to be pre-processed and analyzed both semantically and melodically. The strategy is to find groups of semantically-similar poems and compare them melodic-wise, and repeat the process with not semantically-similar groups. At the end, the comparison could reveal if there is a link between them or not.

All code used for the paper is available at [7].

2. DATA PROCESSING AND EXPLORATION

The data used for the present research is available in the Github repository by Alia Morsi and Miguel García Casado [6]. It is a dataset created from a subset of the Arab Andalusian corpus of the CompMusic project [1], combining scores and textual information from the music [5] and lyrics [9] corpora. In total, there is information of 114 recordings. More specifically, it contains the following relevant folders for this research:

- Lyrics: available in both their original Arabic script (folder ‘original’) and a transliterated version (folder ‘transliterated’, using the American Library of Congress – ALA-LC standard), in TSV and JSON formats. Each file is identified by its MusicBrainz recording ID (MBID).
- Scores-musicxml: directory with the score transcriptions of the corpus recordings, each score named with the MBID of its corresponding recording.
- Metadata-all-nawbas: file containing information –MBID, title, transliterated title, internet archive URL and miscellaneous– about each of the recordings of the corpus.

In order to be able to retrieve the information about each poem from each recording score, the data needs a thorough pre-processing step.

2.1 Pre-Processing

First of all, from all the recordings available, there are some that do not have their lyrics available or that need

completion. An initial filtering of this recordings is performed, resulting in a total of 95 recordings with their corresponding lyrics. It must be noted that there is no alignment of the lyrics to the scores. Scores contain some text annotations indicating when an instrumental section starts (like *Mshalia* or *Tawshiyya*), or when a poem starts by adding the first words of the first line of each poem. Even though all the lyrics are transliterated, these transliterations are automatic, and the annotations in the scores were done manually by Amin Chaachoo, consequently, the two systems do not exactly match. As a workaround, we used a filtering algorithm based on a linear combination of both the edit distance and the common longest sub-string length between poems and their score annotation candidates provided by Chaachoo. This way, we were able to spot the closest string in the lyrics JSON file to a given annotation from the score of the same recording. We also set a threshold, on the longest common sub-string specifically, to discard annotations that were not similar to any poem start – possibly tempo notations, instrumental names or errors. This threshold has been set empirically: when the longest common sub-string was smaller than 3 characters, the annotation and the automatic transliteration were considered completely distant.

Finally, it has to be taken into account that the lyrics of this corpus can be in classical Arabic –not exactly the same to Modern Standard Arabic–, in the Andalusian dialect of Arabic or in the Romanic language spoken in Iberian Al-Andalus at that time. For this reason, this research focuses on the *qasidah* form, which uses classical Arabic.

The result of the pre-processing step is a folder with 95 text files –identified by their MBID– containing a list for each poem with (i) manual annotation, (ii) transliterated poem identifier, (iii) poem number within the recording, (iv) starting measure number, (v) original lyrics, (vi) transliterated lyrics and (vii) poem form. Annotations that are not detected as lyrics are also included with the label "Not Retrieved/Instrumental".

2.2 Statistical Descriptive Analysis

A preliminary descriptive analysis is performed in order to have a general idea of the data. Figure 1 shows the distribution of poetic forms in the whole lyrics corpus of Arab-Andalusian music. The great majority of the poems are in the *zajal* form. On the other hand, there is a minority of *birūlatin* forms, and the total number of poems with the *qasidah* form is 147. Other side outcomes of this analysis are attached in the Appendix section A.1.

3. LYRICS ANALYSIS

Extracting the general subject or "semantic field" of a text is a problem that has thoroughly been studied in the context of English Natural Language Processing (NLP). As described above, we have at our disposal two different versions of each poem: the original transcription, which most of the time happens to be in classical Arabic; and a transliterated version with a harmonized syntax. In this section, we describe some of the experiments that have been carried out in order to build a measure of semantic similarity

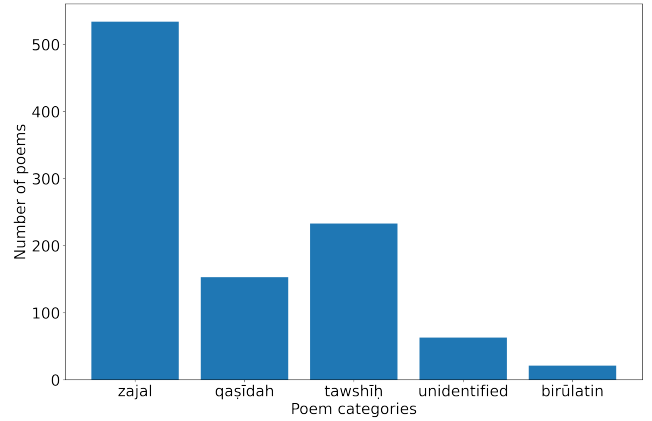


Figure 1. Poetic Forms Distribution

across poems. We also take a deeper look at the method we have chosen to use for the rest of our work, our motivations, and its possible limitations.

From the first attempt described in the Appendix section A.2, we retained three important factors that we assumed had greatly impacted the overall quality of our results:

- The parsing method: Although being easily implementable, dividing each poem word by word seemed to considerably discard important information. In Arabic especially, the context plays a crucial role. Therefore, a word can have a very different meaning from one sentence to another based on its position in the sentence and its surroundings.
- The poems grouping strategy: Using a clustering algorithm, in this context, did not yield any relevant results as there was a high semantic variability within each cluster.
- The translation step was also very likely to induce some errors in our keywords selection process, as poems from our dataset were not always word for word translatable.

Therefore, we decided to operate with an analogous yet different approach. First, as the keywords selection method, we decided to employ a Bag-of-Words based [2] algorithm which not only spots the most relevant words in a corpus, but also looks for n-grams with n ranging from 1 to 6 in order to make sure to not discard any relevant contextual information. This method resulted in a total of 103 and 192 detected keywords for the original and transliterated lyric versions respectively. In order to find a trade-off between a sparse and efficient representation of the poems, a minimum threshold of appearance equal to 4% and a maximum one equal to 95% have been set. Regarding the poems grouping strategy, we opted for a simplified approach that would consist in a query-based algorithm. For each poem, the method returned the 5 most similar ones based on their vector representation obtained from the previous step described above. One important assumption that

we made here is that similarity across poems only depends on the semantic content they share. In other words, the more words or n-grams two poems share, the more semantically similar they will be considered. Our final distance measure results in counting the number of non-zero common components of two poems' vector representation. From a general perspective, to preserve the integrity of the text and keep any pertinent semantic information, we did not use any translation tool throughout the process. We finally repeated a similar pipeline to group non-semantically close poems together, by selecting for each poem, a random sample of 5 of its least similar ones.

4. MELODY ANALYSIS

Regarding the melodic analysis, once the starting measure of each poem is known, the notes corresponding to each poem can be retrieved from the scores. Afterwards, a simple melodic analysis is applied: counting the number of intervals (m2, M2, m3, etc.) between consecutive notes for each poem with music21 (see code details in the available repository [7]).

To be able to compare the melody of different poems, the number of intervals per interval type and per poem is normalized. This way, all interval types are contained within the range 0-1.

Lastly, to compute an objective difference between two melodies, we use an L1-based distance measure. We take the absolute difference between each interval type (for instance, m2: $0.5 - 0.4 = 0.1$) and then sum all differences up, to have a single number that represents how similar they are. Once the semantic analysis has produced groups of similar and non-similar poems, we compute an average inner-group distance – resulting in a single value per group.

5. RESULTS AND DISCUSSION

As a final step, the goal is to compare any potential influence between the lyrics content and melodic patterns. By combining the grouping strategy between similar and dissimilar poems, with the melodic distance calculation method, our final results will try to measure whether the grouping strategy has an impact on the melodic cohesiveness of grouped poems. As we applied our experiment for both versions of the poems lyrics, we now end up with 4 distinct groups: one group per lyric version, and one group for each similarity/dissimilarity category. Figure 2 and Figure 3 show the distribution of the average melodic distance within each group, for semantically similar and dissimilar poems in their original and transliterated version respectively.

One first noticeable element is how much the lyrics version influences the final results of the melodic similarity analysis. When using transliterated lyrics, our definitions of semantic and melodic similarity seem to align, as the semantically similar category seems to have a lower average inter-group melodic distance than the other category. However, when using lyrics in their original version, this trend is inverted and both of these definitions contradict each

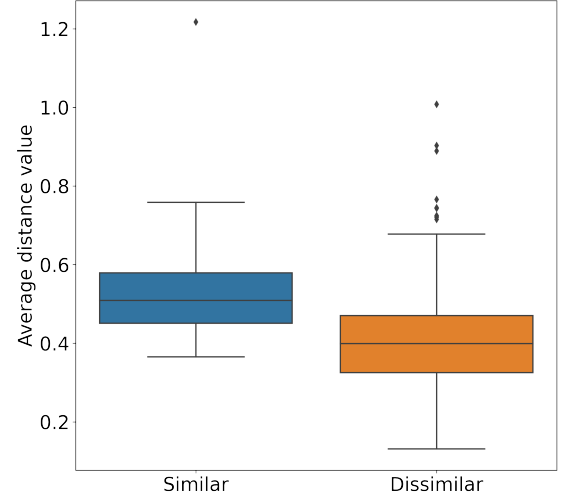


Figure 2. Melodic Distance Boxplot (Original Lyrics)

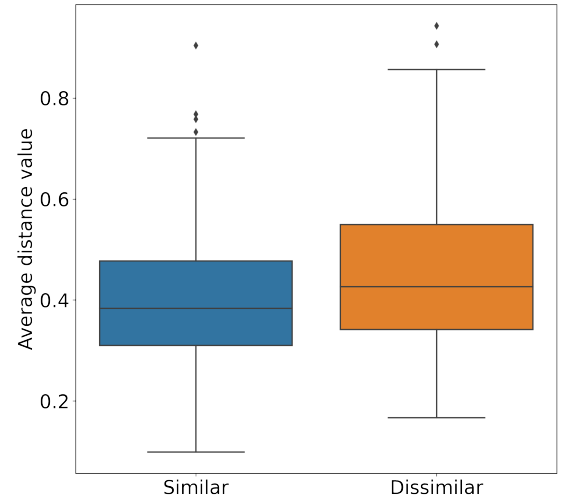


Figure 3. Melodic Distance Boxplot (Transliterated Lyrics)

other. In order to measure whether these results are significant or not, and infer about any actual difference between Similar and Dissimilar categories, we choose to carry out two T-Tests. We specifically choose to perform independent T-tests as our observations (i.e our groupings of poems) never happen to be in both categories, even though they are formed using the same "population" of poems. Before doing so, we make sure that our data comply with the normality assumption required by T-Tests. As an example, Figure 6 in Appendix section A.3 shows that the melodic distance distribution for semantically similar group using transliterated lyrics seems to follow a Gaussian-like distribution. We finally use an equal variance type of test as

we have the same number of groups in each category (one group per poem), which represents a total of 147 groups per category in total. We formulate our hypotheses as follows:

- The null hypothesis (H0): Asserts that there is no significant difference between the means of the two categories. This case would imply that no significant mutual influence might exist between the lyrical content and the melodic patterns.
- The alternate hypothesis (H1): Asserts that there is a significant difference between the two means, and therefore, a possible relationship between the semantic information contained in the lyrics and the melodic content.

	Original	Transliterated
t-statistic	6.46	-2.89
p-value	$4.47e - 10 < .05$	$0.004 < .05$

Table 1. T-Tests Results

From our results reported in Table 1 using a 95% confidence interval, there seems to be, regardless of the lyric version used, a significant difference in the average melodic distance within groups across categories. Therefore, we can assume that the grouping strategy has an impact on the melodic cohesiveness of the groups. However, this same grouping strategy is directly linked to our assumption about semantic similarity. This leads us to conclude that semantic similarity among poems and their melodic content, in Arab-Andalusian music, might have a mutual influence. Although the performed tests highlight a statistical significance underlying the data, there are a few things that have to be taken into account while considering the aforementioned conclusion. First, the process has been applied only to the *qasīdah* form, which as mentioned before, does not represent the majority of the lyrical content available in the dataset. Second, the numerous constraints related to the Arabic language computational processing and the lack of efficient tools to address it led us to choose a quite simplistic definition of semantic similarity. Therefore, it is very probable that some semantically relevant poems have been discarded during the grouping process. Moreover, we also lack of a formal and reliable evaluation process, as our only way to qualitatively assess the grouping process was to use inaccurate translation tools. We have also highlighted in Figure 4 and Figure 5 that the grouping step is subject to variations based on the lyrics version we use, as it is a crucial part of our overall methodology, this can have direct consequences on the final results. Third, the melodic pattern analysis has been restrained to a fairly basic approach, while not taking other more complex indicators into account. However, this work marks a first attempt towards drawing a link between the lyrical and melodic content in Arab-Andalusian music.

6. FUTURE RESEARCH DIRECTIONS

In this section, we propose a few improvements that could possibly be brought on top of the existing work. Since our analysis is essentially exploratory, there are several limitations that could be addressed in the future:

- As researchers who do not know Arabic, it is difficult to evaluate the semantic analysis correctly. We have used tools such as the Google Translator API, but this is clearly far from ideal. Further analysis coming from native speakers could directly improve the reliability of the results.
- A pre-trained NLP model in Arabic could be used to better analyze the semantics of the poems, building for example a semantic embedding to group similar words or poems together.
- A harmonization of the lyrics could be performed – adapt all dialects to the same variety of Arabic (classical Arabic or Modern Standard Arabic)–, so all poems can then be automatically analyzed.
- If there is an actual, mutual influence between lyrics and melodic variations, it could be studied more in detail: whether there are certain motives related to specific topics, whether the lyrics condition the melody or the other way around, etc.
- The melodic analysis performed in this paper is fairly simple, more complex and richer analysis could be carried out in the future.
- The normalization step of the number of occurrences may be misleading. We may be comparing melodies that are not that comparable due to the number of notes. This can be further researched.
- We have found sporadic errors along the lyrics corpus such as empty lines coded as empty sections (see file `25857838-1a8a-4102-a818-88bca8be5cf6.tsv`) or poem forms included as lyrics (`04b42450-8838-4a20-9e79-d42f03f8cc51.tsv`). Moreover, the edit distance method used to find poems start measures also produces some errors. This could be manually revised or a more reliable method could be created in the future.

7. CONCLUSION

We have proposed an exploratory approach to address the relationship of lyrics’ semantic fields and melodic intervals in Arab-Andalusian Music. Results seem to show that semantic similarity among poems and their melodic content might have a mutual influence. However, the employed methodology has various limitations that need to be considered: (i) the process has been applied only to the *qasīdah* form, which does not represent the majority of the lyrical content available in the dataset, (ii) we lack of a formal and reliable evaluation process for the semantic analysis and (iii) the melodic pattern analysis carried out is

fairly simple. Nonetheless, this work has unveiled new research questions that could potentially deepen our general understanding of Arab-Andalusian music.

8. REFERENCES

- [1] Compmusic: Research corpora. <https://compmusic.upf.edu/corpora>. Accessed: 13-03-2021.
- [2] Count vectorizer: Scikit-learn reference. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html. Accessed: 16-03-2021.
- [3] Jewish music research centre: Andalusian nuba. <https://jewish-music.huji.ac.il/content/andalusian-nuba>. Accessed: 16-03-2021.
- [4] Amin Chaachoo. *La musica andalusi al-Ala*. Al-muzara, Cordoba., 2011.
- [5] Amin Chaachoo, Mohamed Sordo, Niccolo Pretto, Rafael Caro Repetto, Baris Bozkurt, and Xavier Serra. Arab-andalusian music corpus, June 2018. This work also received the support of the Musical Bridges project, funded by RecerCaixa.
- [6] Alia Morsi and Miguel García Casado. Github repository: Arab-andalusian music. <https://github.com/MTG/arab-andalusian-music>, 2021. Accessed: 14-03-2021.
- [7] Gonzalo Nieto and Morgan Buisson. Github repository: Relationship between lyrics’ semantic fields and melodic intervals in arab-andalusian music. <https://github.com/gonznm/aam-lyrics-melody>, 2021. Accessed: 18-03-2021.
- [8] Niccolò Pretto, Bariş Bozkurt, Rafael Caro Repetto, and Xavier Serra. Nawba recognition for Arab-Andalusian music using templates from music scores. *Proceedings of the 15th Sound and Music Computing Conference: Sonic Crossings, SMC 2018*, pages 394–399, 2018.
- [9] Mohamed Sordo, Mehdi Chaachoo, and Xavier Serra. Arab-andalusian music lyrics dataset, June 2018. New version prepared by Alia Morsi.

A. APPENDIX

A.1 Further Statistical Descriptive Analysis

Figure 4 presents a correlation matrix between the different forms of all poems in the corpus. There seems to be no correlation between them, as the maximum presented is 0.3. This indicates that the distribution of each poetic form within each poem does not seem to follow any kind of trend or pattern.

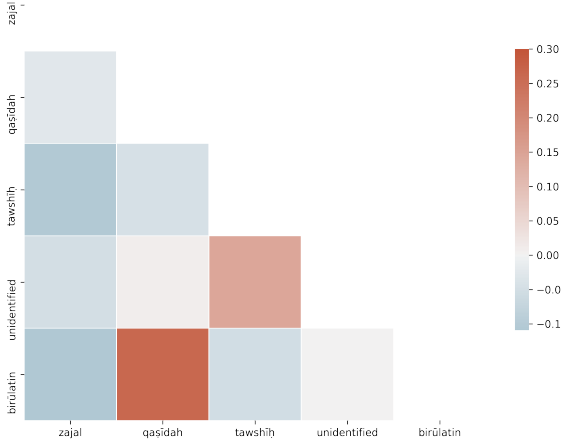


Figure 4. Poetic Forms Correlation

As a side outcome of the research, a list of possible instrumental names has been gathered. These are annotations from the all scores that are not detected as lyrics using the aforementioned edit distance method. Figure 5 shows the number of occurrences of each annotation for the whole corpus of 95 recordings. One can see that tempo notations such as "accel.", "Larghetto" or "rit.", as well as empty strings, are very frequent. However, taking out these annotations, the method seems to reveal names of instrumentals that were not previously known by the authors, as "TAQSIM" or "MAWWAL". There also seems to be errors with specific poems possibly not being interpreted as such ("Ana elmusi 'u linafsi" or "Arsala Laho ilayna").

A.2 Semantic Experiments

The problem of comparing poems semantically can be approached in various ways. However, our main assumption in this work has been the following: If two poems share some semantic information, then they are more likely to have words in common. From this hypothesis, our first objective has been to build a list of keywords from our poems corpus. To this end, the most simple method we started with has been to list all the words present in our

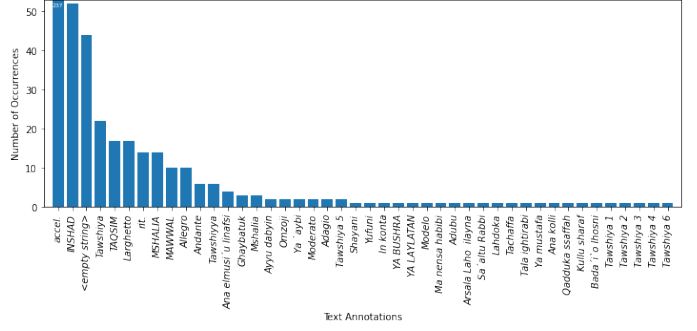


Figure 5. Annotations from the scores that are not automatically detected as lyrics

database, and sorting them based on their number of appearances. Then, in order to select only relevant and meaningful words, we chose to discard the ones that would appear too often, and the ones that would not appear in enough poems. We also used the Google Translate API in order to select nouns and verbs that carried the most information (based on their English translation). Given this final list of words, each poem was then converted to a vector representation, of which each component i would be either 0 or 1 whether it contained the corresponding i th keyword from the list. Using their respective mathematical representation, it was then possible to calculate a distance value between two poems. We carried out this calculation for each possible pair of poems from our corpus and obtained a distance matrix where each component (i, j) represented the distance between poems i and j . Finally, using this distance matrix, we applied a hierarchical clustering algorithm that iteratively merged together closest elements from our corpus to eventually obtain n cluster of poems. With a silhouette score analysis and a qualitative evaluation using the Google Translate API, it turned out that the resulting clusters for both original and transliterated lyric versions did not seem semantically relevant.

A.3 Other Figures

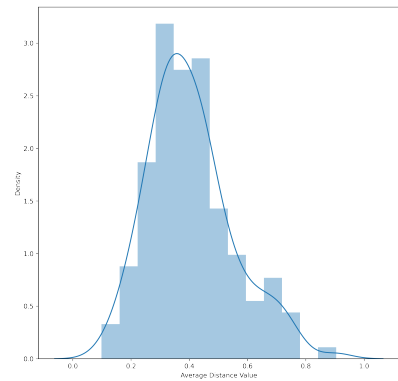


Figure 6. Melodic Distance Distribution (Original Lyrics, Similar Group)