Université de Neuchâtel
Faculté des sciences économiques

Computational exploration and analysis of meters and patterns in Arabic metrical poetry:

Detection, visualization, and clustering of structures and compositions

---

PhD thesis submitted to the Faculty of Economics and Business

Information Management Institute

University of Neuchâtel, Switzerland

For the PhD degree in Computer Science

by

Abdelmalek BERKANI

Approved by the dissertation committee:

Prof. Adrian HOLZER, University of Neuchâtel, thesis director

Dr. MER. Paul Cotofrei, University of Neuchâtel

Prof. Mohamed Khaldi, Université Abdelmalek Essaadi, Morocco

Prof. Bruno Paoli, Université Lumière Lyon 2, France

Defended on December 12th, 2024

Université de Neuchâtel
Faculté des sciences économiques

IMPRIMATUR POUR LA THÈSE

Computational exploration and analysis of meters and patterns in Arabic metrical poetry: Detection, visualization, and clustering of structures and compositions.

Abdelmalek BERKANI

UNIVERSITÉ DE NEUCHÂTEL

FACULTÉ DES SCIENCES ÉCONOMIQUES

La Faculté des sciences économiques,

sur le rapport des membres du jury

Prof. Adrian HOLZER, Université de Neuchâtel, directeur de thèse

Dr. MER Paul COTOFREI, Université de Neuchâtel

Prof. Bruno PAOLI, Université Lumière Lyon 2, France

Prof. Mohamed KHALDI, Ecole Normale Supérieure, Université Abdelmalek Essaadi, Maroc

autorise l'impression de la présente thèse.

Neuchâtel, le 11 février 2025

Le doyen
Peter Fiechter

Faculté des sciences économiques
Avenue du 1er-Mars 26
CH-2000 Neuchâtel
Email: secretariat.seco@unine.ch
Tel: +41 32 718 15 00
www.unine.ch/seco

Abstract

This thesis addresses the issue of insufficiency of data and natural language processing (NLP) tools related to the metrical system of Arabic poetry, which makes the exploration of poetic compositions challenging. Although the metrical system has been extensively studied from a linguistic perspective, it has received far less attention in computational linguistics, where structured data and appropriate processing techniques are essential for effective analysis. To help address this problem, the thesis provides solutions across three key areas: Arabic metrical system rules, NLP techniques, and the analysis and visualization of poetic structures. We introduce *Arabic Meters Identification System (AMIS)*, a structured framework organized into specialized modules. At its core, *AMIS* integrates the data and rules of the Arabic metrical system, which characterizes all Arabic meters. Additional modules focus on the automatic detection of meter in poems or individual verses, achieving a reliable accuracy of 99.97%. Other modules are dedicated to the structural analysis and visualization of poetic compositions, including comparative assessments. Furthermore, this thesis identifies the key features of metered Arabic poetic structures through a clustering approach based on five metrics: *Evenness*, *Variability*, *Repetition*, *Pattern Potential*, and *Pattern Usage*. Our findings show that *Repetition*, *Variability*, and *Pattern Usage* are the most significant metrics, with just two attributes being sufficient to characterize a poem's structure. Additionally, we explore the potential of visual representation of Arabic poetry, making its structure visible and accessible to both Arabic and non-Arabic speakers. These visualization techniques uncovered new metrics that improve clustering outcomes, providing deeper insights into the structure of Arabic poetry.

Keywords: Arabic poetry, Meters and patterns, Clustering, Visualization.

Résumé

Cette thèse aborde le problème de l'insuffisance de données et d'outils de traitement automatique du langage naturel liés au système métrique de la poésie arabe, ce qui rend l'exploration des compositions poétiques difficile. Bien que le système métrique ait été largement étudié d'un point de vue linguistique, il a reçu beaucoup moins d'attention en linguistique computationnelle, où des données structurées et des techniques de traitement appropriées sont essentielles pour une analyse efficace. Pour aider à résoudre ce problème, la thèse propose des solutions dans trois domaines clés : les règles du système métrique arabe, les techniques de traitement automatique du langage naturel (NLP) et l'analyse et la visualisation des structures poétiques. Nous présentons *Arabic Meters Identification System (AMIS)*, un framework structuré organisé en modules spécialisés. Au cœur de *AMIS*, les données et les règles du système métrique arabe sont intégrées, caractérisant ainsi tous les mètres arabes. Des modules supplémentaires se concentrent sur la détection automatique des mètres dans les poèmes ou les vers individuels, avec une précision fiable de 99,97%. D'autres modules sont dédiés à l'analyse structurelle et à la visualisation des compositions poétiques, y compris des évaluations comparatives. En outre, cette thèse identifie les principales caractéristiques des structures poétiques arabes métrées grâce à une approche de Clustering basée sur cinq métriques : *Evenness*, *Variability*, *Repetition*, *Pattern Potential* et *Pattern Usage*. Nos résultats montrent que *Repetition*, *Variability* et *Pattern Usage* sont les métriques les plus significatives, avec seulement deux attributs suffisants pour caractériser la structure d'un poème. De plus, nous explorons le potentiel de la représentation visuelle de la structure de la poésie arabe, la rendant accessible aux locuteurs arabophones et non arabophones. Ces techniques de visualisation ont permis de découvrir de nouvelles métriques qui améliorent les résultats du Clustering, offrant des perspectives plus approfondies sur la structure de la poésie arabe.

Mots-clés: poésie arabe, mètres et patterns, Clustering, visualisation.

Acknowledgements

I would like to express my sincere gratitude to my thesis supervisor, Professor Adrian Holzer, for his invaluable guidance throughout this journey and for his dedicated effort in embracing a subject deeply rooted in the complexities of the Arabic language. I am deeply grateful for his availability and insightful advice on all aspects of the research process.

I also extend my heartfelt thanks to Professor Paul Cotofrei for his guidance on the mathematical and technical aspects of this work, as well as for his time and meticulous review in evaluating the thesis.

Furthermore, I am profoundly grateful to Professors Mohamed Khaldi and Bruno Paoli for their expertise and generous time in reading and evaluating this research.

This thesis, completed entirely during family time in the evenings, weekends, and holidays, would not have been possible without the unwavering support of my wife, Rquia, and my sons, Ilyass and Ziad. Their patience, constant encouragement, and sacrifices in enduring my absence, as well as the indirect pressures of my research, were instrumental in bringing this project to completion. I owe them a debt of gratitude that words cannot fully express.

A special thanks goes to my longtime friend, Mohamed Nekcha, with whom I share a passion for Arabic literature, poetry, philosophy, and music. Our many conversations over countless pizzas have been a continuous source of inspiration and encouragement throughout this journey.

My heartfelt thanks to my childhood friend, Mr. Abdelkader Nejjari, and his wife, Amel, for their unwavering encouragement.

Finally, I am deeply grateful to all members of my family in Morocco, whose belief in me has been a constant source of strength. Their encouragement has been invaluable, and I deeply cherish their presence in my life.

Contents

Abstract **v**

Résumé **vii**

Acknowledgements **ix**

Contents **xi**

List of Figures **xvii**

List of Tables **xix**

1 Introduction **1**

1.1 Arabic language 1

1.2 Arabic poetry 3

1.3 Arabic metrical system 4

1.4 Problem statement 5

1.4.1 Metrical system data 6

1.4.2 Meter detection 6

1.4.3 Structure analysis 7

1.4.4 Poetry visualization 8

1.4.5 Poetry corpus 9

1.5 Contributions 10

1.5.1 Metrical system data 11

1.5.2 Meter detection 13

1.5.3 Structure analysis 14

1.5.4 Poetry corpus 15

1.6 Form and structure 15

References 17

I Arabic metrical system and meter detection 21

2 Rule based meter detection 23

2.1 Introduction 23

2.2 Background 25

2.2.1 Syllables 25

2.2.2 Patterns 26

2.2.3 Meters 27

Original meters 28

Meter variants 29

2.3 Related work 31

2.4 Arabic Meters Identification System 33

2.4.1 Data and rules collection 33

Pattern combinations collection 35

Single poetry verse 35

Verse parts concordance 37

Meter ambiguities 38

2.4.2 Meter detection processing 40

Text preparation 41

Syllabic segmentation 43

Exact pattern matching 44

Similarity matching 45

2.5 Poetry corpus collection 46

2.5.1 Challenges in poetry collection 46

2.5.2 Used corpus 48

2.6 Results 48

2.6.1 Meters 49

2.6.2 Evaluation 54

2.7 Discussion 55

2.8 Conclusion 58

References 59

II Structure analysis **63**

3 Clustering of metered Arabic poetry **65**

3.1 Introduction 65

3.2 Related work 67

3.2.1 Classification of non Arabic poetry 67

3.2.2 Classification of Arabic poetry 69

3.3 Poem structure metrics 70

3.4 Data selection 71

3.4.1 Corpus 71

3.4.2 Poems data 72

3.4.3 Evenness 73

3.4.4 Variability 74

3.4.5 Repetition 74

3.4.6 Pattern Potential 74

3.4.7 Pattern Usage 75

3.5 Clustering analysis 76

3.5.1 Clustering approach 78

3.5.2 Clustering tendency 79

3.5.3 Clustering options 80
3.6 Results 82
3.7 Discussion 86
3.8 Conclusion 87
References 88

III Visualization 93

4 Visualization of poetic compositions 95
4.1 Introduction 95
4.2 Related work 97
4.3 Exploratory of poetry visualization 99
4.3.1 Verse regularity 99
4.3.2 Pattern regularity 100
4.3.3 Visualizing a poem summary 101
4.3.4 Visualizing a poem details 103
4.3.5 Case study 103
4.4 Materials and methods 109
4.5 Visualization 111
4.5.1 Patterns with distinct colors 112
4.5.2 Patterns with shared colors 112
4.5.3 Dissimilarities 113
Histogram 113
Structural Similarity Index Measure (SSIM) 113
Highlighting differences 114
4.6 Experiment 118
4.6.1 Distribution across columns 119
4.6.2 Distribution across rows 119

4.6.3 Distribution across cells 120
4.6.4 Positions across columns 120
4.6.5 Positions across rows 120
4.6.6 Clustering 121
4.6.7 Evaluation 124
4.7 Discussion 125
4.8 Conclusion 126
References 127

5 Conclusion **131**
5.1 Research assessment 131
5.2 Future research 134
References 136

IV Appendices **137**

A Meter ambiguities: pruning code **139**

B Metrical system data: view by meter **143**

C Metrical system data: view by pattern **153**

D Visualization samples **163**

E Visualization experiments **201**

List of Figures

1.1 *AMIS* Framework 11

2.1 *AMIS* data and rules collection 34

2.2 *AMIS* meter detection processing 41

2.3 Example of meter detection process 45

3.1 Attributes correlation matrix 72

3.2 Scaled data summary 76

3.3 Clustering option 1 82

3.4 Clustering option 2 83

3.5 Clustering options 4 and 5 84

3.6 Clustering options 3 and 6 85

4.1 Example of Arabic poetry visualization 95

4.2 Summary visualization of poetry composition 102

4.3 Shapes of syllable visualization 103

4.4 Summary visualization of the poems: *al-Mu'allaqāt* 105

4.15 Research approach overview 109

4.16 Distinct colors option 112

4.17 Shared colors option 112

4.18 Dissimilarities: different meters, patterns with distinct colors 115

4.19 Dissimilarities: different meters, patterns sharing colors 116

4.20 Dissimilarities: same meters, patterns with distinct colors 117

4.21 Dissimilarities: same meters, patterns sharing colors 118
4.22 Clustering by positions across columns: option k=3 122
4.23 Clustering by positions across columns: option k=4 123

List of Tables

2.1 Syllables notation . . . 26

2.2 Original poetry patterns . . . 27

2.3 Original poetry meters with 8 positions . . . 28

2.4 Original poetry meters with 6 positions . . . 28

2.5 Original poetry meters with 4 positions . . . 29

2.6 Structure of Arabic metrical system dataset . . . 35

2.7 Example of combinations for single verse detection . . . 36

2.8 Concordance of verse hemstitchs: view by syllables . . . 37

2.9 Concordance of verse hemistichs: view by names . . . 38

2.10 Remaining meter ambiguities . . . 40

2.11 Poetry corpus . . . 48

2.12 Arabic metrical system: data summary . . . 49

2.13 Arabic metrical system: combinations . . . 50

2.14 Arabic metrical system: available patterns . . . 51

2.15 Arabic metrical system: syllables counts ranging from 7 to 30 . . . 51

2.16 Arabic metrical system: patterns frequency . . . 52

2.17 Distribution across positions of 12 patterns representing 85.35% of the total frequency . . . 53

2.18 Distribution across positions of 12 patterns representing 10.35% of the total frequency . . . 53

2.19 Distribution across positions of 19 patterns representing 4.30% of the total frequency . . . 54

2.20 Meter detection precision 55

3.1 Retained poem metrics 73

3.2 Data summary before standardizing 75

3.3 Attributes combinations and experiments 79

3.4 Clustering tendency assessment using Hopkins Statistic *(h)* . . . 80

3.5 Clustering baseline where Silhouette >= 0.50 81

3.6 Clustering options 81

4.1 Example of verse regularity calculation 100

4.2 Example of pattern regularity calculation 101

4.3 Example of data used for summary visualization 101

4.4 Case study using a corpus subset: *al-Mu'allaqāt* 104

4.5 Example of a poem's metrical structure represented as patterns and their syllabic quantities 111

4.6 Example of a normalized poem's metrical structure with 16 dummy syllables (N) 111

4.7 Cluster membership and average syllabic quantity (Q) for k=3 and k=4 options 123

4.8 Comparison of clustering scores using "Positions across columns" metric with past values 124

Chapter 1

Introduction

1.1 Arabic language

Arabic, a Semitic language closely related to Hebrew, Aramaic, and Amharic, was initially the language of Bedouins in the deserts of Arabia. However, within two centuries, it emerged as one of the major languages in the history of human culture [1]. With a rich literary tradition dating back to the pre-Islamic era, Arabic has played a central role in shaping linguistic and cultural heritage. The Arabic language originates from the *koinè* used by poets of Central and Eastern Arabia [2]. During the expansion of the Islamic caliphate from the 7th to 12th centuries AD, Arabic evolved from a regional language into the official administrative language of the empire. It also became a primary language of scholarly and scientific discourse, bridging intellectual traditions across regions and continents [3]. Today, Arabic is the native language of over 470 million people [4] and serves as the liturgical language for over a billion Muslims worldwide. Its profound cultural and historical significance continues to resonate in modern academia.

The standardization of literary Arabic began in the 8th century AD with the establishment of grammatical norms. This standardized form, referred to as *Standard Arabic*, includes both *Classical Arabic* and *Modern Standard Arabic*

(MSA), representing its medieval and modern variations, respectively. Despite differences in the lexicon and stylistic features of *MSA*, its morphology and syntax have remained relatively stable [5]. Unlike many other diglossic languages, standard Arabic is not acquired as a mother tongue in the Arab world but is learned formally through education. It is primarily used in formal written and spoken contexts, whereas vernacular dialects dominate everyday communication [6]. Over time, the gap between Standard Arabic and regional or social colloquial varieties has widened significantly, particularly in terms of phonology, morphology, syntax, and lexicon. The various dialects of Arabic differ significantly from one another, to the extent that mutual intelligibility is challenging between the most distinct forms. However, only one of these dialects, Maltese, has developed into a separate written language [7].

Arabic is written from right to left and its alphabet consists of 28 consonants and 3 vowels, which can be either short or long. The short vowels, represented by diacritics above or below the consonants, are known as *Fatha*, *Damma*, and *Kasra*, while the absence of a vowel is marked by the diacritic *Sukun*. Each short vowel has a corresponding long vowel: *Alif* for *Fatha*, *Waw* for *Damma*, and *Yae* for *Kasra* [8].

However, Arabic text is generally not vowelized in everyday writing. Instead, readers must infer the correct vowels based on context, a characteristic feature of the *abjad* writing system, where each letter typically represents a consonant, leaving the vowels to be supplied by the reader [9].

Arabic's derivational morphology is based on a root and pattern system. Roots, typically composed of three consonants, form the semantic core, while patterns comprising vowel sequences and affixes modify the root to create specific words. For instance, the root *(ktb)* conveys the concept of writing and, depending on the pattern, can form words like *kataba* (he wrote), *maktab* (office), and *maktaba* (library). Further complexity arises from Arabic's inflectional morphology, where verbs and nouns are modified to express tense, number, gender, and case (nominative, accusative, or genitive). The lack of written vowels creates significant ambiguity, as many words with the same consonantal root can differ only in their vowels. For example, the root *(ktb)* can be interpreted as *kataba* (he wrote), *kutiba* (it was written), or *kutub* (books). This absence of vowel markers makes context essential for determining meaning, presenting substantial challenges for natural language processing (NLP) tasks, where accurate vowel restoration and word disambiguation are critical for text analysis and machine translation [10].

1.2 Arabic poetry

The Classical Arabic era, typically traced back to the 6th century AD, witnessed a flourishing of literary expression, particularly in the form of poetry. This period saw the rise of public recitation and oral composition, culminating in a highly refined and formalized poetic tradition. The Arabic ode, or *qasīda*, evolved during this era into its most sophisticated and structured form, becoming the benchmark of literary eloquence and artistic expression [11]. As such, Arabic poetry is widely regarded as the earliest form of Arabic literature, playing a foundational role in the literary tradition [12].

Arabic poetry is primarily divided into three main types. *Metered poetry* is the classical form that adheres to strict meter and rhyme schemes based on the Arabic metrical system where verses follow consistent rhythmic patterns and rhyme throughout the poem. Each verse is composed by two hemistichs: the first part *al-Sadr*, and the second part *al-'Ajoz*. In rare exceptions, a verse may not have a second part. *Free verse* emerged in the mid-20th century, allowing poets to break from traditional meter while maintaining rhythm and thematic coherence, providing greater flexibility in poetic expression. *Prose poetry*, a modernist form, abandons meter and rhyme entirely, blending elements of prose and poetry. It focuses on imagery, emotion, and musicality, challenging classical conventions even more than free verse [13].

1.3 Arabic metrical system

In metered Arabic poetry, the field of study related to poetry meters is called *al-'Arud*. It focuses on the theoretical rules governing composition in classical Arabic poetry, with the aim of capturing and describing its metrical structure [14]. *al-'Arud* is the prosody system essential to classical Arabic poetry, where poets must adhere to strict rhythmic patterns, unlike many free-verse traditions. The goal of this system is to preserve the musicality and structure of Arabic poetry. The creation of this theory is attributed to *al-Khalīl ibn Ahmad al-Farāhīdī* (died between 776 and 791 AD). *al-'Arud* is not only used for Arabic poetry but also for Ottoman, Persian, Urdu, and other Eastern languages [15].

Although this theory dominates the specifications of metered poetic compositions, it has limitations that several authors have pointed out for various reasons. The metrics of Arabic poetry have long been considered only from a theoretical perspective and wrongly judged as a faithful reflection of the poets' practice [16]. Arabic poetry existed before this theory, and the real poetic compositions are found in the poems themselves, rather than being uniformly dictated by the theory. Some poetic forms are very rare, even though the theory presents them in the same way as more frequent forms [17]. Additionally, the theory is complex [18] and often justifies its own choices without adequately considering real practice. For example, meters like *al-Muqtadab* and *al-Mudari'* are extremely rare in poetry, and their existence is almost purely theoretical [19].

1.4 Problem statement

While studies on Arabic metrical poetry have examined both semantic and structural dimensions, they have primarily done so from a linguistic perspective. Few studies have explored automatic computational approaches for analyzing Arabic metrical poetry structures [20]. Specifically, there is a notable gap in research regarding the finite sets of data on meters, patterns, and potential combinations needed for deeper exploration. Further research and technical advancements are essential to make metered Arabic poetry more accessible for educational and research purposes.

The complexity of the Arabic metrical system, coupled with computational challenges associated with the language, has contributed to the scarcity of studies examining the structural aspects of Arabic metrical poetry. Furthermore, there is a clear bias in computational poetry analysis towards English and other widely studied languages, leaving Arabic, despite being a major global language spoken by millions, underrepresented in this field.

This thesis addresses this gap by focusing on metered Arabic poetry to advance natural language processing techniques for this rich literary tradition. It aims to enhance the accessibility of Arabic poetry through exploration, visualization, and analysis, specifically targeting the structural components such as meters and patterns, while deliberately excluding the semantic aspects. To the best of our knowledge, this thesis is the first to systematically address the data requirements for identifying and analyzing Arabic metrical poetry and to present methods for its visualization, thereby making it accessible to a broader audience, including non-native Arabic speakers.

The overall problem statement of this thesis can be summarized as the following question: What type of data and processing framework could alleviate the analysis of metered Arabic poetry? To address this question, we break it down into the following five sub-questions:

Q1: How can meters in Arabic poetry be characterized? **Q2**: How can a meter be automatically detected? **Q3**: What are the relevant data metrics that allow for the comparison of different poetic compositions? **Q4**: What are the relevant visualizations that allow for the comparison of different poetic compositions?, **Q5**: What is an adequate methodology to devise and use an Arabic poetry corpus?

1.4.1 Metrical system data

Before detecting the meter of a poem or verse, we need a well-structured reference of the Arabic metrical system, which serves as the ground-truth for meter detection. This system organizes and stores the characterization of all meters. The reference data should encompass the fundamental elements, such as syllables, patterns, meters, and their possible combinations as documented in Arabic literary tradition. Additionally, the collection of reference data must go beyond established theoretical principles and anticipate other variants found in the literature. To ensure accuracy and efficiency, it is crucial to have a comprehensive set of data and metadata objects, with their associations clearly defined and organized. This approach helps avoid redundancies while maintaining the dataset at an optimal level, neither excessive nor insufficient. This consideration leads to the question:
Q1: How can meters in Arabic poetry be characterized?

1.4.2 Meter detection

The exploration and analysis of a metered Arabic poetic structure inevitably require the detection of this structure, which involves identifying the meter of the verse or poem in question. This process entails several steps, ranging from text preparation to the actual detection. The preparation phase poses challenges due to the need for vocalization of the text, as well as the requirement to retain only what is pronounceable and rewrite the text accordingly before extracting the syllabic sequences, which serve as input data for the meter identification modules. These modules use different techniques, including exact matching verification and similarity matching. This raises the question:

Q2: How can a meter be automatically detected?

1.4.3 Structure analysis

Poetic structures vary based on the meter, the number of positions, and the arrangement of patterns. Variations can occur between different poems or even between verses within the same poem, presenting challenges in terms of order and complexity in poetic composition.

Metrics are therefore needed to quantify these features, whether through data analysis or visualizations, to distinguish compositions effectively. Attributes such as variability, repetition, and pattern usage may be considered, with only the most relevant metrics retained to characterize the compositions accurately. Since metrics directly impact analysis results, selecting the appropriate ones is crucial and must be tailored to the specific context being studied.

Given the variety of meters and patterns, both in terms of quantity and positioning, several data points may appear missing for certain positions or even for a hemistich, particularly in meters that lack one. However, this absence reflects a reality that must be acknowledged. In this context, missing values represent valuable information and do not need to be substituted. Another challenge arises from the variation in the number of verses within poems, which causes meters used in longer poems to be overrepresented compared to shorter ones, especially when the latter employ less common meters. Since the data is not balanced for valid reasons, metrics should account for this imbalance in both calculations and interpretation of the results. We address these issues through the following question:

Q3: What are the relevant data metrics that allow for the comparison of different poetic compositions?

1.4.4 Poetry visualization

Visualizing the structure of metered Arabic poetry serves both artistic and analytical purposes. Metered poetry, governed by rhythmic patterns, can be difficult to fully appreciate or analyze through text alone. Visualization provides a means to highlight these structures, making them more accessible for interpretation and aesthetic appreciation. By transforming the rhythmic patterns of Arabic poetry into visual elements such as shapes and colors, it offers a new medium for experiencing the musicality and rhythm of the poetry, enhancing its aesthetic value and engaging with its structure beyond linguistic boundaries.

From an analytical perspective, visualization is crucial for examining the complex patterns of metered poetry, which are structured around precise combinations of long and short syllables. It allows researchers to map the metrical system graphically, facilitating the identification of structural similarities, the comparison of different poems, and the discovery of new metrics that may not be immediately apparent through traditional text-based methods.

For students and researchers, visualization simplifies the complexities of Arabic meter by providing an intuitive approach to engaging with the prosodic structure. It offers a clear, visual breakdown of metrical forms, aiding in the understanding of how these patterns are constructed and function within the poem. This makes it an effective educational tool for grasping the intricacies of Arabic poetry's metrical systems.

However, visualization presents several challenges. Accurately representing metrical structures requires precise depiction of syllables and patterns, ensuring that complex forms remain clear and easy to interpret. Additionally, the lack of readily available digital corpora for Arabic poetry complicates the automation of visualization processes. Furthermore, the success of visualization depends on how effectively it conveys the structural elements of poetry. Ensuring that graphical representations of poetic structure are easily interpretable is essential for meaningful analysis and educational use. This allows us to raise the following question:

Q4: What are the relevant visualizations that allow for the comparison of different poetic compositions?

1.4.5 Poetry corpus

A comprehensive Arabic poetry corpus is crucial for research in computational linguistics, metrical analysis, and natural language processing (NLP). It serves as a foundational resource for validating the metrical system, testing meter detection algorithms, visualizing and analyzing poetic structures. A well-developed poetry corpus enables the verification of pattern combinations that define each meter, ensuring adherence to metrical rules while identifying deviations.

One key application of the corpus is testing algorithms for automatic meter detection, which rely on accurate syllabic segmentation and pattern recognition. However, building a comprehensive corpus presents significant challenges. The lack of vowelization in most available corpora is a critical issue, as vowels are essential for metrical scansion and identifying the correct meter. While manual vowelization is accurate, it is time-consuming. Automated tools exist but are often imperfect, particularly in poetry, where language tends to be more flexible.

Additionally, creating a large and diverse corpus is labor-intensive, involving the collection, digitization of poems across various meters, poets, and eras. Another challenge is the imbalance in meter representation: frequent meters dominate, while rarer ones are underrepresented, limiting the analysis of less common forms. The selection of poems is further complicated by variations in length, requiring careful curation to ensure that the corpus represents the broader poetic tradition accurately.

The existence of multiple versions of the same poem also complicates corpus creation, making the choice of which version to include a critical decision. Moreover, most existing corpora are limited to textual data, lacking annotations that could support additional analyses, such as semantic exploration. Finally, whether to use a whole corpus or a partial selection depends on the specific research objective. This raises the question:

Q5: What is an adequate methodology to devise and use an Arabic poetry corpus?

1.5 Contributions

The contributions of this thesis are bundled together in the framework *Arabic Meters Identification System (AMIS)*, which evolves from meter detection to poetry visualization and analysis. *AMIS* is a set of data and software components designed for processing metered Arabic poetry. Figure 1.1 provides a layered overview of the various contributions encompassed by *AMIS*, along with references to the chapters where they are described.

The main contribution is the Arabic metrical system data, detailed in Chapter 2. This fundamental data component structures the data and metadata of

Arabic meters and patterns. A second contribution, discussed in Chapter 2, is the natural language processing components, which prepare poetry text for accurate meter detection using different matching techniques. A third contribution is the analysis of poetry structures, particularly the clustering of poetic compositions, described in Chapter 3. The final contribution, outlined in Chapter 4, focuses on the visualization of poetic structures and how it can be used to compare different compositions and uncover new metrics for data analysis.

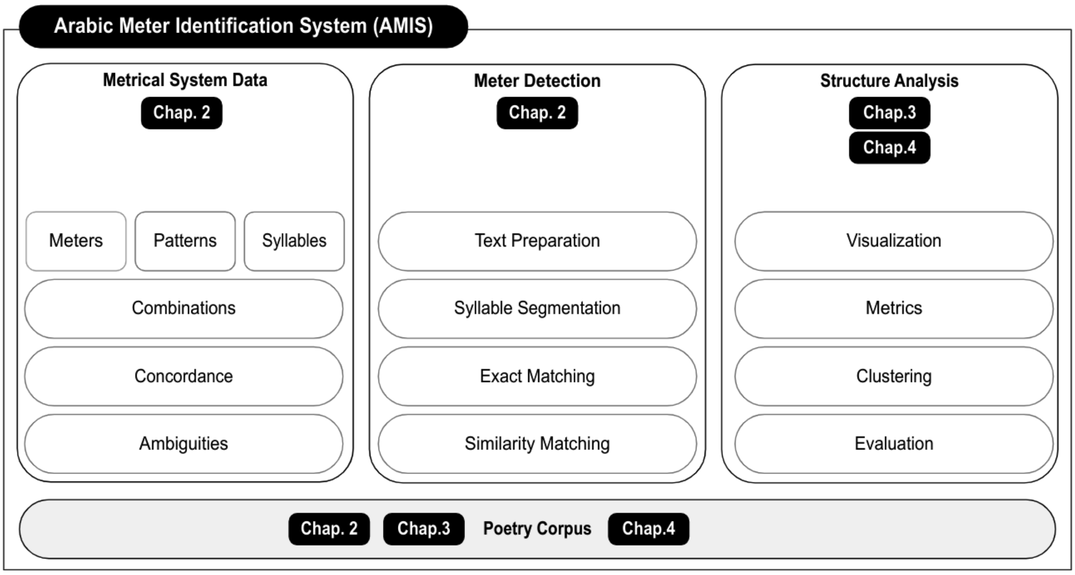


FIGURE 1.1: *AMIS* Framework

Additionally, this thesis offers two further contributions in the form of poetry corpora in both textual and graphical formats (see Chapters 2, 3, and 4).

1.5.1 Metrical system data

This contribution addresses the research question **Q1**: How can meters in Arabic poetry be characterized? It focuses on the individual data elements that serve as the building blocks of the metrical system, alongside the possible associations between them as described in the literature. After completing the data collection and pruning processes, the final set of exhaustive combinations serves as a reference dataset for meter detection, structure analysis, and composition visualization.

Meters. The 16 canonical meters from traditional theory formed the starting point for data collection. Additionally, 13 variant meters, which result from alterations to the original meters for poetic purposes, were also included.

Patterns. Beyond the 8 original theoretical patterns, 35 additional patterns were identified through an assessment of the metrical system, using a poetry corpus that encompasses all known meters.

Syllables. A key decision was made regarding the representation of syllables before they were incorporated. Syllables are the fundamental elements used to construct patterns, which are essentially sequences of syllables.

Combinations. The structure of the metrical system data follows a straightforward hierarchy. Syllables combine to form patterns, and the number and arrangement of patterns determine meters. The integration of these elements creates a dataset that captures the essential and unique sequences required for metrical analysis in accordance with the conventions of Arabic poetic literature.

Concordance. In metered poetry, rhyme and rhythm are often dictated by the meter. As such, specific rules must be followed to ensure that the verse is well-formed according to the requirements of the chosen meter.

Ambiguities. Since certain patterns are shared across multiple meters, there are cases where a verse's syllabic sequence may be identical for different meters. Such cases are treated as ambiguities and must be resolved using a set of priority and dominance rules.

1.5.2 Meter detection

This contribution addresses research question **Q2**: How can a meter be automatically detected? It includes all the necessary processing components, from raw poetry text to the display of detection results. These components form a continuous workflow, where the output of one step serves as the input for the subsequent step.

Text preparation. This component prepares and cleans the raw text, retaining only the necessary elements as pronounced in poetry. It reformulates the text according to well-established rules in Arabic poetic literature. Additionally, it corrects certain vocalization issues to produce a text that is formatted appropriately for the next stage: syllabic segmentation.

Syllabic segmentation. This component performs the metrical scansion. It begins with the prepared text from the previous stage and encodes it to enable syllable recognition. The output is a syllabic sequence for each verse, forming the basis for further metrical analysis.

Exact and similarity matching. Once the metrical scansion is complete and the text is represented as a syllabic sequence, this component attempts to match the sequence against the metrical system data. It uses regular expressions to control the matching process, ensuring that the number and arrangement of syllables correspond with the reference data. Initially, the system attempts an exact match, but if the meter is not detected, it switches to a similarity-based approach, allowing for a tolerance of up to two mismatched positions. However, the total number of syllables between the input and the reference must remain identical.

1.5.3 Structure analysis

This contribution emphasizes the necessity of extracting features from poetic compositions and computing metrics that enable meaningful comparisons between poems. Identifying suitable metrics is a challenging task, which can be approached through textual data exploration, visualization, or a combination of both.

Metrics, clustering, and evaluation. This contribution addresses research question **Q3**: What are the relevant data metrics that allow for the comparison of different poetic compositions? Chapter 3 details the process of clustering poetic structures based on a set of attributes selected after a rigorous assessment process to retain only those that truly characterize the structural patterns of poems. Additionally, Chapter 4 extends this work by exploring further attribute selection through visualization techniques. In both approaches, the clustering results are evaluated using established clustering evaluation metrics, ensuring the validity of the analysis.

Visualization. This contribution addresses research question **Q4**: What are the relevant visualizations that allow for the comparison of different poetic compositions? As detailed in Chapter 4, the goal is to demonstrate how visualizing poetic structures can uncover previously hidden metrics, thereby enhancing the overall analysis of poetry. Key decisions are made regarding the shapes and colors used for visualization, as these visual elements play a crucial role in interpreting poetic patterns. Additionally, the length of poems poses a challenge, as it directly impacts the size and clarity of the generated images. Despite these challenges, the resulting visualizations significantly improve clustering outcomes by offering a new perspective on structural patterns. Thus, visualization proves to be essential in poetry analysis, complementing traditional feature extraction methods based solely on textual data.

1.5.4 Poetry corpus

This contribution addresses research question **Q5**: What is an adequate methodology to devise and use an Arabic poetry corpus? The poetry corpus plays a central role throughout this thesis, appearing in nearly all chapters. Its use varies depending on the research objective: sometimes the entire corpus is utilized, while in other instances only a subset is employed. The corpus has evolved over the course of this research to accommodate the validation of results and the refinement of methods. Notably, the final version of the corpus is in the form of images, produced through visualization processes that facilitate the calculation of similarities and dissimilarities between poems. This image-based corpus allows for a novel approach to structural analysis by converting textual patterns into visual representations.

1.6 Form and structure

This thesis adopts a hybrid structure, combining a monograph format with a collection of three peer-reviewed articles. This approach allows for a cohesive presentation of the overall research objectives and methodology, while the articles provide focused, detailed analysis of specific aspects of the study. This format ensures both comprehensive coverage of the topic and the opportunity to publish key findings in a structured, scholarly manner. The thesis is divided into four parts:

Part I: Arabic metrical system and meter detection. In this part, Chapter 2 surveys current efforts in Arabic poetry meter detection and provides detailed insights into the structure and data collection of the Arabic metrical system, along with the processing steps for meter detection. This chapter addresses the challenge of automatic meter detection in metered Arabic poetry, proposing a solution based on data, rules, and processes, from the preparation of raw poetic text to the detection of specific meter variants and their related patterns.

Part II: Structure analysis. In this part, Chapter 3 reviews the existing literature on poetry clustering and classification before presenting clustering experiments on poetry compositions using metrics such as variability, repetition, and pattern usage, evaluated for their ability to distinguish one poem from another. The experiments involved various combinations of attributes with two different algorithms, assessed using established clustering validation scores.

Part III: Visualization. In this part, Chapter 4 explores how to visualize Arabic poetry compositions to uncover new insights that may lead to the discovery of new patterns. This chapter covers the preparation of syllabic data, conversion of poems into images, calculation and display of image dissimilarities, and the assessment of newly found metrics. As an evaluation approach, clustering was applied to the same corpus to compare the results obtained in Chapter 3.

Part IV: Appendices. The decision to combine a monograph with a collection of articles allowed certain content to be relocated to enhance overall readability. Some of this material, such as the code for meter ambiguity pruning (Appendix A) and detailed data on the meters and patterns of the Arabic metrical system (Appendices B and C), was originally part of the published article that forms the basis of Chapter 2. The appendices also provide supplementary research content, improving the clarity and readability of sample visualizations from the poetry corpus (Appendix D), which were moved to streamline Chapter 4, along with results from visualization experiments (Appendix E).

References

[1] Djamel Eddine Kouloughli. *L'arabe (The Arabic)*. French. Paris: Presses universitaires de France, 2007.

[2] R Blanchère and M Gaudefroy-Demombynes. *Grammaire de l'arabe classique: morphologie et syntaxe (Grammar of classical Arabic: morphology and syntax)*. French. Maisonneuve & Larose, 1975.

[3] Karin C Ryding. *A reference grammar of Modern Standard Arabic*. Cambridge University Press, 2005.

[4] World Bank. *Arab World Data*. `https://data.worldbank.org/region/arab-world`. Accessed: 2024-10-05. 2024.

[5] Janet CE Watson. *The phonology and morphology of Arabic*. Oxford University Press, USA, 2002.

[6] Elabbas Benmamoun and Reem Bassiouney. *The Routledge handbook of Arabic linguistics*. Routledge, 2018.

[7] Bernard Comrie. "Languages of the world". In: *The handbook of linguistics* (2003), pp. 19–42.

[8] Djamel Eddine Kouloughli. *grammaire de l'arabe d'aujourd'hui (Grammar of Today's Arabic)*. French. POCKET, 1994.

[9] Elinor Saiegh-Haddad and Roni Henkin-Roitfarb. "The structure of Arabic language and orthography". In: *Handbook of Arabic literacy: Insights and perspectives* (2014), pp. 3–28.

[10] Nizar Y Habash. *Introduction to Arabic natural language processing.* New York: Morgan & Claypool Publishers, 2010.

[11] Michael Zwettler et al. "The oral tradition of classical Arabic poetry". In: *Columbus, Ohio* 173 (1978).

[12] Reynold A Nicholson. *A literary history of the Arabs.* Routledge, 2013.

[13] Huda J Fakhreddine. *The Arabic prose poem: poetic theory and practice.* Edinburgh University Press, 2021.

[14] Ahmad Al Hachimi. *The Golden balance in the craft of Arabic poetry.* Arabic. Cairo: Maktabat Al Adab, 1997, p. 164.

[15] Atakan Kurt and Mehmet Kara. "An algorithm for the detection and analysis of arud meter in Diwan poetry". In: *Turkish Journal of Electrical Engineering and Computer Sciences* 20.6 (2012), pp. 948–963.

[16] Bruno Paoli. *De la théorie à l'usage: essai de reconstitution du système de la métrique arabe ancienne (From theory to practice: an essay on the reconstruction of the ancient Arabic metric system).* French. Vol. 234. Institut français du Proche-Orient, 2008.

[17] Mustapha Harakat. *Poetry rhythms.* Arabic. Al Dar Al Thaqafiya LiNachr, 1998, p. 156.

[18] Kamal Abou Al Dib. *On the rhythmic structure in Arabic poetry.* Arabic. Beirut: Dar Al'Ilm Lilmalayin, 1974.

[19] Dima Choukr and Bruno Paoli. "La métrique arabe au xiiie siècle après Al Khalil: entre tradition et innovation (Arabic metrics in the 13th Century after Al Khalil: between tradition and innovation)". French. In: *Bulletin d'études orientales* Tome LIX (2010), pp. 11–15.

[20] Djamel Eddine Kouloughli. "Traitement automatique de la métrique arabe: réalisations et perspectives (Automatic Processing of Arabic Metrics: Achievements and Perspectives)". French. In: *Bulletin d'études orientales* 59 (2010), pp. 17–32.

Part I

Arabic metrical system and meter detection

Chapter 2

Rule based meter detection

Based on, with extensions:
Pattern matching in meter detection
of Arabic classical poetry
17th ACS/IEEE International
Conference on Computer Systems
and Applications (AICCSA)
November, 2020.

2.1 Introduction

Meter is a fundamental feature in Arabic classical poetry, defining the rhythmic structure and harmony of verses. Its detection is crucial for both academic and practical applications. In teaching, meter helps students understand the formal aspects of poetry and how rhythm shapes meaning and emotional tone [1]. In poem classification and authorship recognition, it serves as a key attribute for categorizing and attributing works based on their rhythmic patterns [2]. Additionally, meter plays an important role in computational aesthetics, aiding in the evaluation of poetic qualities and the generation of new compositions [3].

Understanding structure is fundamental in any field, as it enables systematic data collection, exploration, and analysis. In poetry, recognizing meter not only advances research in literature, linguistics, and computational poetics, but also enhances analysis and visualization. It facilitates comparisons between compositions, uncovers the rhythmic framework, and provides valuable insights into how poets structure their works.

Automatic meter detection in Arabic poetry presents unique challenges that require a comprehensive approach to both data collection and processing. In this article, we introduce the *Arabic Meters Identification System (AMIS)*, a framework that leverages an exhaustive dataset of pattern combinations to characterize the Arabic metrical system and employs pattern matching and similarity measures for meter detection. One of the key complexities in meter detection arises from the fact that each meter can have numerous pattern variations, which often lead to sequence redundancies and conflicts between meters [4]. Therefore, it is crucial to disambiguate these meter variants while ensuring concordance between the two hemistichs (parts) of the verse.

The intricate morphology of Arabic poses significant challenges for computational natural language processing (NLP). The ambiguity in the writing system, combined with Arabic's rich morphology and complex word formation, makes developing computational approaches both difficult and resource-intensive [5].

In Arabic metered poetry, even if a verse part ends with a short vowel, it is often pronounced and processed as a long vowel [6]. Consequently, the final phoneme of each verse part must be adapted accordingly [7]. When dealing with a single verse written as a continuous sentence, the main challenge is distinguishing between the two verse parts. The metrical system dataset must be adapted to handle cases where the first part ends with either a short or long vowel.

As of the publication of this article in 2020, the best reported accuracy for meter detection in a single poetry verse was 75% [8]. *AMIS* enhances previous work by addressing both data and processing aspects. On the data side, it builds an exhaustive dataset of pattern combinations for all meter variants, and prunes it by eliminating meter conflicts and non-concordant patterns between verse parts. On the processing side, *AMIS* focuses on phonological preparation of verses, syllabic segmentation, exact pattern matching, and similarity measures to mitigate imperfections in text preparation.

This chapter aims at addressing the following research questions:

RQ1: How can meters in Arabic poetry be characterized?

RQ2: How can a meter be automatically detected?

This chapter is structured as follows: Section 2.2 gives an overview of Arabic meters, Section 2.3 presents previous work related to this research, Section 2.4 describes *AMIS* framework approach, Section 2.5 highlights the methodology of poetry corpus collection, Section 2.6 summarizes results, Section 2.7 discusses the findings and finally, Section 2.8 wraps up with a conclusion.

2.2 Background

In classical Arabic poetry, syllables, patterns, and meters are the fundamental building blocks of the poetic composition.

2.2.1 Syllables

In prosody and phonology, syllables can be represented in various ways to capture their structure and length. One common representation is the CV notation, where C represents a consonant and V represents a vowel. For simplification, we use an alternate notation with three letters: (L) for long, (S) for short, and (A) to indicate the absence of vocalization, which is rare in Arabic poetry [6]. The mapping between the two notations is based on the prosodic tradition in metered Arabic poetry, as shown in Table 2.1.

In this research, we opted to use syllables as the basic units of analysis instead of the traditional *pegs* (*watid*) and *cords* (*sabab*) found in classical Arabic prosody. In the context of Arabic poetry, a peg is either the syllable sequence *LS* or *SL*, and a cord represents the syllable sequence *SS* or *L*. While these elements are foundational in classical Arabic poetry, we chose syllables for simplicity and clarity, ensuring that the analysis remains accessible and consistent across different metrical patterns. Further, according to Paoli [9]: "*in the current state of our knowledge, there is no reason to believe that the Arabic metric system was anything other than purely quantitative, consisting of the coded succession of short and long syllables, or, more precisely, of fixed and variable positions*".

Table 2.1: Syllables notation

Notation	Equivalent representation
CV	*S*
CVV	*L*
CVC	*L*
CVVC	*LA*
CVCC	*LA*

Although the Arabic language is written from right to left, the syllables and patterns are represented from left to right to make the reading of structures more accessible.

2.2.2 Patterns

A pattern refers to a named sequence of syllables; for example, the pattern *mafā'īlun* is composed of the four syllables *SLLL*.

As part of this research, we utilized the full set of patterns displayed in Table 2.2, along with additional variants, excluding two specific patterns that include two elements separated by a space character: *LLS L (mustaf'i lun)* and *LS LL (fā'ilā tun)*. These patterns are phonetically identical to *LLSL* *(mustaf'ilun)* and *LSLL (fā'ilātun)*, respectively, and their inclusion would not provide any computational advantage. Although these patterns theoretically exist, their actual usage is exceedingly rare. Their inclusion in the system is largely to maintain consistency with the peg-and-cord division prescribed by traditional Arabic metrical theory [6].

Table 2.2: Original poetry patterns

Pattern	**Syllables**
fā'ilātun	*LSLL*
fā'i lātun	*LS LL*
fā'ilun	*LSL*
fa'ūlun	*SLL*
mafā'īlun	*SLLL*
maf'ūlātu	*LLLS*
mufā'alatun	*SLSSL*
mustaf'ilun	*LLSL*
mustaf'i lun	*LLS L*
mutafā'ilun	*SSLSL*

2.2.3 Meters

Meters are identified by unique names and serve to establish the rhythm of the verse. The meter is a set of ordered patterns in a verse. In most poems, the verses adhere to one consistent meter. The occurrence of multiple meters within a single poem is uncommon. The meter is related to an entire verse, and the number of pattern positions in the verse depends on the meter. Since a pattern may exist in more than one meter at one or more positions, any change in verse patterns arrangement may have an impact on the verse's structure. For instance, the syllables sequence *(SSLSL) (SSLSL) (SSLSL) (SSLSL) (SSLSL) (SSLSL)* represents an instance of the meter *al-Kāmil* with six pattern positions, three in each hemistich. In this particular case, the pattern *mutafā'ilun (SSLSL)* is repeated six times.

Original meters

Tables 2.3, 2.4 and 2.5 show the 16 original meters of the Arabic metrical system *al-'Arud* [7].

TABLE 2.3: Original poetry meters with 8 positions

Meter	First part	Second part
al-Tawīl	*fa'ūlun mafā'īlun fa'ūlun mafā'īlun*	*fa'ūlun mafā'īlun fa'ūlun mafā'īlun*
al-Basīt	*mustaf'ilun fā'ilun mustaf'ilun fā'ilun*	*mustaf'ilun fā'ilun mustaf'ilun fā'ilun*
al-Mutaqārib	*fa'ūlun fa'ūlun fa'ūlun fa'ūlun*	*fa'ūlun fa'ūlun fa'ūlun fa'ūlun fa'ulun*
al-Mutadārak	*fā'ilun fā'ilun fā'ilun fā'ilun*	*fā'ilun fā'ilun fā'ilun fā'ilun*

TABLE 2.4: Original poetry meters with 6 positions

Meter	First part	Second part
al-Madīd	*fā'ilātun fā'ilun fā'ilātun*	*fā'ilātun fā'ilun fā'ilātun*
al-Wāfir	*mufā'alatun mufā'alatun fa'ūlun*	*mufā'alatun mufā'alatun fa'ūlun*
al-Kāmil	*mutafā'ilun mutafā'ilun mutafā'ilun*	*mutafā'ilun mutafā'ilun mutafā'ilun*
al-Rajaz	*mustaf'ilun mustaf'ilun mustaf'ilun*	*mustaf'ilun mustaf'ilun mustaf'ilun*
al-Ramal	*fā'ilātun fā'ilātun fā'ilātun*	*fā'ilātun fā'ilātun fā'ilātun*
al-Sarī'	*mustaf'ilun mustaf'ilun fā'ilun*	*mustaf'ilun mustaf'ilun fā'ilun*
al-Munsarih	*mustaf'ilun maf'ūlātu mustaf'ilun*	*mustaf'ilun maf'ūlātu mustaf'ilun*
al-Khafīf	*fā'ilātun mostaf'i-lon fā'ilātun*	*fā'ilātun mostaf'i-lon fā'ilātun*

Table 2.5: Original poetry meters with 4 positions

Meter	First part	Second part
al-Hazaj	*mafā'īlun mafā'īlun*	*mafā'īlun mafā'īlun*
al-Mudāri'	*mafā'īlun faa'i-laaton*	*mafā'īlun faa'i-laaton*
al-Muqtadab	*maf'ūlātu mustaf'ilun*	*maf'ūlātu mustaf'ilun*
al-Mujtath	*mostaf'i-lon fā'ilātun*	*mostaf'i-lon fā'ilātun*

Meter variants

Most of poetry verses do not conform to the original form of meters [4]. Poets use variants instead. Knowing that the Arabic metrical system can theoretically extend to other variants, we limited the scope of our research to variants commonly found in the literature, relying on five different bibliographical sources [4, 7, 10–12].

In the Arabic metrical system, in addition to their names, meters are also assigned a qualifier based on the number of positions available to accommodate patterns. A meter is considered "Complete" (*al-Tām*) when it retains all of its positions. It is classified as "Partial" in three cases: *al-Majzū'* when it loses one position per hemistich, *al-Mashtūr* when it loses half of its positions (the entire second hemistich), and *al-Manhūk* when it loses two-thirds of its positions [10]. For the sake of simplification and readability, we use the suffix "C" for complete meters and the letter "P" followed by a number for partial variants. For example, the meter *al-Rajaz* can appear in all four forms, with 6, 4, 3, and 2 positions corresponding to *al-Rajaz.C (al-Rajaz al-Tām)*, *al-Rajaz.P1 (al-Rajaz al-Majzū')*, *al-Rajaz.P2 (al-Rajaz al-Mashtūr)*, and *al-Rajaz.P3 (al-Rajaz al-Manhūk)* forms, respectively.

As indicated in the list of 29 meter variants below, the number of positions ranges from 2 to 8, with an equal distribution between the first and second hemistich in most cases. The exceptions are *al-Rajaz.P2* and *al-Rajaz.P3*, which lack a second hemistich.

- *al-Ṭawīl.C* (*al-Ṭawīl al-Tām*), 8 positions, (4,4).
- *al-Madīd.C* (*al-Madīd al-Tām*), 6 positions, (3,3).
- *al-Madīd.P* (*al-Madīd al-Majzū'*), 4 positions, (2,2).
- *al-Basīt.C* (*al-Basīt al-Tām*), 8 positions, (4,4).
- *al-Basīt.P1* (*al-Basīt al-Majzū'*), 6 positions, (3,3).
- *al-Basīt.P2* (*al-Basīt al-Mukhla'*), 6 positions, (3,3).
- *al-Wāfir.C* (*al-Wāfir al-Tām*), 6 positions, (3,3).
- *al-Wāfir.P* (*al-Wāfir al-Majzū'*), 4 positions, (2,2).
- *al-Kāmil.C* (*al-Kāmil al-Tām*), 6 positions, (3,3).
- *al-Kāmil.P* (*al-Kāmil al-Majzū'*), 4 positions, (2,2).
- *al-Hazaj.C* (*al-Hazaj al-Tām*), 4 positions, (2,2).
- *al-Rajaz.C* (*al-Rajaz al-Tām*), 6 positions, (3,3).
- *al-Rajaz.P1* (*al-Rajaz al-Majzū'*), 4 positions, (2,2).
- *al-Rajaz.P2* (*al-Rajaz al-Mashṭūr*), 3 positions, (3,0): no second hemistich.
- *al-Rajaz.P3* (*al-Rajaz al-Manhūk*), 2 positions, (2,0): no second hemistich.
- *al-Ramal.C* (*al-Ramal al-Tām*), 6 positions, (3,3).
- *al-Ramal.P* (*al-Ramal al-Majzū'*), 4 positions, (2,2).
- *al-Sarī'.C* (*al-Sarī' al-Tām*), 6 positions, (3,3).
- *al-Munsarih.C* (*al-Munsarih al-Tām*), 6 positions, (3,3).
- *al-Khafīf.C* (*al-Khafīf al-Tām*), 6 positions, (3,3).
- *al-Khafīf.P* (*al-Khafīf al-Majzū'*), 4 positions, (2,2).
- *al-Muḍāri'.C* (*al-Muḍāri' al-Tām*), 4 positions, (2,2).
- *al-Muqtaḍab.C* (*al-Muqtaḍab al-Tām*), 4 positions, (2,2).
- *al-Mujtath.C* (*al-Mujtath al-Tām*), 4 positions, (2,2).
- *al-Mutaqārib.C* (*al-Mutaqārib al-Tām*), 8 positions, (4,4).

- *al-Mutaqārib.P* (*al-Mutaqārib al-Majzū'*), 6 positions, (3,3).
- *al-Mutadārak.C* (*al-Mutadārak al-Tām*), 8 positions, (4,4).
- *al-Mutadārak.P1* (*al-Mutadārak al-Majzū'*), 6 positions, (3,3).
- *al-Mutadārak.P2* (*al-Mutadārak al-Mashtūr*), 4 positions, (2,2).

2.3 Related work

Alnagdawi et al. [8] proposed meter detection methods based on the sixteen theoretical meters and patterns in three steps. The first step is text conversion in order to keep only pronounced letters. The second step is the segmentation phase where text is converted to syllables. Meter is detected in the last step by comparing the syllables sequence with the grammar stored previously. The dataset used to evaluate meter detection consists of 128 verses from different Arabic poems with a success rate of 75%. Yousef et al. [13] used recurrent neural network (RNN) to detect sixteen poetry Arabic meters and four English meters with an overall accuracy of 96.38% and 82.31%, respectively.

Further research goes beyond the theoretical meters and considers variants of original meters: Ismail et al. [14], Alabbas et al. [15], Abuata and Al-Omari [16], and Omer and Oakes [17]. For instance Ismail et al. [14] proposed a detection based on editing, consultation and knowledge base modules. They evaluate the system on 20 poems and report good results without giving figures on detection accuracy. Further, Alabbas et al. [15] encoded the prosody of each input text using *Khashan*'s method called *"numerical prosody"*. Authors report an overall accuracy of 98.6% based on the whole poem evaluation.

The rule based approach presented by Abuata and Al-Omari [16] describes an algorithm that detects the correct meter in five steps. The algorithm is based on predefined rules for text conversion in prosody form. It uses only the first part of the verse. The algorithm was evaluated on a sample of classical Arabic poems and achieves an accuracy of 82%.

In terms of meter usage and value, the authors Omer and Oakes [17] show that linguistic features based on the Arabic poetry meters are good attributes for authorship attribution. Authors argue that meter-based features outperform the usual linguistic features commonly used in authorship studies like word frequencies. They have also shown that features of Arabic classical poetry meters are suitable to distinguish authors in English as well as Arabic. Scott [18] and Ahmed et al. [19] also used the Arabic poetry meter in authorship attribution. They used meter as a feature to distinguish authors.

The main difference of the approaches is the detection phase. Some of them use only the first part of the verse [16] with the risk to detect a prose sentence or free poetry as a classical poetry verse.

Some methods use the whole poem or more than one verse [1] or rely on the writing style [20], special characters or spaces as criteria to separate the first part and the second part. This approach is challenging because of the possible different styles in the same poem.

Other methods use only theoretical set of patterns that compose each meter. These theoretical sets are rarely used. Most of them are altered and many variants exist for each meter [4].

Researches in Arabic meter identification expand more upon the detection process than the building of patterns dataset they have used as reference. Making this dataset exhaustive and available, is a step forward in research on Arabic text analysis.

2.4 Arabic Meters Identification System

AMIS is a comprehensive system of data, processes, and applications developed to accurately detect the meter of a poem or isolated verse in metered Arabic poetry. It forms the foundation of an integrated solution for the exploration and analysis of Arabic poetry. *AMIS* consists of two primary modules: one dedicated to Arabic metrical system data and rules collection and the second to meter detection processing, each further divided into specialized sub-modules.

2.4.1 Data and rules collection

The data collection module, illustrated in Figure 2.1, is designed to build a dataset of essential pattern combinations that constitute the Arabic metrical system. This dataset serves as a reference repository for accurate meter detection in any poetic verse. It is crucial that the dataset contains only the necessary entries, neither too many nor too few, to avoid redundancies, prevent meter conflicts, and ensure the uniqueness of pattern sequences. The process begins by collecting combinations from the literature, then adding additional combinations to address the case of individual verses. Finally, the dataset is pruned to maintain concordance between the parts of the verse and to remove ambiguities between meters.

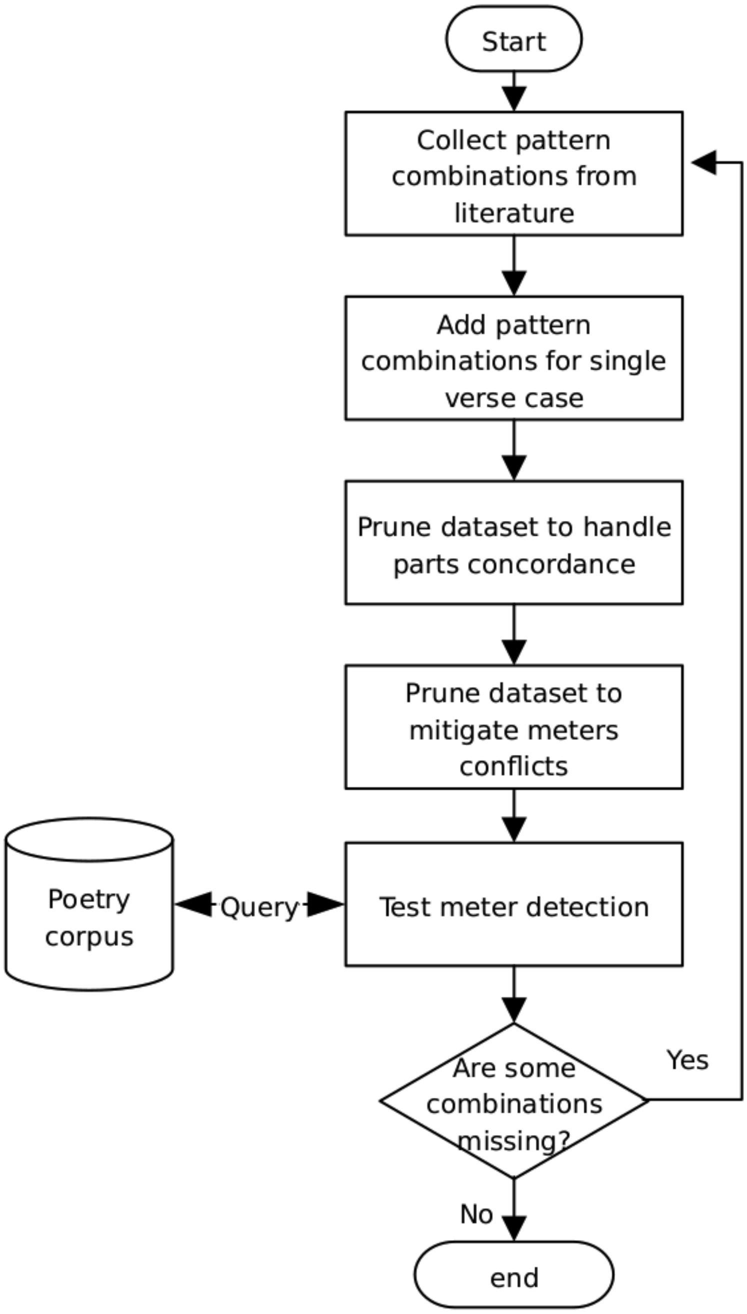


Figure 2.1: *AMIS* data and rules collection

Each row in the Arabic metrical system dataset represents an occurrence of one possible arrangement of patterns for a given meter. The dataset contains unique occurrence identification number, meter name, patterns and their corresponding syllables as shown in Table 2.6. The dataset allows up to eight patterns per row, which corresponds to the maximum number of pattern positions in metered Arabic poetry. At the end of the dataset construction, the concatenation of syllables for each meter is unique.

TABLE 2.6: Structure of Arabic metrical system dataset

Attribute	**Value**	**Attribute**	**Value**
ID	Unique ID	Meter	Meter name
First part			
fp1p	Pattern 1.1	fp1s	syllables of Pattern 1.1
fp2p	Pattern 1.2	fp2s	syllables of Pattern 1.2
fp3p	Pattern 1.3	fp3s	syllables of Pattern 1.3
fp4p	Pattern 1.4	fp4s	syllables of Pattern 1.4
Second part			
sp1p	Pattern 2.1	sp1s	syllables of Pattern 2.1
sp2p	Pattern 2.2	sp2s	syllables of Pattern 2.2
sp3p	Pattern 2.3	sp3s	syllables of Pattern 2.3
sp4p	Pattern 2.4	sp4s	syllables of Pattern 2.4

Pattern combinations collection

After structuring the dataset to accommodate the possible combinations of metrical patterns, we first inserted all the basic theoretical combinations as outlined in *al-'Arud* theory, shown in Tables 2.3, 2.4, and 2.5. We then incorporated all possible variations found in the literature [4, 7, 10–12]. For each addition, we tested the meter detection by applying it to a poetry corpus with known meters to check for redundancies and validate the combinations.

Single poetry verse

The two hemistichs of a metered Arabic poetry verse are often presented either on the same line or on separate lines. Sometimes, they are embedded within prose or formatted with various stylistic or indentation choices, making style-based automated separation unreliable. Another complication arises from the structure itself: while the number of patterns in each hemistich often matches, the syllable count may differ between them, complicating the distinction between the first and second hemistichs.

An additional challenge is found in the treatment of the final positions in each hemistich. According to a rule known as *al-ishbāʿ* (roughly translated as *Saturation*), the last position in both hemistichs often ends with a long vowel. This phenomenon occurs when a short vowel, typically on a final consonant, is extended to a long vowel, ensuring that the verse adheres to the prescribed meter and maintains rhythmic balance. It also supports rhyme consistency and allows the verse to align with the designated syllable structure without altering its meaning.

Although a vowel at the end of a verse can be lengthened to meet metrical requirements, detecting the final vowel in the first hemistich is challenging due to the lack of clear separation between the two parts. Consequently, meter detection may be inaccurate.

To address this issue, the metrical dataset accounts for the first hemistich potentially ending with either a short or a long vowel. Each primary combination ending with a long vowel in the last position of the first hemistich has a corresponding secondary combination ending with a short vowel at the same position, as shown in Table 2.7 about a combination of the meter *al-Basīt.C*. By adding a secondary combination for each primary one, the dataset's occurrences tend to double, increasing coverage to support accurate meter detection.

TABLE 2.7: Example of combinations for single verse detection

Combination	**First Part**	**Second Part**
Primary	LLSL SSL LLSL L**L**	SLSL LSL LLSL L**L**
Secondary	LLSL SSL LLSL L**S**	SLSL LSL LLSL L**L**

Verse parts concordance

In order to preserve verse harmony, there is a concordance, for some meters, between the last pattern of the first hemistich and the last pattern of the second hemistich [10]. Non-concordant combinations are not relevant for meter detection and are thus excluded from consideration. The concordant patterns at the end position of each hemistich are shown in Tables 2.8 and 2.9. An example of this concordance is found in all variants of the meter *Al-Mutadarak*, where there is no mixing, within the same verse, between the patterns *LSL, LSS (fā'ilun, fā'ilu)* and the patterns *LL, LS (fi'lun, fi'lu)*.

TABLE 2.8: Concordance of verse hemstitchs: view by syllables

Meter	**First part end**	**Second part end**
al-Tawīl.C	*SLLL, SLLS, SLSL, SLSS* *SLL, SLS*	*SLLL, SLSL* *SLL*
al-Madīd.C	*LSLL, LSLS, SSLL, SSLS* *LSL, LSS* *SSL, SSS*	*LSLL, SSLL* *LSL, LSLA, SSL, LL* *LSL, SSL, LL*
al-Madīd.P	*LSL, LSS* *SSL, SSS*	*LSL, LSLA, SSL, LL* *LSL, SSL, LL*
al-Basīt.C	*SSL, SSS* *LL, LS*	*SSL, LL* *LL*
al-Basīt.P2.	*LLL, LLS* *SLL, SLS*	*LLL* *SLL*
al-Kāmil.C	*SSL, SSS*	*SSL, LL*
al-Rajaz.C	*LLL, LLS, SLL, SLS*	*LLL, SLL*
al-Rajaz.P1	*LLLA, SLLA*	*LLLA, SLLA*
al-Sarī'.C	*LSL, LSS* *SSL, SSS* *LL, LS* *LSLA*	*LSL, LSLA, LL* *SSL, LL* *LL* *LSLA*
al-Khafīf.C	*LSL, LSS*	*LSL*

Table 2.9: Concordance of verse hemistichs: view by names

Meter	First part end	Second part end
al-Ṭawīl.C	*mafāʼīlun,mafāʼīlu,mafāʼilun,mafāʼilu* *faʼūlun,faʼūlu*	*mafāʼīlun,mafāʼilun* *faʼūlun*
al-Madīd.C	*fāʼilātun,fāʼilātu,faʼilātun,faʼilātu* *fāʼilun,fāʼilu* *faʼilun,faʼilu*	*fāʼilātun,faʼilātun* *fāʼilun,fāʼilān,faʼilun,fiʼlun* *fāʼilun,faʼilun,fiʼlun*
al-Madīd.P	*fāʼilun,fāʼilu* *faʼilun,faʼilu*	*fāʼilun,fāʼilān,faʼilun,fiʼlun* *fāʼilun,faʼilun,fiʼlun*
al-Basīt.C	*faʼilun,faʼilu* *fiʼlun,fiʼlu*	*faʼilun,fiʼlun* *fiʼlun*
al-Basīt.P2.	*mafʼūlun,mafʼūlu* *faʼūlun,faʼūlu*	*mafʼūlun* *faʼūlun*
al-Kāmil.C	*faʼilun,faʼilu*	*faʼilun,fiʼlun*
al-Rajaz.C	*mafʼūlun,mafʼūlu,faʼūlun,faʼūlu*	*mafʼūlun,faʼūlun*
al-Rajaz.P1	*mafʼūlān,faʼūlān*	*mafʼūlān,faʼūlān*
al-Sarīʼ.C	*fāʼilun,fāʼilu* *faʼilun,faʼilu* *fiʼlun,fiʼlu* *fāʼilān*	*fāʼilun,fāʼilān,fiʼlun* *faʼilun,fiʼlun* *fiʼlun* *fāʼilān*
al-Khafīf.C	*fāʼilun,fāʼilu*	*fāʼilun*

Meter ambiguities

Pattern alterations may lead to sequence redundancies across different meters [12]. For example, a sequence repeating the pattern *LLSL (mustafʼilun)* six times appears in both meters *al-Kāmil.C* and *al-Rajaz.C*. We can choose to display both meters as a result or remove one of them. To provide the most relevant meter as the result, we used pattern dominance based on the original theoretical meter compositions shown in Tables 2.3, 2.4, and 2.5. The dominant pattern is defined as the most frequently occurring pattern within the meter's sequence. The pruning code that removes ambiguous occurrences from the patterns combination dataset is detailed in Appendix A.

- *al-Kāmil.C* and *al-Rajaz.C*: Pattern *LLSL (mustafʼilun)* is more present in *al-Rajaz.C* than *al-Kāmil.C*.

- *al-Kāmil.C* and *al-Sarī'.C*: Pattern *LLSL (mustaf'ilun)* is more present in *in al-Sarī'.C* than *al-Kāmil.C*.
- *al-Kāmil.P* and *al-Rajaz.P1.*: Pattern *LLSL (mustaf'ilun)* is more present in *al-Rajaz.P1* than *al-Kāmil.P*.
- *al-Wāfir.P* and *al-Hazaj.C*: Pattern *SLLL (mafā'īlun)* is more present in *al-Hazaj.C* than *al-Wāfir.P*.
- *al-Rajaz.P1* and *al-Hazaj.C*: Pattern *SLSL (mafā'ilun)* is more present in *al-Hazaj.C* than *al-Rajaz.P1*.
- *al-Ramal.P* and *al-Madīd.P*: Pattern *LSLL (fā'ilātun)* is more present in *al-Ramal.P* than *al-Madīd.P*.

The dominance approach does not resolve all ambiguities. Table 2.10 illustrates the remaining five ambiguous cases, all of which occur within the first three positions of the first and second hemistich. In these cases, the concatenation of syllable sequences from both hemistichs leads to identical overall sequences. For instance, in case 1, both *al-Kāmil.P* and *al-Basīt.P2* produce the same sequence, *LLSLSSLSLLLLSLSSLSLL*. This creates ambiguity in detecting the correct meter because the system cannot determine whether the first or second sequence is the accurate match. Currently, the system resolves this ambiguity by displaying the first detected sequence during analysis. While this approach ensures that one plausible sequence is presented to the user, it does not guarantee that the selected sequence corresponds to the true metrical structure intended by the poet. Future improvements will aim to address these ambiguities more effectively, potentially by incorporating additional linguistic or structural rules to refine the meter detection process.

Table 2.10: Remaining meter ambiguities

Case	Meter	FP1S	FP2S	FP3S	SP1S	SP2S	SP3S
1	*al-Kāmil.P*	*LLSL*	*SSLSLL*		*LLSL*	*SSLSLL*	
1	*al-Basīt.P2*	*LLSL*	*SSL*	*SLL*	*LLSL*	*SSL*	*SLL*
2	*al-Basīt.P2*	*LLSL*	*SSL*	*SLS*	*LLSL*	*SSL*	*SLL*
2	*al-Kāmil.P*	*LLSL*	*SSLSLS*		*LLSL*	*SSLSLL*	
3	*al-Basīt.P2*	*LLSL*	*SSL*	*SLS*	*SLSL*	*SSL*	*SLL*
3	*al-Kāmil.P*	*LLSL*	*SSLSL*		*SSLSL*	*SSLSLL*	
4	*al-Kāmil.P*	*LLSL*	*SSLSLS*		*SSLSL*	*SSLSLL*	
4	*al-Basīt.P1*	*LLSL*	*SSL*	*SLSS*	*SLSL*	*SSL*	*SLL*
5	*al-Muqtadab.C*	*SLLS*	*LSSL*		*SLLS*	*LSSL*	
5	*al-Mutaqārib.P*	*SLL*	*SLS*	*SL*	*SLL*	*SLS*	*SL*

2.4.2 Meter detection processing

Beside the patterns combinations dataset that characterizes the Arabic metrical system, *AMIS* handles all the processing steps of meter detection as shown in Figure 2.2.

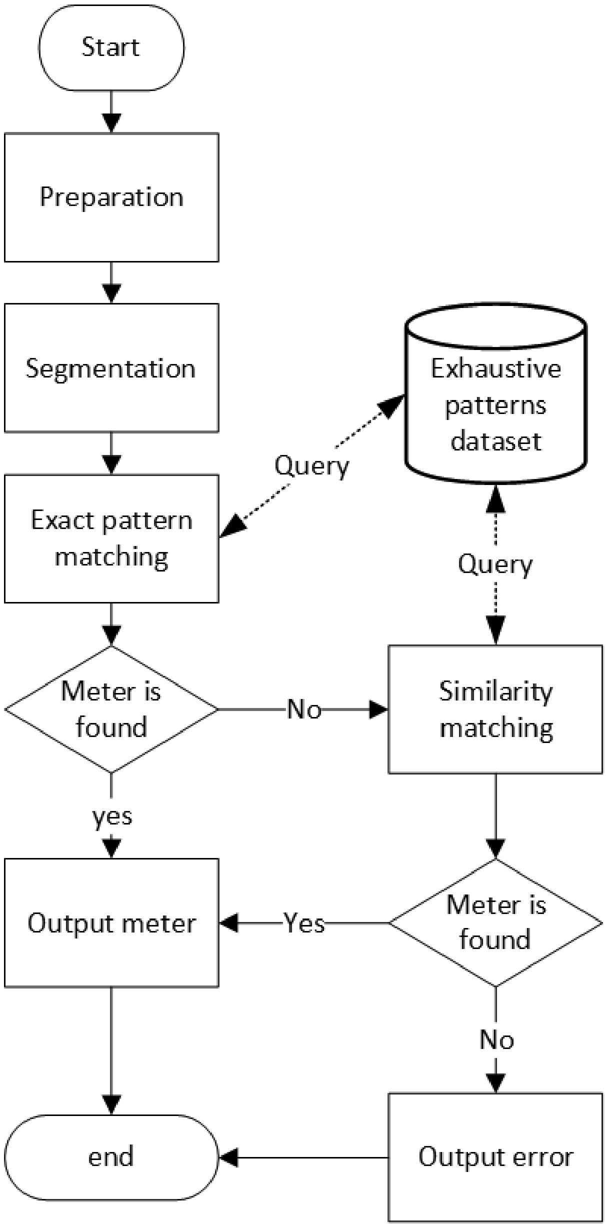


FIGURE 2.2: *AMIS* meter detection processing

Text preparation

The objective of this sub-module is to retain only the phonetically pronounced components of the text. The output is formatted specifically for prosodic analysis, focusing on preserving phonemes to facilitate accurate syllabic segmentation in subsequent steps. This step is crucial in removing non-pronounced elements, ensuring that only the sounds relevant to meter detection are retained, thereby optimizing the text for precise metrical analysis in metered Arabic poetry.

The text preparation process consists of several essential steps. The initial step focuses on cleaning the raw verse text by preserving only Arabic characters, spaces, and diacritics while removing any extraneous symbols, numbers, or punctuation. This cleaning step ensures the text is properly formatted for subsequent analysis and meter detection. By restricting the content to Arabic script and diacritics, the process eliminates potential sources of noise or interference, thereby enhancing the accuracy of meter detection and other linguistic analyses.

The second step focuses on letter normalization to standardize character representation. For instance, the letter *Alif Wasla* (Unicode: 0671), which appears at the beginning of certain words but is silent, is converted to a regular *Alif* (Unicode: 0627). Similarly, the elongated *Alif* (Unicode: 0622) is normalized to an *Alif* followed by a long vowel. This step ensures uniformity in letter representation and eliminates potential inconsistencies that could result from variations in character encodings or representations within the text.

The third step, known as *Nunation*, involves replacing the double diacritics (*tanween*) at the end of a word with the consonant carrying the double diacritic, followed by the letter “N”, a simple diacritic, and a vowel absence marker (*Sukun*). This transformation is essential for aligning the metrical pattern of the verse, especially in cases where the final syllable requires adjustment to meet the prosodic rules of metered Arabic poetry.

The fourth step addresses vowel elongation for certain words typically written with short vowels but pronounced with long vowels. For example, the word *Lakin* (however), written with a short vowel on the letter “L”, becomes *Lākin* with a long vowel. This process relies on a dictionary of such words, containing both their original and adapted elongated forms.

The fifth step focuses on rewriting the definite article to account for the type of letter that follows: either a *lunar* or *solar* letter [21]. The Arabic definite article consists of the two letters *"al"*. When a lunar letter follows, the article remains unchanged. However, when a solar letter follows, the "l" in *"al"* is assimilated into the initial consonant of the noun, resulting in a doubled consonant sound [22]. For example, we pronounce *an-Nār* (fire) instead of *al-Nār*. Additionally, adjustments are made to the definite article depending on the word's position in the text and whether it is preceded by a particle, ensuring the prosodic structure remains intact for accurate meter detection.

The sixth step involves reformulating the emphasis character, *Shadda*, by doubling the consonant that bears the emphasis. This adjustment ensures that the correct phonetic representation is maintained.

The seventh step addresses sandhi, acknowledging that pronunciation in Arabic extends beyond individual word boundaries. This step considers how the final sound of one word can influence or merge with the initial sound of the following word to create a seamless flow of pronunciation. This adjustment is crucial for maintaining the rhythmic integrity of the verse, as it ensures that phonetic transitions align with the intended meter and flow of the poetry.

The final step manages vowel elongation, or letter saturation, by converting short vowels into long vowels at the end of a verse. Additionally, this process addresses specific cases, such as the elongation of the letter "H" in masculine possessive pronouns and "M" in plural forms. These adjustments are essential for maintaining the rhythmic and metrical integrity of the verse.

Syllabic segmentation

The syllabic segmentation or metrical scansion sub-module consists of two steps. The first step processes the input text in its prosodic form, encoding vowels as 0 and 1 to capture their length and occurrence. The second step converts the resulting binary code into syllables, ensuring accurate syllable identification. All verse preparation steps are detailed in Algorithm 1.

Algorithm 1 Verse preparation: from raw text to syllables

```
/* Text preparation */
text ← raw(verse)
text ← removeNonArabicChars(text)
text ← removeSpecialChars(text)
text ← removePunctuation(text)
text ← removeDigits(text)
text ← normalizeLetters(text)
text ← rewriteNunation(text)
text ← elongateVowels(text)
text ← rewriteDefiniteArticle(text)
text ← rewriteEmphasis(text)
text ← handleSandhi(text)
text ← handleSaturation(text)

/* Vowels encoding */
for all vowels in text do
  if vowel = "Short" then
    code ← code + "1"
  else
    code ← code + "0"
  end if
end for

/* Syllabic segmentation */
syllables ← replace(code, "10", "L")
syllables ← replace(syllables, "1", "S")
syllables ← replace(syllables, "0", "A")
```

Exact pattern matching

This sub-module queries the Arabic metrical dataset to identify an occurrence that meets three conditions. The first condition ensures that the input syllable count matches the syllable count of at least one occurrence stored in the dataset. The second and third conditions verify that the syllables of the verse align with regular expressions applied to the stored pattern sequences, checking both from start to end and from end to start. The output includes the meter name, the verse composition, and an indication of exact matching without errors. For example, consider the following syllable sequence: *LLSLLLSLSSLSLSSLSLSSLSLSSLSL*. The system searches for occurrences with a length of 28 syllables, ensuring that the sequence aligns with the regular expression checks in both directions.

Similarity matching

When exact similarity matching fails, the system relaxes the strict matching rule and attempts a similarity comparison, allowing for approximate matches with a tolerance of up to two differences. In this relaxed search, the system identifies the closest pattern combinations that share the same number of syllables and positions, permitting at most two discrepancies. The output result displays the meter name along with an indication of one or two mismatches. For example, consider the following syllable sequence:
*SSLSLSSLSLSSLSLLLSLS**L**LSLLLSL*. The closest sequence in the dataset is:
*SSLSLSSLSLSSLSLLLSLS**S**LSLLLSL*, differing by only one syllable.

Verse	ذُو العَقْلِ يَشْقَى فِي النَّعِيمِ بِعَقْلِهِ وَأَخُ الجَهَالَةِ فِي الشَّقَاوَةِ يَنْعَمُ
Preparation	ذُلْ عَقْلِ يَشْقَى فِنْ نَعِيمِ بِعَقْلِهِي وَأَخُلْ جَهَالَةِ فِشْ شَقَاوَةِ يَنْعَمُو
Code	10101101010101101110110111011011101101110110
Syllables	LLSLLLSLSSLSLSSLSLSSLSLSSLSL
Detection	[0]: Exact matching
Meter	*al-Kāmil.C*
First part	*mustaf'ilun mustaf'ilun mutafā'ilun* LLSL LLSL SSLSL
Second part	*mutafā'ilun mutafā'ilun mutafā'ilun* SSLSL SSLSL SSLSL

Figure 2.3: Example of meter detection process

Figure 2.3 illustrates each step in the meter detection process for a known verse composed by *al-Mutanabbī*, which translates to *"The wise man suffers in bliss with his mind, while the ignorant revels in hardship"*.

2.5 Poetry corpus collection

A poetry corpus is an essential dataset for evaluating meter detection and assessing combinations within the Arabic metrical system. We created a vocalized corpus of metered Arabic poetry using available online sources, where the meter of each poem is specified [23], and supplemented missing information about verses and vocalization with alternative sources [24] and print books.

2.5.1 Challenges in poetry collection

Building such a corpus is a time-consuming task due to the numerous challenges encountered.

- A ready-to-use poetry corpus in a format suitable for natural language processing (NLP) was not available, necessitating manual collection.
- All poems in the corpus needed to be fully vocalized. This was done manually, as no pre-vocalized version existed.
- Some poems contained spelling mistakes, which were identified and corrected.
- The corpus was constructed to represent the majority of known meters in Arabic poetry. However, the distribution of meters within the corpus may be unbalanced due to poets' preferences for using certain meters more frequently, resulting in overrepresentation of common meters and underrepresentation of rare ones.

- For certain meters, such as *al-Mutadārak.P1* (*al-Mutadārak al-Majzū'*), no valid verses were found at the time of corpus compilation, and poems labeled with this meter were discovered to be incorrectly attributed.
- When multiple versions of a poem existed, decisions were made regarding which version to include. These versions sometimes differed in the number of verses or their sequence. We chose the version most frequently cited in the literature.
- In some cases, different sources attributed the same poem to multiple poets, requiring a decision on which attribution to follow. This was done by comparing various sources and choosing the poet most consistently identified as the author.
- A decision was made on how many verses to include per poem, setting minimum and maximum limits to maintain uniformity across the corpus. We decided to include all the verses and allow the flexibility for other processing modules in *AMIS* to use either the entire poem or a subset of the verses.
- Titles of poems were not always consistent; in some instances, the title was the first part of the first verse, while in other cases, a distinct title was used.
- The announced meter for some poems might be incorrect, requiring careful validation during the collection process.
- The announced meter is always given for the entire poem, without specifying the meter for individual verses, which can lead to inconsistencies.
- Not all poets have their poetry collections published, making the acquisition of their work more difficult.

2.5.2 Used corpus

In the initial version of *AMIS*, we used a poetry corpus comprising 107 poems, 2,711 verses, 61 distinct poets, 28 meter variants, and nine poetic eras, spanning from the pre-Islamic to the modern period.

In the latest version of *AMIS*, the corpus was expanded to 3,450 verses, with corrections to vocalization errors in some poems and improvements to the text preparation module. The updated corpus now includes 126 poems by 69 distinct poets. Table 2.11 shows the distribution of verses per meter in the poetry corpus.

TABLE 2.11: Poetry corpus

Meter	**Verses**	**Meter**	**Verses**
al-Ṭawīl.C	364	*al-Ramal.C*	127
al-Madīd.C	120	*al-Ramal.P*	75
al-Madīd.P	60	*al-Sarī'.C*	295
al-Basīt.C	322	*al-Munsarih.C*	101
al-Basīt.P1	5	*al-Khafīf.C*	171
al-Basīt.P2	159	*al-Khafīf.P*	37
al-Wāfir.C	134	*al-Mudāri'.C*	39
al-Wāfir.P	85	*al-Muqtadab.C*	102
al-Kāmil.C	257	*al-Mujtath.C*	88
al-Kāmil.P	108	*al-Mutaqārib.C*	102
al-Hazaj.C	135	*al-Mutaqārib.P*	55
al-Rajaz.C	110	*al-Mutadārak.C*	79
al-Rajaz.P1	174	*al-Mutadārak.P1*	0
al-Rajaz.P2	76	*al-Mutadārak.P2*	20
al-Rajaz.P3	50		

2.6 Results

In this research, we used the 16 original meters known in literature, along with all their variants, to characterize Arabic meters. The Arabic metrical system contains essential components that allow for accurate detection of the meter of any single verse in metered Arabic poetry. In the first version (V1) of *AMIS*, we identified 23,118 pattern combinations within the Arabic metrical system. Upon reviewing the dataset for the second version (V2), we added 400 additional combinations. As a result, the dataset of pattern combinations in the Arabic metrical system now contains 23,518 occurrences, as shown in Table 2.12.

Table 2.12: Arabic metrical system: data summary

Phase	Verses (V1)	Verses (V2)
Collection from literature	31,402	32,570
Single verse case	25,502	25,502
Verse parts concordance	-33,732	-34,500
Meter ambiguities	-54	-54
Total	**23,118**	**23,518**

2.6.1 Meters

Detailed Arabic metrical system data is provided in Appendix B and Appendix C, presenting views organized by meter and pattern, respectively. This content is an essential part of the research findings and is placed there to enhance the readability of the main document.

Table 2.13: Arabic metrical system: combinations

Meter	Comb.	Syl. Range	Meter	Comb.	Syl. Range
al-Tawīl.C	640	0	*al-Ramal.C*	1,536	2
al-Madīd.C	352	1	*al-Ramal.P*	384	2
al-Madīd.P	16	1	*al-Sarī'.C*	3,328	1
al-Basīt.C	1,944	1	*al-Munsarih.C*	648	1
al-Basīt.P1	1,008	2	*al-Khafīf.C*	416	1
al-Basīt.P2	324	0	*al-Khafīf.P*	48	1
al-Wāfir.C	32	2	*al-Mudāri'.C*	8	0
al-Wāfir.P	30	2	*al-Muqtadab.C*	18	0
al-Kāmil.C	442	4	*al-Mujtath.C*	48	1
al-Kāmil.P	128	3	*al-Mutaqārib.C*	1,280	2
al-Hazaj.C	288	1	*al-Mutaqārib.P*	320	2
al-Rajaz.C	7,680	0	*al-Mutadārak.C*	1,022	4
al-Rajaz.P1	858	1	*al-Mutadārak.P1*	474	4
al-Rajaz.P2	112	n/a	*al-Mutadārak.P2*	114	3
al-Rajaz.P3	20	n/a			

Total combinations: 23,518

Table 2.13 presents all pattern combinations in the Arabic metrical system data, highlighting the key possibilities for detecting the meter of any single verse in metered Arabic poetry. The syllabic range refers to the maximum difference in the number of syllables between the first and second hemistichs.

Table 2.14: Arabic metrical system: available patterns

Meter	Available patterns	Meter	Available patterns
al-Tawīl.C	6	*al-Ramal.C*	12
al-Madīd.C	10	*al-Ramal.P*	12
al-Madīd.P	7	*al-Sarī'.C*	11
al-Basīt.C	8	*al-Munsarih.C*	9
al-Basīt.P1	11	*al-Khafīf.C*	10
al-Basīt.P2	9	*al-Khafīf.P*	7
al-Wāfir.C	4	*al-Mudāri'.C*	4
al-Wāfir.P	4	*al-Muqtadab.C*	5
al-Kāmil.C	10	*al-Mujtath.C*	7
al-Kāmil.P	9	*al-Mutaqārib.C*	5
al-Hazaj.C	6	*al-Mutaqārib.P*	5
al-Rajaz.C	13	*al-Mutadārak.C*	6
al-Rajaz.P1	13	*al-Mutadārak.P1*	7
al-Rajaz.P2	8	*al-Mutadārak.P2*	7
al-Rajaz.P3	6		

Table 2.15: Arabic metrical system: syllables counts ranging from 7 to 30

Meter	Syl. Min	Syl. Max	Meter	Syl. Min	Syl. Max
al-Tawīl.C	26	28	*al-Ramal.C*	22	26
al-Madīd.C	19	22	*al-Ramal.P*	14	18
al-Madīd.P	13	13	*al-Sarī'.C*	20	24
al-Basīt.C	26	28	*al-Munsarih.C*	23	24
al-Basīt.P1	21	24	*al-Khafīf.C*	22	24
al-Basīt.P2	20	20	*al-Khafīf.P*	15	16
al-Wāfir.C	22	26	*al-Mudāri'.C*	16	16
al-Wāfir.P	17	20	*al-Muqtadab.C*	16	16
al-Kāmil.C	22	30	*al-Mujtath.C*	15	16
al-Kāmil.P	17	22	*al-Mutaqārib.C*	20	24
al-Hazaj.C	14	16	*al-Mutaqārib.P*	14	18
al-Rajaz.C	22	24	*al-Mutadārak.C*	16	24
al-Rajaz.P1	14	16	*al-Mutadārak.P1*	12	20
al-Rajaz.P2	11	12	*al-Mutadārak.P2*	8	14
al-Rajaz.P3	7	8			

Table 2.16: Arabic metrical system: patterns frequency

Pattern	Freq.	Freq. (%)	Pattern	Freq.	Freq. (%)
LLSL	20,168	13.759	*SLLS*	779	0.531
SLSL	19,877	13.560	*LS*	768	0.524
LSSL	18,602	12.690	*SL*	720	0.491
SSL	11,595	7.910	*L*	720	0.491
SSSL	11,501	7.846	*SLA*	720	0.491
SLL	9,118	6.220	*LSLS*	716	0.488
LSL	8,096	5.523	*LLS*	689	0.470
SLS	6,832	4.661	*LLLS*	444	0.303
LL	6,142	4.190	*SSLLA*	400	0.273
SSLL	5,132	3.501	*SSLA*	400	0.273
LSLL	4,984	3.400	*LSLLA*	400	0.273
LLL	3,055	2.084	*LLSLA*	396	0.270
SSS	1,773	1.210	*SSLS*	332	0.226
LSLA	1,766	1.205	*SLSSL*	120	0.082
SLLL	1,433	0.978	*SSLSS*	80	0.055
SLSS	1,354	0.924	*LLSLL*	48	0.033
LLSS	1,277	0.871	*SSLSLL*	48	0.033
SSLSL	1,232	0.84	*SSLSLA*	32	0.022
LLLA	1,200	0.819	*LLSLS*	16	0.011
LSS	1,200	0.819	*SSLSLS*	16	0.011
SLLA	1,200	0.819	*SLSSS*	8	0.005
LSSS	1,195	0.815			

Total pattern frequency: 146,584

Table 2.17: Distribution across positions of 12 patterns representing 85.35% of the total frequency

Pattern	**FP1S**	**FP2S**	**FP3S**	**FP4S**	**SP1S**	**SP2S**	**SP3S**	**SP4S**	**Total**
LLSL	4,615	3,350	1,801		4,582	3,290	2,530		20,168
SLSL	4,399	3,482	1,576	128	4,366	3,486	2,184	256	19,877
LSSL	4,309	2,905	1,738		4,276	2,866	2,508		18,602
SSL	798	2,562	1,322	903	798	2,589	1,465	1,158	11,595
SSSL	3,001	2,780			2,968	2,752			11,501
SLL	1,120	883	1,649	320	1,120	1,068	2,510	448	9,118
LSL	406	2,184	1,280	128	406	2,204	1,232	256	8,096
SLS	1,120	879	1,633	320	1,120	800	960		6,832
LL	406	388	560	452	406	412	1,966	1,552	6,142
SSLL	1,376	812	288		1,376	832	448		5,132
LSLL	1,376	816	224		1,376	840	352		4,984
LLL		144	673			284	1,954		3,055

Total pattern frequency: 146,584

Table 2.18: Distribution across positions of 12 patterns representing 10.35% of the total frequency

Pattern	**FP1S**	**FP2S**	**FP3S**	**FP4S**	**SP1S**	**SP2S**	**SP3S**	**SP4S**	**Total**
SSS		58	812	903					1,773
LSLA		54	478			85	1,149		1,766
SLLL	106	391		128	106	446		256	1,433
SLSS	72	170	912	128	72				1,354
LLSS		140	1,137						1,277
SSLSL	288	240	64		288	256	96		1,232
LLLA		32	528			128	512		1,200
LSS		48	1,024	128					1,200
SLLA		32	528			128	512		1,200
LSSS		121	1,074						1,195
SLLS	82	271		128	82	216			779
LS		12	304	452					768

Total pattern frequency: 146,584

Table 2.19: Distribution across positions of 19 patterns representing 4.30% of the total frequency

Pattern	FP1S	FP2S	FP3S	FP4S	SP1S	SP2S	SP3S	SP4S	Total
L			64	256			80	320	720
SL			64	256			80	320	720
SLA			64	256			80	320	720
LSLS	6	264	224		6	216			716
LLS		32	657						689
LLLS	6	216			6	216			444
LSLLA		32	128			48	192		400
SSLA		32	128			48	192		400
SSLLA		32	128			48	192		400
LLSLA			144				252		396
SSLS		44	288						332
SLSSL	32	24			32	32			120
SSLSS		16	64						80
LLSLL		16				32			48
SSLSLL		16				32			48
SSLSLA						32			32
LLSLS		16							16
SSLSLS		16							16
SLSSS		8							8

Total pattern frequency: 146,584

2.6.2 Evaluation

To evaluate the meter identification performance, we calculate the precision by dividing the number of correct detection by the number of all verses in the poetry corpus.

$$\text{Precision} = \frac{\text{Detected Verses}}{\text{All Verses}} \tag{2.1}$$

As shown Table 2.20, in the initial release of *AMIS*, 2,597 out of 2,711 verses were successfully identified during the exact matching phase, resulting in a precision of 95.80%. Subsequently, in the similarity phase, 95 more verses were detected, with 19 remaining undetected. Consequently, the overall precision, combining exact matching and similarity detection, reached 99.30%.

In the latest version of *AMIS*, 3,319 out of 3,450 verses were identified during the exact matching phase, achieving a precision of 96.20%. Further, in the similarity phase, an additional 130 verses were successfully detected. Only 1 verse remained undetected in this version. The combined precision of exact matching and similarity detection reached 99.97%.

TABLE 2.20: Meter detection precision

Matching	**V1 (Verses)**	**V1(%)**	**V2 (Verses)**	**V2(%)**
Exact	2,597	95.80	3,319	96.20
Similarity 1 error	78	2.90	109	3.16
Similarity 2 errors	17	0.60	21	0.61
Detected	**2,692**	**99.30**	**3,449**	**99.97**
Undetected	19	0.70	1	0.03
Total	**2,711**	**100**	**3,450**	**100**

2.7 Discussion

It is important to note that when the meter of a verse is detected, whether through exact matching or similarity, it signifies the presence of a corresponding syllabic sequence in the metrical system data. However, this does not necessarily mean that the detected meter matches the one indicated in the poetry corpus. In the corpus, the meter is typically specified for the entire poem, and it is often assumed in the literature that the first verse determines the meter for the whole composition [25]. However, this assumption is not always accurate. Although rare, cases exist where one or more verses within a poem deviate from the stated meter. Therefore, when a meter is detected, it confirms that the text is metered Arabic poetry rather than prose or free verse.

As shown in Table 2.13, the 23,518 essential combinations are distributed over 29 meter variants. The number of combinations vary significantly between meters.

The *al-Rajaz.C* meter holds the record for the highest number of possible combinations, with 7,680 occurrences, and the most available patterns, with 13 possibilities. As shown in Table 2.14, the number of available patterns ranges from 4 to 13. Due to its resemblance to prose and the ease of composition it offers, *al-Rajaz.C* and its partial variants are often referred to in the literature as the *"donkey of poets"* [10]. On the other extreme, there is the *al-Mudāri'.C* meter, which has only 8 possible combinations and 4 available patterns, suggesting that it may be more challenging to compose in this meter. Similarly, the *al-Muqtadab.C* meter offers only 18 combination possibilities with 5 available patterns. The variants *al-Madīd.P* and *al-Rajaz.P3* also provide limited possibilities, with 16 and 20 combinations, respectively, and 7 and 6 available patterns.

However, it is essential to highlight that the abundance or rarity of a meter in the metrical system data does not necessarily reflect its prevalence in actual poetic practice. The number of possibilities mentioned in the metrical dataset simply indicates that the theoretical metric system has identified these combinations for each meter. Only the real poetic practice, through an extensive study of the Arabic poetic corpus, can confirm or deny the frequent or rare use of a meter.

When examining the syllabic range in Table 2.13, we observe that it varies from 0 to 4. A value of 0 indicates that, across all combinations for a given meter, the number of syllables in the first and second hemistichs is identical.

Regarding patterns, we used a total of 43 patterns: 8 original patterns and 35 altered variants. Some patterns are used more frequently than others. Table 2.17 highlights the distribution of 12 distinct metrical patterns, which account for 85.35% of the total frequency (146,584 occurrences) across various positions in the first and second hemistich of Arabic verse. Approximately 80% of the total frequency is held by only 10 patterns: *LLSL*, *SLSL*, *LSSL*,

SSL, *SSSL*, *SLL*, *LSL*, *SLS*, *LL*, and *SSLL*. Patterns such as *LLSL*, *SLSL*, and *LSSL* are particularly frequent, especially at the beginning of both hemistichs, suggesting that these patterns may form the backbone of traditional Arabic meters. Their prominence in the initial positions likely reflects a preference for rhythmic regularity in the opening parts of verses. Additionally, the consistency of these patterns in both the first and second hemistichs indicates a balanced metrical structure between the two parts of the verse. Some positions, however, show no values due to structural constraints, where certain patterns do not appear in specific positions for some meters.

Tables 2.18 and 2.19 display the distribution of metrical patterns that account for 10.35% and 4.30% of the total pattern frequency in Arabic verse, respectively. Notably, some patterns are confined to only one part of the verse. For example, the pattern *SSLSLA* appears exclusively in the second hemistich, while 11 patterns are absent from the second hemistich: *SSLS*, *SSS*, *LSS*, *LS*, *LLS*, *SLSSS*, *LSSS*, *LLSLS*, *LLSS*, *SSLSLS*, and *SSLSS*.

Additionally, we observe that certain patterns are distributed across all eight positions, while others occur only in one or two positions with very low frequency. This suggests that some of these rare patterns may have originated from theoretical constructs rather than actual use in metered Arabic poetry.

Many Arabic meters exhibit close resemblances, and even minor variations in a verse can lead to a different meter variant. Instances where the exact matching sub-module fails in identifying the meter often trace back to the quality of verse preparation.

Text preparation in Arabic poetry poses significant challenges, primarily due to issues related to vowelization. One of the key problems arises from the placement of the emphasis marker *Shadda* and the vowel: sometimes the *Shadda* precedes the vowel, and other times it follows, leading to inconsistency in how the text is processed. Another challenge is the inconsistent use of the absence of a vowel *Sukun* on the letter *"l"* of the definite article; while it is not necessary to mark this in every instance, some texts do so while others omit it. Additionally, the placement of *tanween (nunation)* is often variable, with it sometimes appearing on the consonant and other times on the long vowel that follows. In many cases, vowelization is only partially applied, leaving key parts of the word unvowelized. All these variations complicate the preparation process, making it difficult to consistently detect the meter accurately and creating potential sources of error in meter analysis.

Furthermore, the implementation of the saturation rule faces inherent challenges, relying primarily on morphological and syntactic layers. Accurately detecting possessive pronouns or plurals necessitates a thorough part-of-speech analysis, as not all words ending with the letter *"H"* function as possessive pronouns, and not all those ending with *"M"* indicate plurals [7].

Finally, some meters are not detected, potentially due to historical transmission issues. Since Arabic poetry was primarily transmitted through oral tradition, certain poems may not have been faithfully preserved or recorded. Variations in pronunciation, regional accents, or even the memory of the poets themselves may have altered these poems over time, leading to discrepancies between the original compositions and the written versions we analyze today.

2.8 Conclusion

Meter detection is a complex process that involves multiple steps, such as phonological text preparation, vowel encoding, syllabic segmentation, and pattern matching. This research tackled this challenging task in metered Arabic poetry, focusing on both entire poems and individual verses. Two key contributions were made.

The first is the development of the metrical system data as a fundamental component that structures the data and metadata of Arabic meters and patterns. This comprehensive dataset contains 23,518 pattern combinations from Arabic literature, providing a robust foundation for meter analysis. It ensures concordance between the two hemistichs of a verse while resolving meter ambiguities.

The second contribution is the creation of natural language processing (NLP) components, which prepare the poetry text for accurate meter detection using both exact and similarity matching techniques. This approach achieved a high accuracy of 99.97%, effectively addressing variations that arise during text preparation, making the method highly reliable.

References

[1] Atakan Kurt and Mehmet Kara. "An algorithm for the detection and analysis of arud meter in Diwan poetry". In: *Turkish Journal of Electrical Engineering and Computer Sciences* 20.6 (2012), pp. 948–963.

[2] Hamid R Tizhoosh, Farhang Sahba, and Rozita Dara. "Poetic features for poem recognition: A comparative study". In: *Journal of Pattern Recognition Research* 3.1 (2008), pp. 24–39.

[3] Justine Kao and Dan Jurafsky. "A computational analysis of style, affect, and imagery in contemporary poetry". In: *Proceedings of the NAACL-HLT 2012 workshop on computational linguistics for literature*. 2012, pp. 8–17.

[4] Mustapha Harakat. *Poetry rhythms*. Arabic. Al Dar Al Thaqafiya LiNachr, 1998, p. 156.

[5] Abdelhadi Soudi, Günter Neumann, and Antal van den Bosch. "Arabic computational morphology: knowledge-based and empirical methods". In:

Arabic computational morphology: Knowledge-based and Empirical Methods. Springer, 2007, pp. 3–14.

[6] Ibrahim Anis. *Music of poetry*. Arabic. Cairo: Maktabat Al Anglo Masriya, 1952.

[7] Abdul Aziz Atiq. *The Science of Prosody and Rhyme*. Arabic. Beirut: Dar Al Nahda Al Arabia, 1987, p. 211.

[8] Mohammad A Alnagdawi, Hasan Rashideh, and F Aburumman. "Finding Arabic poem meter using context free grammar". In: *Journal of Communication and Computer Engineering* 3.1 (2013), pp. 52–59.

[9] Bruno Paoli. "Métrique arabe: au-delà du cercle vicieux des théories (Arabic Metrics: beyond the vicious circle of theories)". French. In: *Bulletin d'études orientales* 65 (2017), pp. 177–216.

[10] Ahmad Al Hachimi. *The Golden balance in the craft of Arabic poetry*. Arabic. Cairo: Maktabat Al Adab, 1997, p. 164.

[11] Ghazi Yammout. *Arabic poetry meters*. Arabic. Beirut: Dar Al Fikr Al Lobnani, 1992, p. 263.

[12] Hāshim Sālih Mannā'. *The Healer in Prosody and Rhymes*. Arabic. Beirut: Dar Al Fikr Al Arabi, 2003, p. 312.

[13] Waleed A. Yousef et al. "Learning meters of Arabic and English poems with Recurrent Neural Networks: a step forward for language understanding and synthesis". In: *CoRR* abs/1905.05700 (2019).

[14] MA Ismail et al. "Expert system for testing the harmony of Arabic poetry". In: *Journal of Engineering Sciences* 1 (2010), pp. 401–411.

[15] Maytham Alabbas, Zainab A Khalaf, and Khashan M Khashan. "BASRAH: an automatic system to identify the meter of Arabic poetry". In: *Natural Language Engineering* 20.1 (2014), pp. 131–149.

[16] Belal Abuata and Asma Al-Omari. "A rule-based algorithm for the detection of arud meter in CLASSICAL Arabic poetry". In: *International Arab Journal of Information Technology* 15.4 (2018), pp. 1–5.

[17] Ahmed Ibrahim Ahmed Omer and Michael Philip Oakes. "Arud, the metrical system of arabic poetry, as a feature set for authorship attribution". In: *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*. IEEE. 2017, pp. 431–436.

[18] Hazel Scott. "Pegs, cords, and ghuls: Meter of classical Arabic poetry". PhD thesis. Swarthmore College. Department of Linguistics, 2010.

[19] A-F Ahmed, Ramdani Mohamed, and Bellafkih Mostafa. "Machine learning for authorship attribution in Arabic poetry". In: *Int. J. Future Comput. Commun* 6.2 (2017), pp. 42–46.

[20] Abdulrahman Almuhareb et al. "Recognition of classical Arabic poems". In: *Proceedings of the workshop on computational linguistics for literature*. 2013, pp. 9–16.

[21] Hamada Shehdeh Abid Dawood and Ahmad Atawneh. "Assimilation of consonants in English and assimilation of the definite article in Arabic". In: *American Research Journal of English and Literature* 1.4 (2015), pp. 9–15.

[22] Djamel Eddine Kouloughli. *grammaire de l'arabe d'aujourd'hui (Grammar of Today's Arabic)*. French. POCKET, 1994.

[23] AlDiwan. *Arabic Poetry Encyclopedia*. 2019. URL: `https://www.aldiwan.net`.

[24] Adab. *International Encyclopedia of Arabic Poetry*. 2018. URL: `http://www.adab.com`.

[25] Reynold A Nicholson. *A literary history of the Arabs*. Routledge, 2013.

Part II

Structure analysis

Chapter 3

Clustering of metered Arabic poetry

Based on, with extensions:

Clustering analysis of metered
Arabic poetry compositions
20th ACS/IEEE International
Conference on Computer Systems
and Applications (AICCSA)
December, 2023

3.1 Introduction

The poetic structure is a crucial element that warrants in-depth examination. We investigate metered Arabic poetry structure, aiming to identify distinctive traits that contribute to the unique character of each poem.

This study is situated within the exploring of the notions of order and complexity as they pertain to Arabic poetry. Building upon the findings of previous research (i.e., [1, 2]), we understand order to encompass all aspects of structure and organization, whereas complexity is associated with the quantity and variety of information. Consequently, the number of structural elements contributes to its complexity, while the arrangement of these elements influences its order.

In the context of metered Arabic poetry, specific structural elements are particularly relevant when considering complexity. These elements include used meters and patterns throughout the poem while the arrangement of these structural elements contributes to the overall sense of order. By thoroughly analyzing these components, we aim to extract and evaluate poem metrics related to order and complexity for understanding their impact on poems clustering.

Understanding the intricate interplay between order and complexity in Arabic poetry is important both for research and education purposes. It allows to understand the poetic structures and to explore the potential of their comparison. The structure itself does not necessitate a complete understanding of the language, which paves the path for poetic insights into other languages while considering their unique characteristics and demands concerning meter and patterns. This article aims at contributing to this goal by addressing the following research question:

RQ: What are the relevant data metrics that allow for the comparison of different poetic compositions?

To address this question, this article is structured as follows: Section 3.2 presents previous work related to this research, Section 3.3 describes poems structure metrics, Section 3.4 highlights data selection and used poetry corpus and Section 3.5 details conducted clustering experiments. Then, Sections 3.6 and 3.7 summarize clustering options and insights and finally, Section 3.8 wraps up with a conclusion.

3.2 Related work

3.2.1 Classification of non Arabic poetry

Kao and Jurafsky [3] examined what sets skilled poets apart from amateurs by analyzing a corpus of 200 English poems (half by professionals and half by amateurs) across 16 features in style, sentiment, and imagery. Their findings revealed that professional poets tend to use a wider range of distinct words, with a higher type-token ratio being a key indicator of expertise. They also found that professionals rely less on rhyme and alliteration, suggesting a shift away from sound devices in modern poetry. Most notably, the use of concrete imagery was identified as the strongest predictor of high-quality poetry.

Rakshit et al. [4] used orthographic, syntactic, phonemic and lexical features related to word types to classify Bangla poems into four categories: “Devotional”, “Love”, “Nature”, and “Nationalism”. The study compares the performance of Support Vector Machine (SVM) and Naive Bayes algorithms and finds that SVM, trained on both lexical and stylistic features, outperforms the other method. Similarly, Kaur and Saini [5] categorized Punjabi poems into categories: “Nature and Festival”, “Linguistic and Patriotic”, “Relation and Romantic” and “Philosophy and Spiritual” using different lexical, syntactic and semantic features. In the subject-based poetry classification Lou et al. [6] classified poems in categories such “Love”, “Nature”, “Religion”, “Living”, “Arts & Sciences” using term frequency and probabilistic models (Latent Dirichlet Allocation). The SVM classifier gives an overall accuracy ranging between 0.6 and 0.8. In emotion classification field, Ahmad et al. [7] used deep learning algorithms to categorize a dataset of English poems into emotional states such “Love”, “Joy”, “Hope”, “Sadness”, “Anger” etc. Experimental results obtained by using the C-BiLSTM model achieves an overall accuracy of 88%. Tanasescu et al. [8], Gopidi and Alam [9] direct their attention towards binary classification in order to differentiate between metered and non-metered poems or to distinguish poetry from prose.

Šeļa et al. [10] investigated unsupervised machine learning classification techniques when the categories of poems are unknown. For instance, a study utilized a series of unsupervised classifications based on abstracted semantic features of poems to gain insights into the relationships between poetry structure and text meaning. This research provided formal evidence of the association between poetic meter and semantics in 18th and 19th century European literature, using collections from Czech, German, and Russian poetry, with additional data from English poetry and early modern Dutch songs.

Kaplan and Blei [11] analyzed a collection of American poems and computed weighted metrics based on different features: orthographic, syntactic and phonetic. The findings suggest that these features are better suited for clustering and distinguishing between different poetry styles than traditional word-occurrence features. Brooke et al. [12] used a clustering approach for mining T.S. Eliot's poem *"The Waste Land"* using a slightly modified version of K-means algorithm and achieves better results than baseline approaches.

A recent paper also suggests a clustering approach for mining emotional patterns in Romanian poetry using lexicon-based emotion features. Lupea et al. [13] employed a hierarchical clustering algorithm that yields 50 clusters, which are validated by a Silhouette Index of 0.79. Rather than using machine learning techniques, Kernot et al. [14] analyzed the authorship of the 21-poem collection *"The Passionate Pilgrim"*, associated with William Shakespeare using three statistical techniques: Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Vector Space Method (VSM), along with four neurolinguistic-based features. The study reveals that only 15 poems are authored by Shakespeare.

3.2.2 Classification of Arabic poetry

The Arabic poetry categorization by era conducted by Abbas et al. [15] involved using multiple classifiers on a vast collection of poems ranging from the 5th to 15th century. The best-performing model for this task was the Multinomial Naive Bayes with a word tokenizer and no stop words, achieving an accuracy score of 70.21%. For the same purpose of classifying poems by era Orabi et al. [16] employed deep learning models and discovered that polarity categorization between "Modern" and "Non Modern" poems achieved a high score of over 91% due to significant vocabulary differences. To classify Arabic poetry based on emotions, Alsharif et al. [17] categorized poems into four classes: "Elegy", "Love", "Pride", and "Satire". The study compared four algorithms: Naive Bayes, SVM, Voting Feature Intervals (VFI), and Hyperpipes, and found that Hyperpipes with non-stemmed and non-rooted features achieved the highest precision of 79%. Similarly, Ahmed et al. [18] used three classifiers, namely SVM, Linear Support Vector Classification (SVC), and Naive Bayes to categorize modern Arabic poems into four classes: "Love", "Islamic", "Social" and "Political". The study found that SVC and Naive Bayes outperformed the SVM algorithm. In the same topic, Shahriar et al. [19] investigated the use of deep learning techniques for categorizing Arabic poems into emotional categories, such as "Joy", "Sadness", and "Love". After training several models, the study found that AraBERT, an Arabic variant of Bidirectional Encoder Representations from Transformers (BERT), achieved the highest performance. In the field of authorship recognition, Ahmed et al. [20] used lexical, syntactic, and semantic features extracted from poems written by 73 poets and achieve an overall accuracy of 98.63% when using SVM classifier on all features.

This literature review demonstrates that researchers have predominantly employed various techniques to classify poems using textual features, with a focus on non-Arabic languages. However, there has been minimal research exploring Arabic poetry, particularly in the context of its structural analysis. This study seeks to fill that gap by investigating the structure of metered Arabic poetry for clustering purposes, thereby contributing to the advancement of natural language processing. To the best of our knowledge, this article represents the first attempt to cluster metered Arabic poetry using the concepts of order and complexity.

3.3 Poem structure metrics

Syllabic quantity. Denoted by Q, the syllabic quantity is the fundamental metric for calculating the weight of any sequence of syllables. In this metric, a long syllable (L) is assigned twice the weight of a short syllable (S) [21]. The absence of vocalization (A) is considered to have zero weight. The syllabic quantity Q is therefore computed as:

$$Q = 2n_L + 1n_S + 0n_A \tag{3.1}$$

where n_L represents the number of long syllables, n_S the number of short syllables, and n_A the number of syllables without vocalization. For example, the syllabic quantity of the sequence $LSLA$ is $Q = 5$.

Syllabic quantity is utilized by some of the 10 metrics identified as potential candidates for characterizing poetry compositions.

- **Pattern usage (USG)**: highlights the usage of a few patterns in the poetry composition.
- **Pattern potential (POT)**: the difference between the available patterns of the meter and the patterns used in the poem.

- **Repetition (*REP*)**: the maximum repetition rate of syllabic sequences in the poem.
- **Range of syllabic quantity (*RQ*)**: the Manhattan distance that computes the range between the maximum and minimum syllabic quantity per verse.
- **Shannon Evenness (*EVE*)** [22]: quantifies both the variety and the balance of the occurrence of different patterns in a poem.
- **Simpson Evenness (*SEVE*)** [23]: evaluates how evenly the patterns in the poem are distributed across different categories, providing insight into the uniformity of the poem's structure.
- **Variability (*VAR*)**: the standard deviation of syllabic quantity per verse.
- **Pattern variability (*PVAR*)**: the standard deviation of syllabic quantity per pattern.
- **General similarity (*GSIM*)**: compares the syllable sequences of a verse to the previous verse using Gestalt pattern matching [24].
- **Random similarity (*RSIM*)**: compares the syllable sequences of a verse to those of another random verse using Gestalt pattern matching [24].

3.4 Data selection

3.4.1 Corpus

We utilized a subset of the poetry corpus used by Berkani et al. [25] and detailed in Chapter 2, containing poems with at least 5 verses. This subset includes 119 poems, 3,381 verses, 68 distinct poets, and 27 meter variants (see Chapter 2, Subsection 2.5.2).

3.4.2 Poems data

The 10 metrics are computed using all poems from the corpus described above. Consequently, the resulting dataset contains, in addition to the poet's name and the poem title, 10 columns (one for each metric) and 119 rows representing each individual poem.

Figure 3.1 shows the correlation analysis result conducted after standardizing data using the z-score normalization method, which transforms each column to have a mean of 0 and a standard deviation of 1.

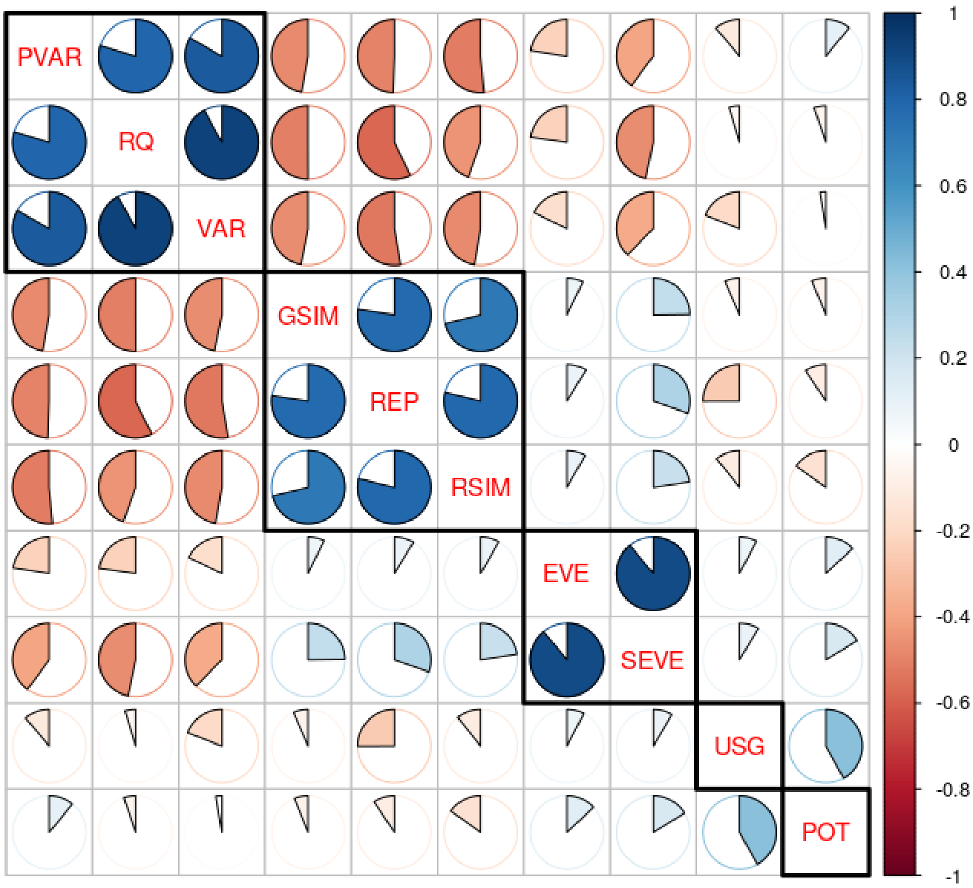


FIGURE 3.1: Attributes correlation matrix

We discarded certain metrics due to their high correlations. For instance, we excluded Simpson's Evenness (*SEVE*) because it is highly correlated with

Shannon's Evenness (*EVE*), making it redundant in this context.

Similarly, we excluded the similarity metrics *GSIM* and *RSIM* due to their high correlations with the repetition metric *REP*, indicating that they capture similar information.

Additionally, we discarded the range metric *RQ* and the pattern variability metric *PVAR*, as both are highly correlated with the variability metric *VAR*, suggesting that these metrics do not provide unique insights beyond what *VAR* already offers.

The 5 selected attributes are shown in Table 3.1.

Table 3.1: Retained poem metrics

Metric	**Att.**	**Type**
Evenness	*EVE*	Complexity
Variability	*VAR*	Order
Repetition	*REP*	Order
Pattern Potential	*POT*	Complexity
Pattern Usage	*USG*	Complexity

3.4.3 Evenness

The Shannon evenness index, also known as Equitability index, is a metric that quantifies the evenness or uniformity of different patterns in the poem. This value ranges from 0 to 1 where 1 indicates complete uniformity.

$$EVE = \frac{H}{\ln(P)} \tag{3.2}$$

$$H = -\sum_{i=1}^{P} (p_i \ln(p_i))$$

EVE is the Shannon Evenness, H is the Shannon entropy, P is the number of distinct patterns in the poem, p_i is the proportion of each pattern i in the poem.

3.4.4 Variability

This metric is computed using the standard deviation of syllabic quantity (Q) of all poem's verses (V). If all verses have the same value (Q), the variability is 0.

$$VAR = \sqrt{\frac{1}{V}\sum_{i=1}^{V}(Q_i - \bar{Q})^2} \tag{3.3}$$

3.4.5 Repetition

This metric evaluates the repetition of syllabic sequences in the poem. Its highest value is 1 when all verses have the same syllabic sequence, as the maximum count of repeated syllables matches the number of verses (V). Conversely, the lowest value is achieved when each verse has its distinct syllabic sequence, resulting in a maximum count of repeated syllables of 1. The repetition metric is then computed as the reciprocal of the number of verses, represented by the division of 1 by the number of verses.

$$REP = \frac{\text{Maximum Count of repeated syllables}}{V} \tag{3.4}$$

3.4.6 Pattern Potential

This metric calculates the difference between the patterns accessible to the poet through the meter (P_a) of the poem and the patterns actually employed in the poem (P_u). It serves as an indicator of the number of unused patterns at the poet's disposal.

$$POT = P_a - P_u \tag{3.5}$$

3.4.7 Pattern Usage

This metric highlights the usage of few patterns in the poem's composition by assuming that composing a limited number of verses with a diverse range of patterns is generally easier than constructing a large number of verses with only a few patterns.

$$USG = \frac{V \cdot POT}{L \cdot P_a} \tag{3.6}$$

USG is the pattern usage metric, V is the number of verses, POT is the pattern potential metric described in equation 3.5, P_a is the number of patterns available in the poem's meter, L is the upper limit of verses processed per poem. In this study, L is set to 100. In most cases, USG varies between 0 and 1 except when the poet uses more than one meter in the same poem.

TABLE 3.2: Data summary before standardizing

	USG	***POT***	***REP***	***EVE***	***VAR***
min	-0.15	-3	0.05	0.19	0.00
median	0.09	3	0.20	0.85	0.86
mean	0.11	3.96	0.29	0.83	0.78
max	0.71	11	1.00	1.00	2.42

The corpus contains some outliers data points that are retained as they could provide insights into unconventional compositions. All data points are centered and scaled.

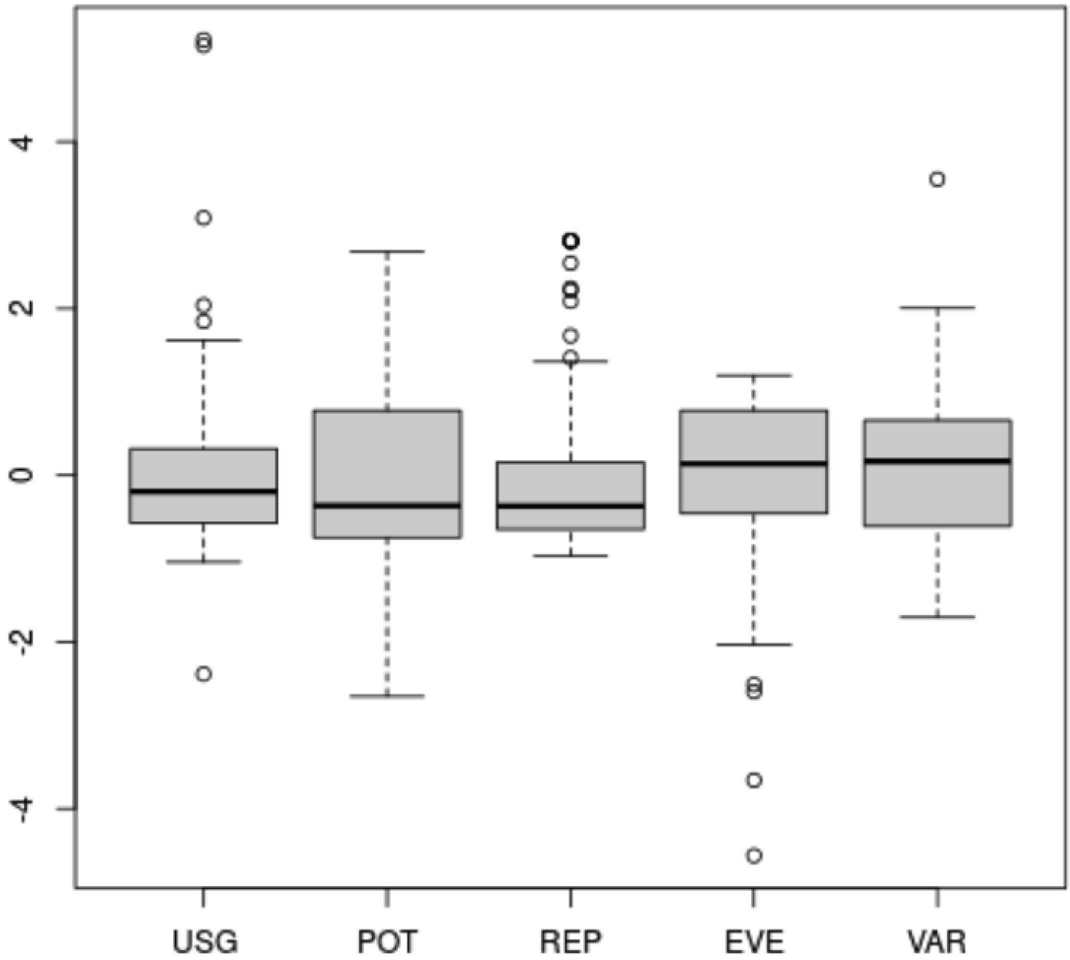


FIGURE 3.2: Scaled data summary

3.5 Clustering analysis

The goal of clustering Arabic poetry compositions is to explore the degree of similarity or dissimilarity among poetic structures, aiming to enhance their characterization and categorization beyond the traditional classification based on meter. Clustering, also known as cluster analysis, consists of grouping objects into a number of clusters in a manner that objects in the same groups are similar to each other and dissimilar to those in other groups [26]. Clustering is a fundamental component of exploratory data analysis. Unlike supervised learning, where models are trained to map input to a specific output, clustering is considered as an unsupervised learning approach [27]. The task of clustering is challenging due to the fact that clustering methods will generate clusters even if the data lacks inherent cluster structure. Consequently, it becomes imperative to assess the cluster tendency of the dataset prior to employing any clustering algorithm. Once the tendency is deemed sufficient, the analysis can proceed by clustering the data, subsequently evaluating the quality of the obtained results. Another hurdle arises in attribute selection, as there is no target output available. This necessitates context dependent approaches [28].

In clustering analysis, five commonly used clustering indicators [29] are computed. The Hopkins statistic is used for clustering tendency. The Silhouette coefficient, the Calinski-Harabasz index, the Davies-Bouldin index and the Dunn index are used for cluster validation. The Hopkins statistic [30] is employed to evaluate the clustering tendency of a dataset, aiming to determine if the data distribution is non-random. The Hopkins statistic value ranges from 0, indicating a lack of discernible clustering possibilities, to 1, denoting a substantial capacity for effective classification. Unlike the Hopkins statistic used for assessing clustering tendency, the Silhouette, Calinski-Harabasz, Davies-Bouldin and Dunn indexes are used for clustering validation. The Silhouette coefficient [31] serves as a metric for evaluating the degree of separation between clusters. It quantifies the proximity of each data point within a cluster to the data points in neighboring clusters. Ranging between -1 and 1, a value approaching 1 signifies that a data point is significantly distant from neighboring clusters. A value of 0 indicates that the data point is in close proximity to neighboring clusters. Conversely, a negative value suggests that the data point may have been wrongly assigned to an incorrect cluster. According to Dalmaijer et al. [32] and Kaufman and Rousseeuw [33], when the Silhouette coefficient exceeds the threshold of 0.50, it indicates a good clustering quality. A value surpassing 0.70 provides strong evidence of excellent clustering quality. The Calinski-Harabasz index [34], also known as the variance ratio criterion, is a clustering evaluation metric for assessing different clustering alternatives. It is a ratio of the between-cluster variance and the within-cluster variance, and it is used to determine the number of clusters that should be used in a clustering solution. The Calinski-Harabasz index ranges from 0 to infinity and should be maximized. The higher the index, the better the solution. The Davies-Bouldin index [35] is a measure used to evaluate the clustering quality. It is based on the average similarity between clusters considering within-cluster dispersion and separation between clusters. It theoretically ranges from 0 to infinity and should be minimized. The lower the index, the better the solution. The Dunn index [36] is another clustering evaluation metric defined as the ratio between the minimum inter-cluster distance and the maximum intra-cluster distance. It theoretically ranges from 0 to infinity and should be maximized. The higher the index, the better the solution.

3.5.1 Clustering approach

In the context of this research and due to the lack of available output attribute, the clustering of all combinations of attributes is considered. Attributes are combined without repetition. Each combination includes at least 2 attributes and up to 5. Two distinct clustering algorithms, namely *hclust* (Agglomerative Hierarchical Clustering) and *K-means* (Partitioning clustering) are applied to each combination. For each algorithm, 8 clustering options are tested by varying the number of clusters from 2 to 9. Consequently, 416 clustering options are tested, as summarized in Table 3.3.

Table 3.3: Attributes combinations and experiments

Atts.	**Comb.**	**Alg.**	**Clusters**	**Exp.**
5	1	2	8	16
4	5	2	8	80
3	10	2	8	160
2	10	2	8	160
	26			**416**

Algorithm 2 Clustering experiments

```
for att = 2 to 5 do
  attributes ← Combinations(att)
  for all comb in attributes do
    h ← Hopkins(comb)
    for all alg in (hclust, kmeans) do
      for k = 2 to 9 do
        c ← Clustering(comb, alg, k)
        sil ← Silhouette(c)
        ch ← CalinskiHarabasz(c)
        db ← DaviesBouldin(c)
        dn ← Dunn(c)
        Result(att, comb, alg, k, h, sil, ch, db, dn)
      end for
    end for
  end for
end for
```

The output of algorithm 2 is a dataset containing 416 clustering results including attributes combination, algorithm, cluster number (k), and all the five clustering metrics: Hopkins statistic (h), Silhouette coefficient (sil), Calinski-Harabasz index (ch), Davies-Bouldin index (db) and Dunn index (dn).

3.5.2 Clustering tendency

We used the Hopkins statistic to assess clustering tendency. Minimum and maximum values for each attribute combinations are shown in Table 3.4.

TABLE 3.4: Clustering tendency assessment using Hopkins Statistic *(h)*

Atts.	**Min(h)**	**Max(h)**
5	0.77	0.77
4	0.75	0.82
3	0.75	0.86
2	0.69	0.88

The Hopkins statistic values indicate that the data distribution is non-random, suggesting discernible clustering possibilities. These values also imply that reducing the number of attributes can still yield strong clustering performance if the correct combinations are chosen. In particular, some 3-attribute and 2-attribute combinations appear to provide a stronger clustering tendency. Thus, using fewer attributes can still result in high clustering effectiveness, if the optimal combinations are selected.

3.5.3 Clustering options

During this phase, the objective is to identify the most promising clustering options. To achieve this, we filtered the clustering results in two steps.

We first established a comparison baseline by setting a threshold for the Silhouette score (*sil*) at 0.50, filtering the dataset to retain only rows where *sil* $>=$ *0.50.* For the remaining entries, we calculated the average values of the Calinski-Harabasz index (*ch*), Davies-Bouldin index (*db*), and Dunn index (*dn*), as shown in Table 3.5. Since, aside from the Silhouette score, there are no established threshold values for these indices in the literature, due to their theoretical tendency towards infinity through maximization or minimization, we decided to compute the averages of these indices within our context and use these average values as thresholds.

TABLE 3.5: Clustering baseline where Silhouette >= 0.50

dn	ch	db
0.23	52.73	1.07
0.15	37.86	1.24
0.38	37.21	0.25
0.18	48.72	0.38
0.06	68.78	0.62
0.05	111.25	0.65
0.08	116.57	0.70
0.07	147.91	0.63
0.04	73.36	1.16
0.03	76.58	0.86
0.10	45.25	0.99
0.08	62.53	0.65
0.09	146.02	0.74
0.03	116.94	1.00
0.10	121.78	0.96
0.11	84.23	0.79
0.11	**84.23**	**0.79**

In the second step, in order to ensure a more comprehensive evaluation and selection of the optimal clustering solution, additional conditions are introduced to avoid relying solely on a single validation index, such as the Silhouette coefficient. It is verified whether at least two other indexes exhibit superior performance compared to the baseline. Specifically, the Calinski-Harabasz and Dunn indexes are expected to exceed the baseline value for maximization, while the Davies-Bouldin index is anticipated to be minimized and thus be lower than the baseline.

TABLE 3.6: Clustering options

alg	att	k	h	sil	ch	db	dn
kmeans	POT-REP	3	0.70	0.53	146.02	0.74	0.09
kmeans	USG-REP	3	0.88	0.58	116.57	0.70	0.08
hclust	REP-EVE	2	0.70	0.59	48.72	0.38	0.18
hclust	REP-VAR	2	0.83	0.58	111.25	0.65	0.05
hclust	REP-VAR	3	0.83	0.56	147.91	0.63	0.07
hclust	USG-EVE	2	0.84	0.73	37.21	0.25	0.38

3.6 Results

The final outcome presented in Table 3.6 retains 6 clustering options.

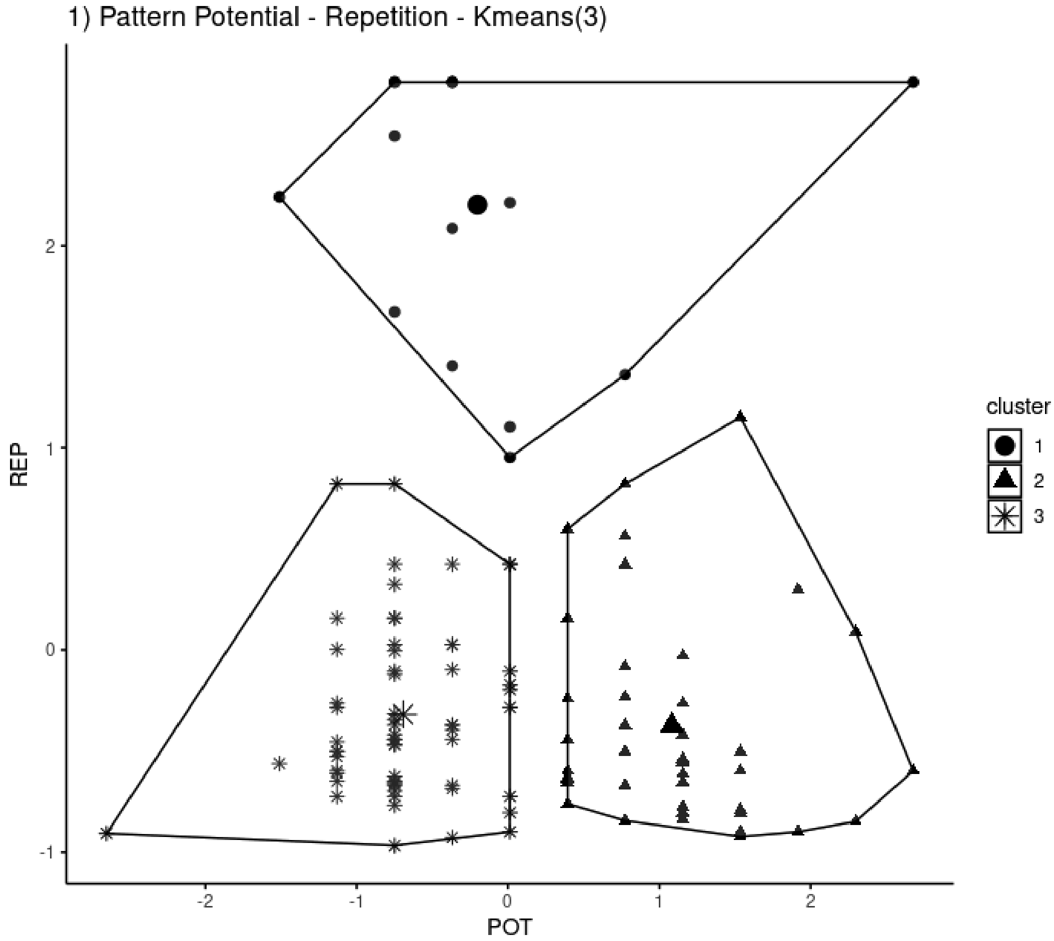


FIGURE 3.3: Clustering option 1

Clustering option 1, shown in Figure 3.3, involves the two attributes *POT* and *REP*, with a division into 3 clusters: C1, C2, and C3, with respective sizes of 16, 42, and 61. C3 primarily contains poetic compositions with low to moderate *POT* and low to moderate *REP*, suggesting that poets in this group used more than half of the patterns available in the meter they chose. Conversely, C2 contains poems that used less than half of the available patterns while also maintaining low *REP*. C1 groups compositions with high *REP* and moderate *POT*, suggesting the use of about half of the available patterns. The extreme case of high *REP* and high *POT* involves a composition that used only a single pattern.

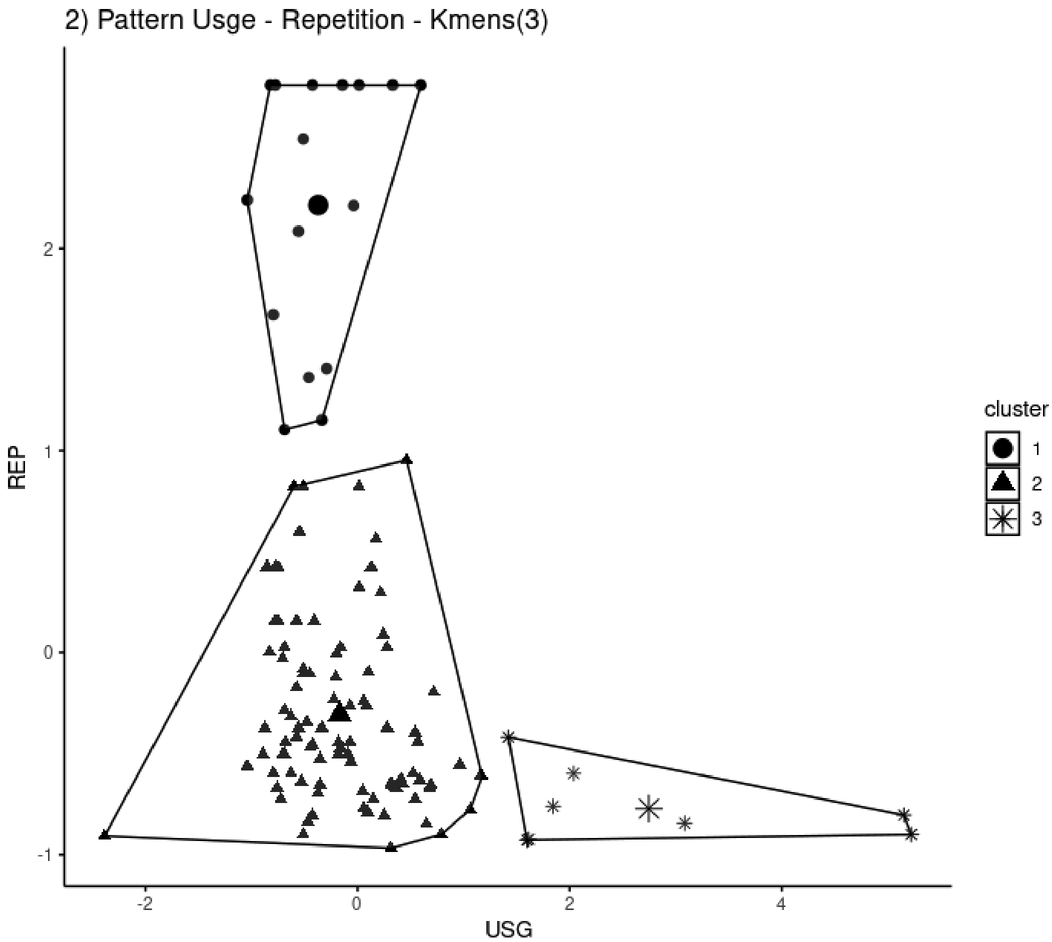


Figure 3.4: Clustering option 2

Option 2, shown in Figure 3.4, focuses on the attributes *USG* and *REP*, with a division into 3 clusters: C1, C2, and C3, with respective sizes of 16, 95, and 8. In this option, poems with medium to high *USG* maintain low repetition. This is the case for C3, which represents complex compositions, as poems in this cluster use few patterns to write many verses while maintaining a low level of repetition. C1 and C2 include poems with a low to moderate number of verses, which partly explains their low *USG*. C2, which contains the majority of compositions, maintains low to moderate *REP*, unlike C1, which has a similar *USG* but with higher repetition.

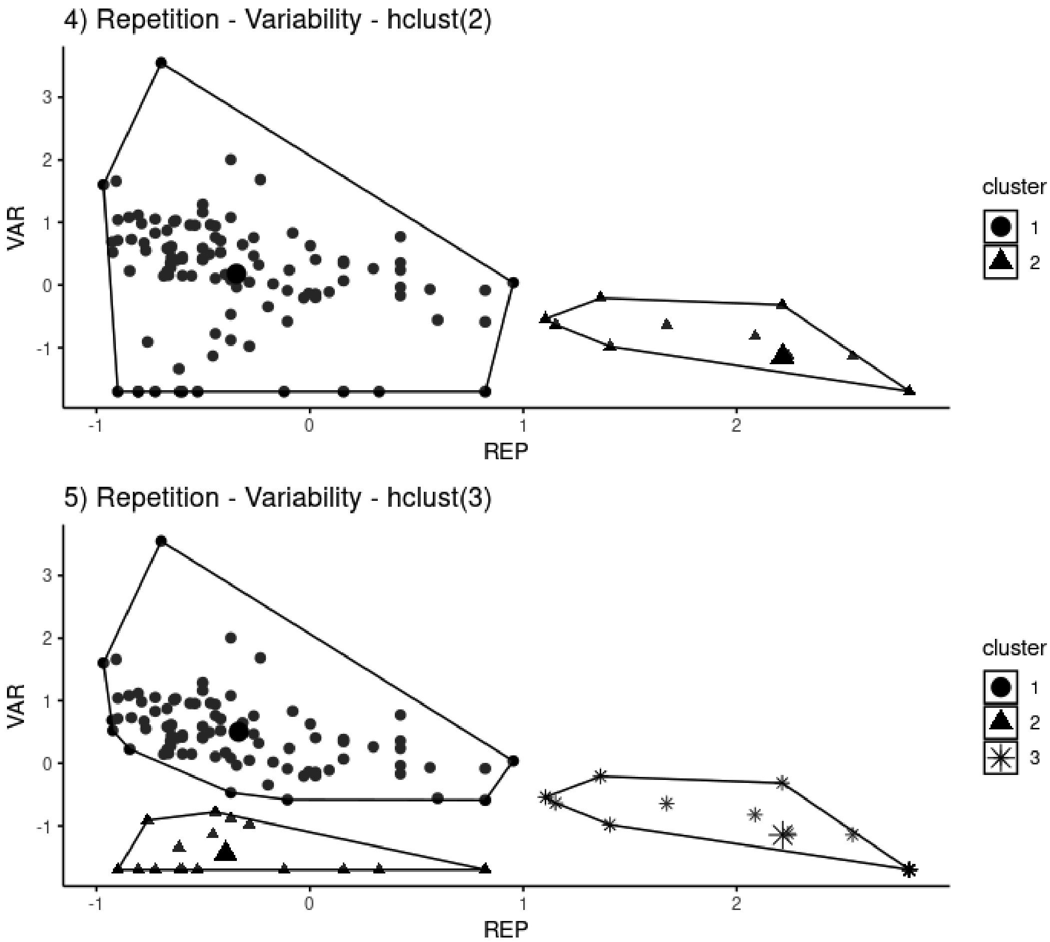


FIGURE 3.5: Clustering options 4 and 5

Options 4 and 5, shown in Figure 3.5, involve the attributes *VAR* and *REP*. In option 4, the data is divided into 2 clusters, C1 and C2, with respective sizes of 103 and 16. C2 is identical to C3 in option 5 and to C1 in option 2, grouping compositions with high *REP* and low *VAR*, suggesting compositions that repeat the same patterns throughout the poem. In contrast, cluster C2 in option 5 is characterized by low *REP* and low *VAR*, indicating a certain complexity, as it involves arranging patterns with similar syllabic quantities while ensuring that the same arrangements are not repeated from verse to verse. C1 in option 4 contains the same poems as the two clusters C1 and C2 in option 5, with respective distributions of 86 and 17. C1 in option 5 suggests that the majority of compositions have low to moderate *VAR* and low to moderate *REP*.

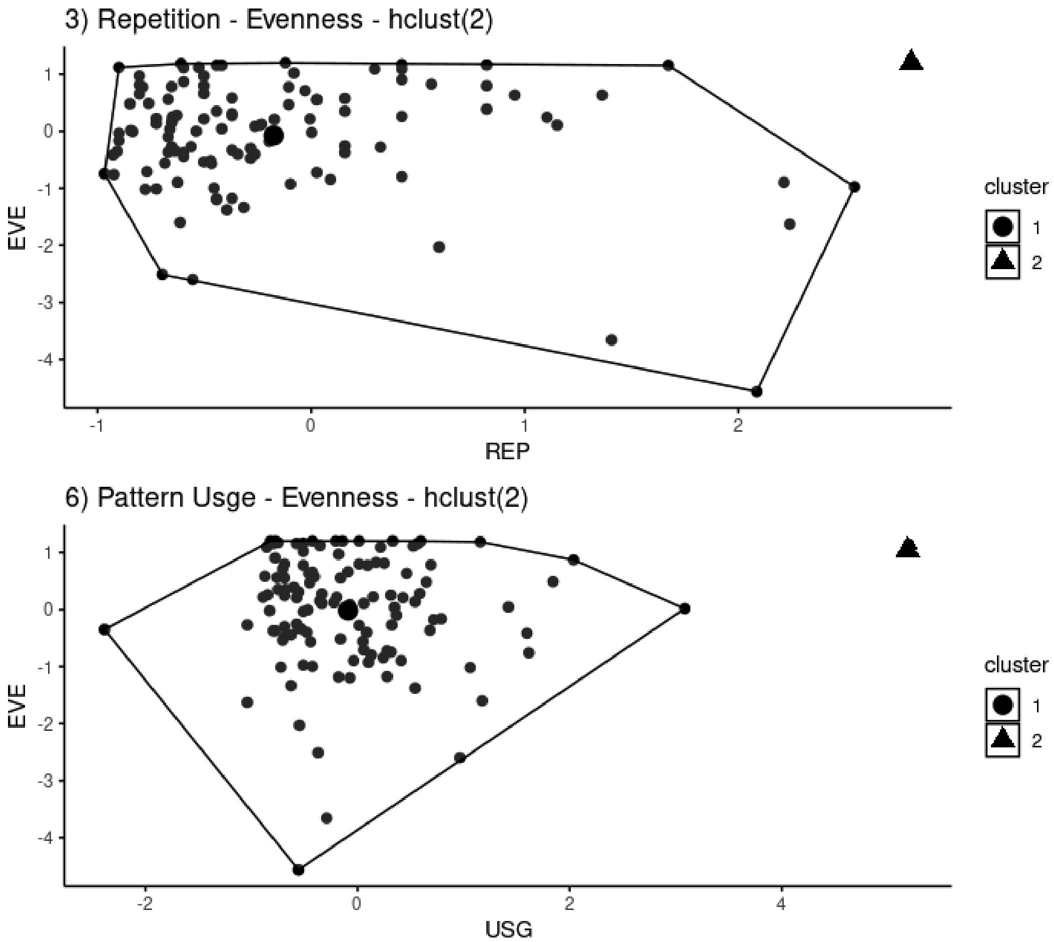


FIGURE 3.6: Clustering options 3 and 6

Options 3 and 6, shown in Figure 3.6, involve the attribute *EVE* with *REP* and *USG*, respectively. The clustering does not seem to provide meaningful insights into the context of metered Arabic poetry. In Option 3, the data is divided into two clusters, C1 and C2, where C1 is dominant with 112 poems compared to 7 in C2, which contains poems that achieved perfect uniformity by using a single pattern or multiple patterns with the same syllabic quantity. Both cluster C1 in option 3 and cluster C1 in option 6 are not very refined, suggesting that approximately two-thirds of the compositions are uniform. In Option 6, the distribution is unbalanced with two clusters, C1 and C2, with respective sizes of 117 and 2. This latter minority cluster, however, includes two compositions distinguished by high *USG*, reflecting long compositions with few patterns.

3.7 Discussion

As indicated by the results in Table 3.6, all of the selected options suggest that only two attributes are necessary for clustering the poetry data, as opposed to the initial five attributes. Among the five options considered, the metric *REP* appears five times, while *USG*, *VAR*, and *EVE* appear twice, and *POT* appears once.

In clusters with the largest number of members, the majority of poems exhibit characteristics such as low to moderate pattern usage *USG*, low to moderate repetition *REP*, and moderate variability *VAR*. However, poets who aim to distinguish their poems from common compositions tend to belong to smaller groups. These groups exhibit either a high repetition rate *REP* (options 1, 2, 3, 4, 5), high pattern usage *USG* (options 2, 6), or low variability *VAR* (option 5). For example, the poets *'Antarah ibn Shaddād* and *Labīd ibn Rabī'a* demonstrated high pattern usage in their poems *al-Mu'allaqāt* [37], by composing 89 and 88 verses, respectively, using only 2 patterns out of the available 10, while maintaining a variability rate of 0 through consistent syllabic quantities across all verses.

Given their clear role in highlighting distinguished compositions, the metrics *REP*, *VAR*, and *USG* stand out as both easier to interpret and essential for clustering metered Arabic poetry. Consequently, these attributes are retained as key factors in the clustering process. The nature of a composition then depends on the poet's skill in effectively combining and balancing these three metrics, which can sometimes be at odds with one another.

The direction of a composition is also influenced by the poet's intended emphasis. The poem's semantics, including its message, emotional tone, and imagery, play a significant role in meter selection. When poets prioritize low variability, they tend to compose verses with consistent syllabic quantities, ensuring a steady rhythmic flow throughout the poem. Furthermore, by selecting and reusing a set of patterns across multiple verses, they can achieve a high pattern usage rate. However, if poets seek to avoid repetitive patterns, composing with a limited number of patterns becomes challenging. Varying their placement poses the risk of diverging from the selected meter and potentially altering syllabic quantities, which could compromise the objective of low variability.

It is therefore advisable to consider these attributes in pairs rather than all three simultaneously, as demonstrated by *USG* with *REP* or *VAR* with *REP* in options 2 and 5. These two options suggest clustering Arabic poetry into three clusters. We anticipate that expanding the corpus could reveal additional clusters. However, metered poetry, unlike free verse or prose, is typically less variable. Consequently, by significantly broadening the study scope, we might expect to see the emergence of a range of clusters, potentially between 6 and a maximum of 9, corresponding to the possible combinations of the three attributes: *VAR* (Low, Medium), *REP* (Low, Medium, High), and *USG* (Low, Medium, High).

3.8 Conclusion

The objective of this research was to identify the structural factors that differentiate one metered Arabic poem from another. We conducted a clustering analysis on 119 metered Arabic poems, considering five independent variables: *Evenness*, *Variability*, *Repetition*, *Pattern Potential*, and *Pattern Usage*. These variables were developed as metrics of order and complexity, with the assumption that the number of structural elements contributes to composition complexity, while the arrangement of these elements influences its order.

By exploring all possible combinations of attributes, we applied two clustering algorithms: *K-means* and *hclust*, and varied the number of clusters from 2 to 9. This approach involved testing 416 attribute combinations and progressively measuring and validating the clustering results. Evaluation metrics, including the Hopkins statistic, Silhouette coefficient, Calinski-Harabasz index, Davies-Bouldin index, and Dunn index, were used to refine the selection process. This iterative approach enabled us to narrow down the clustering options from 416 to 6 by establishing clustering quality thresholds. The results suggest that two attributes are sufficient to characterize the poetry, highlighting the potential effectiveness of a reduced set of attributes for clustering poetry data.

The metrics *Repetition*, *Variability*, and *Pattern Usage* emerged as the most significant factors for clustering metered Arabic poetry. Their capacity to highlight distinctive compositions and ease of interpretation underscores their importance in characterizing metered Arabic poetry compositions.

The maximum number of clusters suggested by the results is 3. However, we hypothesize that significantly expanding the corpus could reveal a clustering range of 6 to 9, particularly when these attributes are considered in pairs rather than simultaneously. This potential range corresponds to combinations of the three attributes: *Variability* (Low, Medium), *Repetition* (Low, Medium, High), and *Pattern Usage* (Low, Medium, High).

References

[1] Eline Van Geert and Johan Wagemans. "Order, complexity, and aesthetic appreciation." In: *Psychology of aesthetics, creativity, and the arts* 14.2 (2020), p. 135.

[2] Eline Van Geert, Christophe Bossens, and Johan Wagemans. "The Order & Complexity Toolbox for Aesthetics (OCTA): A systematic approach to

study the relations between order, complexity, and aesthetic appreciation". In: *Behavior Research Methods* (2022), pp. 1–24.

[3] Justine Kao and Dan Jurafsky. "A computational analysis of style, affect, and imagery in contemporary poetry". In: *Proceedings of the NAACL-HLT 2012 workshop on computational linguistics for literature*. 2012, pp. 8–17.

[4] Geetanjali Rakshit et al. "Automated analysis of bangla poetry for classification and poet identification". In: *Proceedings of the 12th international conference on natural language processing*. 2015, pp. 247–253.

[5] Jasleen Kaur and Jatinderkumar R Saini. "Designing punjabi poetry classifiers using machine learning and different textual features." In: *Int. Arab J. Inf. Technol.* 17.1 (2020), pp. 38–44.

[6] Andres Lou, Diana Inkpen, and Chris Tanasescu. "Multilabel subject-based classification of poetry". In: *Nature* 2218 (2015), pp. 30–7.

[7] Shakeel Ahmad et al. "Classification of poetry text into the emotional states using deep learning technique". In: *IEEE Access* 8 (2020), pp. 73865–73878.

[8] Chris Tanasescu, Bryan Paget, and Diana Inkpen. "Automatic classification of poetry by meter and rhyme". In: *The Twenty-Ninth International Flairs Conference*. 2016.

[9] Amitha Gopidi and Aniket Alam. "Computational analysis of the historical changes in poetry and prose". In: *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*. 2019, pp. 14–22.

[10] Artjoms Šeļa, Petr Plecháč, and Alie Lassche. "Semantics of European poetry is shaped by conservative forces: The relationship between poetic

meter and meaning in accentual-syllabic verse". In: *Plos one* 17.4 (2022), e0266556.

[11] David M Kaplan and David M Blei. "A computational approach to style in American poetry". In: *Seventh IEEE International Conference on Data Mining (ICDM 2007)*. IEEE. 2007, pp. 553–558.

[12] Julian Brooke, Graeme Hirst, and Adam Hammond. "Clustering voices in the Waste Land". In: *Proceedings of the Workshop on Computational Linguistics for Literature*. 2013, pp. 41–46.

[13] Mihaiela Lupea, Anamaria Briciu, and Elena Bostenaru. "Emotion-based hierarchical clustering of romanian poetry". In: *Stud. Inform. Control* 30 (2021), pp. 109–118.

[14] David Kernot, Terry Bossomaier, and Roger Bradbury. "Stylometric Techniques for Multiple Author Clustering". In: *International Journal of Advanced Computer Science and Applications* 8.3 (2017).

[15] Mourad Abbas, Mohamed Lichouri, and Ahmed Zeggada. "Classification of Arabic Poems: from the 5th to the 15th Century". In: *International Conference on Image Analysis and Processing*. Springer. 2019, pp. 179–186.

[16] Mariam Orabi, Hozayfa El Rifai, and Ashraf Elnagar. "Classical Arabic Poetry: Classification based on Era". In: *2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA)*. IEEE. 2020, pp. 1–6.

[17] Ouais Alsharif, Deema Alshamaa, and Nada Ghneim. "Emotion classification in Arabic poetry using machine learning". In: *International Journal of Computer Applications* 65.16 (2013).

[18] Munef Abdullah Ahmed et al. “The classification of the modern arabic poetry using machine learning”. In: *TELKOMNIKA (Telecommunication Computing Electronics and Control)* 17.5 (2019), pp. 2667–2674.

[19] Sakib Shahriar, Noora Al Roken, and Imran Zualkernan. “Classification of Arabic Poetry Emotions Using Deep Learning”. In: *Computers* 12.5 (2023), p. 89.

[20] A-F Ahmed, Ramdani Mohamed, and Bellafkih Mostafa. “Machine learning for authorship attribution in Arabic poetry”. In: *Int. J. Future Comput. Commun* 6.2 (2017), pp. 42–46.

[21] Ibrahim Anis. *Linguistic sounds*. Arabic. Cairo: Maktabat Al Anglo Masriya, 1975.

[22] Claude E Shannon. “A mathematical theory of communication”. In: *The Bell system technical journal* 27.3 (1948), pp. 379–423.

[23] Edward H Simpson. “Measurement of diversity”. In: *nature* 163.4148 (1949), pp. 688–688.

[24] PE Black. *Ratcliff/Obershelp pattern recognition*. `http://www.nist.gov/dads/HTML/ratcliffObershelp.html/`. Accessed: 2022-04-06. 2021.

[25] Abdelmalek Berkani, Adrian Holzer, and Kilian Stoffel. “Pattern matching in meter detection of Arabic classical poetry”. In: *2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA)*. IEEE. 2020, pp. 1–8.

[26] Michael Steinbach, Levent Ertöz, and Vipin Kumar. “The challenges of clustering high dimensional data”. In: *New directions in statistical physics: econophysics, bioinformatics, and pattern recognition* (2004), pp. 273–309.

[27] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. “Data clustering: a review”. In: *ACM computing surveys (CSUR)* 31.3 (1999), pp. 264–323.

[28] Rui Xu and Donald Wunsch. "Survey of clustering algorithms". In: *IEEE Transactions on neural networks* 16.3 (2005), pp. 645–678.

[29] Olatz Arbelaitz et al. "An extensive comparative study of cluster validity indices". In: *Pattern recognition* 46.1 (2013), pp. 243–256.

[30] Brian Hopkins and John Gordon Skellam. "A new method for determining the type of distribution of plant individuals". In: *Annals of Botany* 18.2 (1954), pp. 213–227.

[31] Peter J Rousseeuw. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". In: *Journal of computational and applied mathematics* 20 (1987), pp. 53–65.

[32] Edwin S Dalmaijer, Camilla L Nord, and Duncan E Astle. "Statistical power for cluster analysis". In: *BMC bioinformatics* 23.1 (2022), pp. 1–28.

[33] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

[34] Tadeusz Caliński and Jerzy Harabasz. "A dendrite method for cluster analysis". In: *Communications in Statistics-theory and Methods* 3.1 (1974), pp. 1–27.

[35] David L Davies and Donald W Bouldin. "A cluster separation measure". In: *IEEE transactions on pattern analysis and machine intelligence* 1.2 (1979), pp. 224–227.

[36] J. C. Dunn. "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters". In: *Journal of Cybernetics* 3.3 (1973), pp. 32–57. DOI: 10.1080/01969727308546046.

[37] Al-Zawzani. *Explanation of the Ten Mu'allaqat*. Arabic. Beirut: Dar Maktabat Al Hayat, 1983.

Part III

Visualization

Chapter 4

Visualization of poetic compositions

Based on, with extensions:

Enhancing Arabic Poetic Structure Analysis through Visualization

IEEE VIS 2024: Workshop of natural language processing visualization

October, 2024

4.1 Introduction

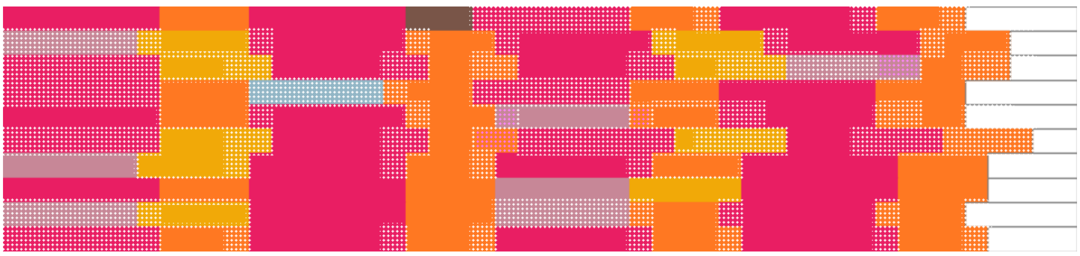


FIGURE 4.1: Example of Arabic poetry visualization

Poetic compositions are fundamentally rooted in two aspects: content and structure. While translation can bridge language gaps and facilitate content comprehension, it falls short in fully conveying the original semantics, message, affect, and imagery as intended by the poet [1]. However, the structure of literary works is universally comprehensible, making visualization an invaluable tool for exploration and analysis. Information visualization provides a powerful means to explore and understand literary works. By translating textual data into a visual format, readers can get insights [2] that may not be immediately discernible from a cursory reading of the text. Visualization aids in identifying structure, patterns, and trends [3]. Moreover, it enables the comparison and contrast of different literary works, highlighting their similarities and differences.

In the context of Arabic poetry, structure visualization offers two benefits. Firstly, it provides a visual medium that enables both Arabic and non-Arabic readers to understand the components of poetry and comprehend the poet's arrangement. Secondly, it facilitates the analysis and comparison of various compositions, aiding in the extraction of potential insights. This necessitates an examination of the relationship between the presentation of poems and the resulting insights, and the impact of appearance on insight outcomes. This research supplements our previous effort [4] to enhance clustering results. Until now, only textual visualizations of the data have been employed. For the current article, we have chosen graphical visualization to extract insights that were previously unnoticed.

This article seeks to contribute to this objective by addressing the following research question:

RQ: What are the relevant visualizations that allow for the comparison of different poetic compositions?

To address this question, this chapter is structured as follows: Section 4.2 presents previous work related to this research, Section 4.3 highlights preliminary explorations of poetry visualizations, Section 4.4 outlines the research method and approach, Section 4.5 introduces patterns and dissimilarity visualizations, Section 4.6 details the clustering experiment, Section 4.7 discusses the results and insights, and finally, Section 4.8 provides a summary.

4.2 Related work

Various methods for visualizing poetry structures have been explored in previous work, serving educational, analytical, and artistic purposes [5]. One such tool, "Emily" presented by Madnani [6], represents poems through colored lines and bars, focusing on the 19th-century verse-letters and poetry of Emily Dickinson.

A suite of visualization tools proposed by Meneses and Furuta [7] assists scholars in conducting critical poetry analysis, synthesizing and highlighting key poem elements. These tools employ graphic shapes like squares, circles, and lines in different colors to create wave-like patterns.

Another example is the "Colors of poetry", a system presented by Robinson [8], which analyzes Spanish poetry and uses linguistic features of text such as syllabification, intonation, rhyme, meter and pause. The tool displays verses annotated and highlighted with several attributes, most notably the stress and phonemes.

The "Poem Viewer" designed by Abdul-Rahman et al. [9] is a user-centered design interface for poetry visualization, allowing for the display of poetry text and its associated semantic and phonetic elements. Visualization is achieved using variously colored circles, rectangles, and lines.

McCurdy et al. [10] proposed "Poemage", an interactive tool for exploring a poem's sonic topology. It investigates the role and influence of technology on the close reading of a poem, representing elements like rhyme types, alliterations, and anagrams in various shapes and colors.

Delmonte [11] presented “SPARSAR”, an application for poetry analysis and visualization, using variable shapes and colors to visualize 75 of Shakespeare's sonnets. It provides three views: “Phonetic Relational”, “Poetic Relational”, and “Semantic Relational”, each based on different features of the poem.

Mittmann et al. [12] proposed a paper on multi-way poetry visualizations as a scheme to visualize poetry at four levels, with a case study based on Portuguese poetry.

Musaoglu et al. [13] proposed “Poetry Barcode”, an application that visualizes a poem's structure, including verb usage, emotional tone, alliterations, adjectives, and word length. The case study focused on the collection of Turkish poet *Nâzım Hikmet Ran*.

Risha et al. [14] presented “Soliloquy”, an interface that visualizes expert thoughts during poem reading and interpretation, aiming to provide insight into its design and potential learning impact.

Other researchers focus on links between poetry and art. Gao et al. [15] proposed a framework for turning poetry into painting aims to study the communication between the artist and the audience. They identify three levels namely technical, semantic and effectiveness. Lerdahl [16] explored the link between poetry and music and proposed an approach to visualize poetry sounds as musical notes.

In our initial attempt at poetry clustering [4], we explored the unique characteristics of metered Arabic poetry to enable comparative analysis of poetry structures. The study applied a clustering method using two different algorithms and six metrics calculated on the poems in the corpus: *Evenness*, *Variability*, *Repetition*, *Pattern Potential*, and *Pattern Usage*. The findings reveal that *Repetition*, *Variability*, and *Pattern Usage* are the most significant metrics, and that only two attributes are necessary to characterize the poetry, rather than all six.

To the best of our knowledge, this article represents one of the first research efforts in clustering metered Arabic poetry compositions, offering a visual and analytical exploration of its intricate structures.

4.3 Exploratory of poetry visualization

Before proceeding with the analysis of metrics through visualization, we conducted several exploratory visualization experiments to evaluate the potential for visualizing metered Arabic poetry. This exploration focuses specifically on the structural aspects of the poetry, leading us to employ simple shapes. This approach allowed us to address the challenges of visualizing poems with different meters and varying verse counts.

In this exploratory phase, we investigated two primary dimensions: the first involved visualizing an entire poem as a summary, while the second focused on the detailed arrangement of syllables and patterns within a subset of the poem. For both dimensions, we calculated two regularity metrics related to verse structure and pattern consistency, respectively.

The measures used for poem visualization are based on the metric *Syllabic Quantity (Q)*, as explained in Equation 3.1 in Chapter 3.

4.3.1 Verse regularity

For verse regularity (*Reg*), we evaluate, for each verse, whether the current syllabic quantity matches that of the preceding verse. For each comparison, we assign a value d, with 1 indicating equality and 0 indicating a difference. The *Reg* value is then calculated by summing all d values across the verses and dividing by the total number of verses, r, excluding the first verse, as it has no preceding verse for comparison.

$$Reg = \frac{\sum_{i=2}^{r} d}{r-1} \tag{4.1}$$

TABLE 4.1: Example of verse regularity calculation

ID	ID.pr	Verse	Verse.pr	Q	Q.pr	d
1	1	LLSL	LLSL	7	7	1
2	1	LSSL	LLSL	6	7	0
3	2	SSLL	LSSL	6	6	1
4	3	SLSL	SSLL	6	6	1
5	4	SLLL	SLSL	7	6	0
					Reg:	**0.50**

Note: "pr." stands for "preceding."

4.3.2 Pattern regularity

For pattern regularity, we used the same regularity formula (Equation 4.1), but instead of comparing verses, we compared patterns occupying the same position in consecutive verses. For example, the pattern in position 1 of verse 1 is compared to the pattern in position 1 of verse 2, and so on. We calculate the regularity for each pattern and take the average as a global pattern regularity measure, as shown in Table 4.2.

TABLE 4.2: Example of pattern regularity calculation

P1	P1.pr	P2	P2.pr	Q.P1	Q.P1.pr	Q.P2	Q.P2.pr	d.P1	d.P2
LSS	LSS	SSL	SSL	4	4	4	4	1	1
LLS	LSS	SSL	SSL	5	4	4	4	0	1
LLS	LLS	SLL	SSL	5	5	5	4	1	0
LSL	LLS	SLL	SLL	5	5	5	5	1	1
LSL	LSL	LSS	SLL	5	5	4	5	1	0
							Reg:	**0.75**	**0.50**

Note: "pr." stands for "preceding."

4.3.3 Visualizing a poem summary

The poem summary visualization includes up to 100 verses per poem and is based on the maximum and minimum syllabic quantities across all verses and pattern positions within the poem, as shown in Table 4.3. Data is scaled to fit the width and height of the visualization shape.

TABLE 4.3: Example of data used for summary visualization

Verse	Total	P1	P2	P3	P4	P5	P6
1	37	7	6	5	7	7	5
2	35	6	7	5	6	6	5
3	36	6	7	5	6	7	5
4	35	6	7	5	6	6	5
5	35	6	6	5	6	7	5
Max (Q)	37	7	7	5	7	7	5
Min (Q)	35	6	6	5	6	6	5

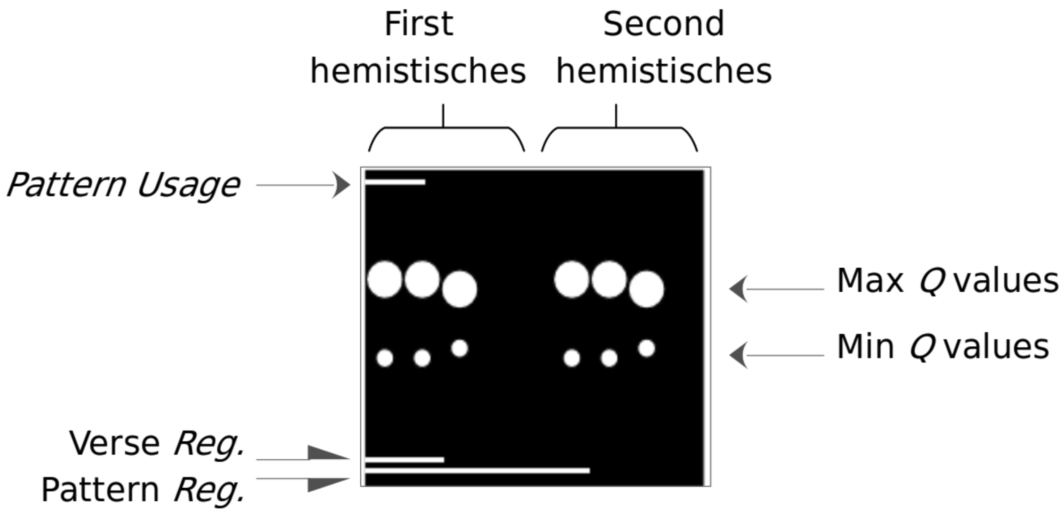


FIGURE 4.2: Summary visualization of poetry composition

Figure 4.2 presents a set of global indicators summarizing poem composition. Verse regularity illustrates how consistently the poet maintains structure across verses. Pattern regularity provides insight into the distribution of patterns within each verse, highlighting the poet's ability to sustain coherence. Pattern usage offers an idea of the composition's complexity, based on the assumption that composing a few verses with numerous patterns is generally simpler than composing many verses with fewer patterns, while still maintaining high poetic quality and semantic coherence. *Pattern Usage* metric is explained in Equation 3.6 in Chapter 3.

The number of circles represents the number of meter positions in the poem. Large and small circles indicate the symmetry of regularity between the two hemistiches, particularly in terms of pattern positioning. The greater the distance between the large and small circles, the weaker the poem's regularity. Perfect regularity is achieved when the large and small circles overlap.

4.3.4 Visualizing a poem details

Since poems may vary significantly in number of verses and in order to have some comparable view between poems, the choice is made to display only the first ten verses of the poem.

Details visualization of a poem relies on its patterns arrangement. Since pattern is a sequence of syllables, the later form the building block of detailed visualization. The shape's form and dimension depend on the syllabic quantity as shown in Figure 4.3). A short syllable (S) is represented by a filled circle with a diameter of x. The absence of vocalization (A) is represented by an empty circle of diameter x. A long syllable (L) is represented by a filled rectangle of width $2x$.

FIGURE 4.3: Shapes of syllable visualization

4.3.5 Case study

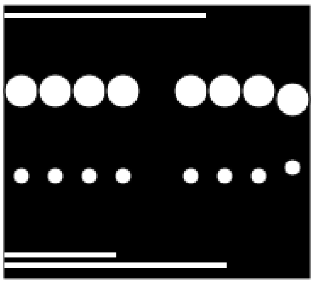
To illustrate the application of these concepts to a set of poems, we used the ten classical Arabic poems known as *al-Mu'allaqāt*, literally *The Ten Suspended Odes*. It is a collection of ten pre-Islamic Arabic poems that are considered to be the finest examples of poetry in the history of Arabic literature [17]. They were so highly appreciated that they were written in gold ink and suspended on the walls of the Kaaba in Mecca [18].

Despite some discussions about author authenticity [19], there is a consensus that theses poems are composed by ten of the most famous Arab poets in the pre-Islamic period. These poems are still celebrated for their literary and cultural significance today. They have been studied extensively by scholars and continue to inspire new generations of poets. We choose these poems because they are almost equivalent in term of semantics, affect and imagery.

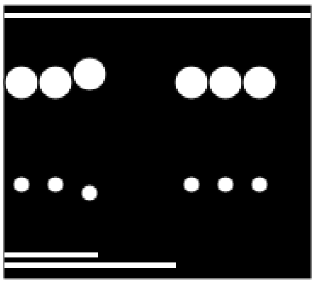
Overall, these poems count a total of 793 verses as shown in Table 4.4.

TABLE 4.4: Case study using a corpus subset: *al-Mu'allaqāt*

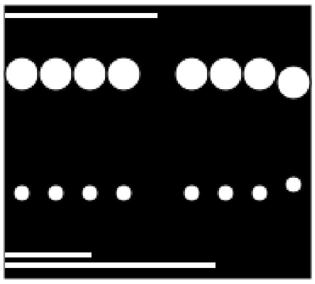
Poet	**Meter**	**Verses**
Al-A'shā	*al-Basīt.C*	66
Al-Hārith ibn Hillizah	*al-Khafīf.C*	101
Al-Nābighah al-Dhubyānī	*al-Basīt.C*	50
'Amr ibn Kulthūm	*al-Wāfir.C*	103
'Antarah ibn Shaddād	*al-Kāmil.C*	89
Labīd ibn Rabī'a	*al-Kāmil.C*	88
'Ubayd ibn al-Abras	*al-Basīt.P2*	46
Imru' al-Qays	*al-Tawīl.C*	81
Tarafa ibn al-'Abd	*al-Tawīl.C*	103
Zuhayr ibn Abī Sulmā	*al-Tawīl.C*	66
	Total:	**793**



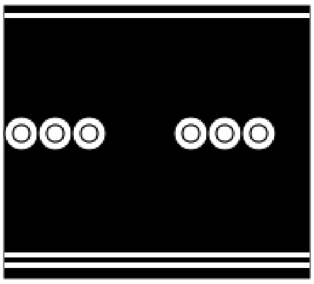
Al-A'shā (al-Basīt.C).



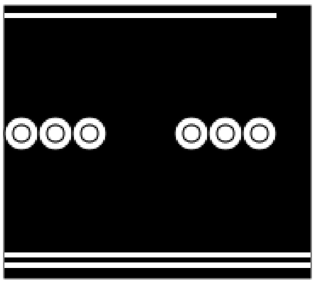
Al-Hārith (al-Khafīf.C).



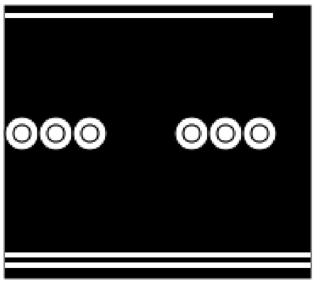
Al-Nābighah (al-Basīt.C).



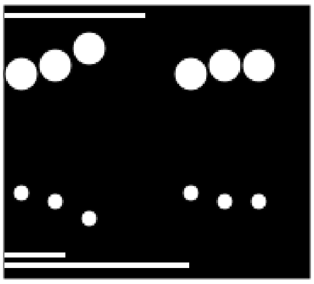
'Amr (al-Wāfir.C).



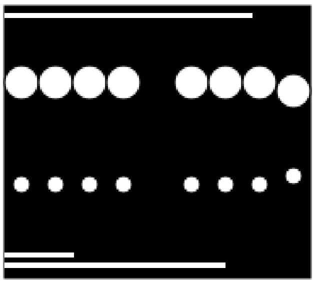
'Antarah (al-Kāmil.C).



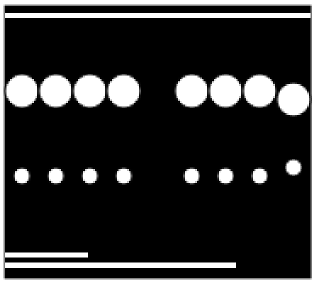
Labīd (al-Kāmil.C).



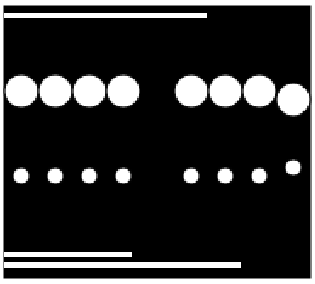
'Ubayd (al-Basīt.P2).



Imru' al-Qays (al-Tawīl.C).



Tarafa (al-Tawīl.C).



Zuhayr (al-Tawīl.C).

FIGURE 4.4: Summary visualization of the poems: *al-Mu'allaqāt*

Figure 4.4 offers insights into the poets' skills from a structural perspective. Two distinguishing elements emerge in the compositions: regularity and pattern usage. *Imru' al-Qays*, *Tarafa*, and *Zuhayr* utilize the same meter and exhibit similar structures with minimal variation. In contrast, *'Amr*, *'Antarah*, and *Labīd* display perfect regularity and employ fewer patterns. On the other hand, *Al-'A'shā*, *Al-Hārith*, and *Al-Nābighah* use more patterns and are less regular. The most irregular composition belongs to *'Ubayd*, who incorporated both forms of the *al-Basīt* meter, *al-Basīt.P1* and *al-Basīt.P2*, within his poem, likely accounting for the variation. This may reflect either a deliberate poetic choice to deviate from traditional metrical patterns or variations introduced over time through transmission. Commenting on *'Ubayd*'s poem, Al-Zawzani [18] remarked: *"His Mu'allaqa combines elements of wisdom, moral lessons, and descriptions, and it is finely crafted. However, its beauty was diminished by the meter the poet chose, a meter in which neither the ancient poets nor the early transitional poets who attempted it succeeded".*

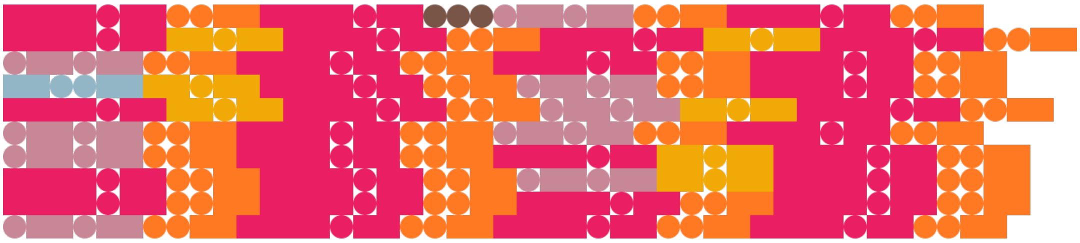


FIGURE 4.5: *Al-A'shā (al-Basīt.C)*

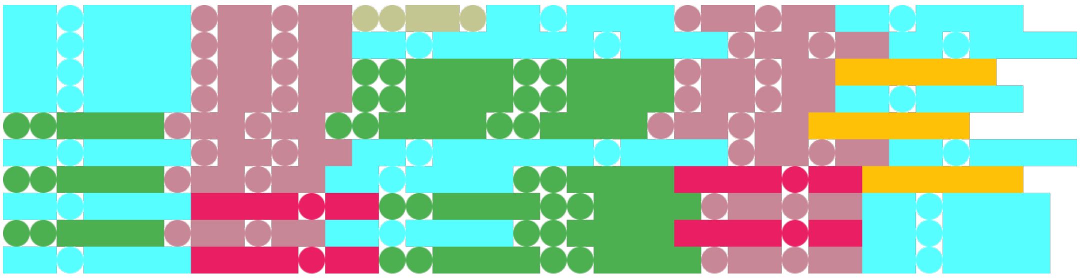


FIGURE 4.6: *Al-Hārith (al-Khafīf.C)*

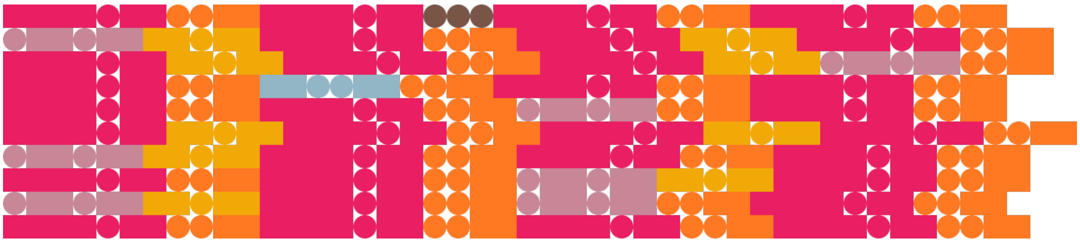


FIGURE 4.7: *Al-Nābighah (al-Basīt.C)*

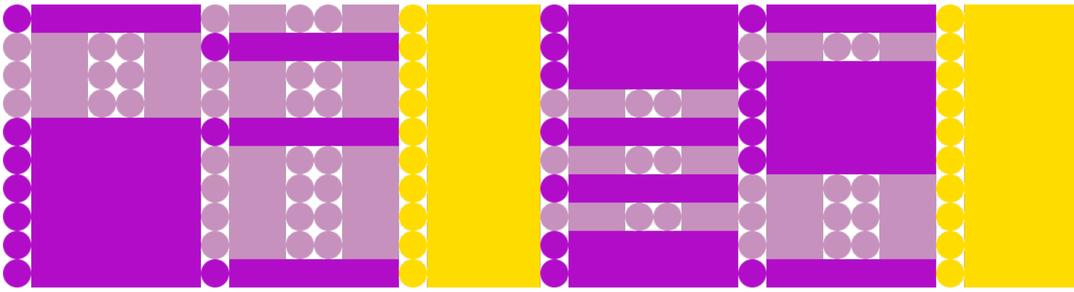


FIGURE 4.8: *'Amr (al-Wāfir.C)*

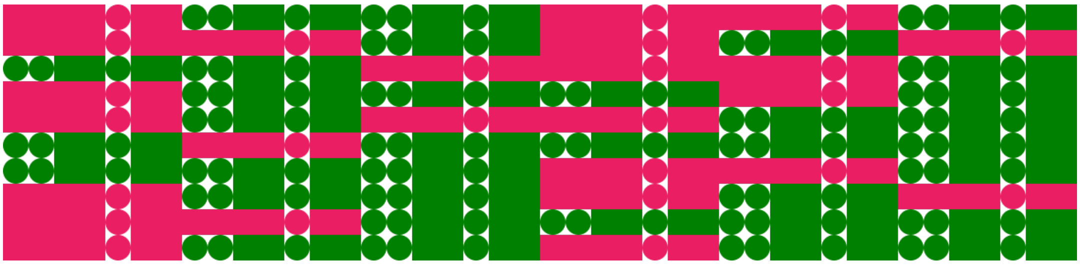


FIGURE 4.9: *'Antarah (al-Kāmil.C)*

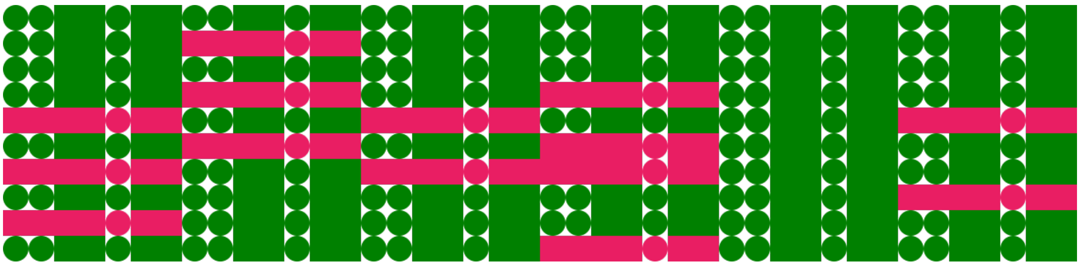


FIGURE 4.10: *Labīd (al-Kāmil.C)*

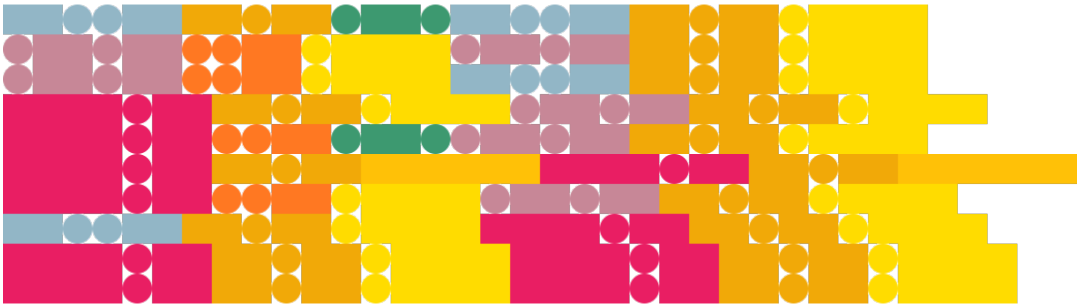


FIGURE 4.11: *'Ubayd (al-Basīt.P2)*

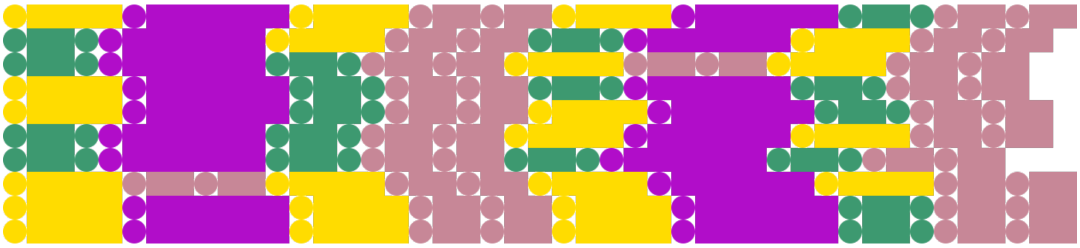


FIGURE 4.12: *Imru' al-Qays (al-Tawīl.C)*

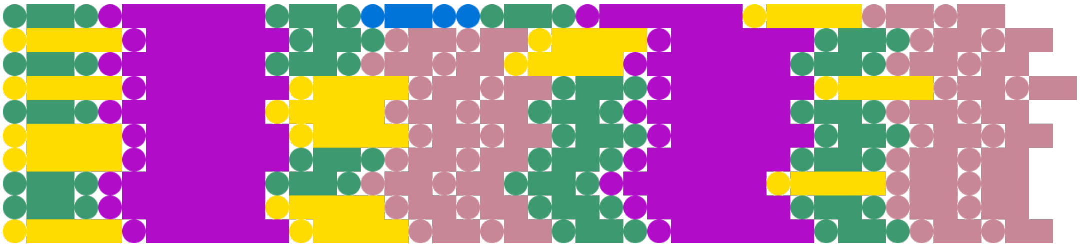


FIGURE 4.13: *Tarafa (al-Tawīl.C)*

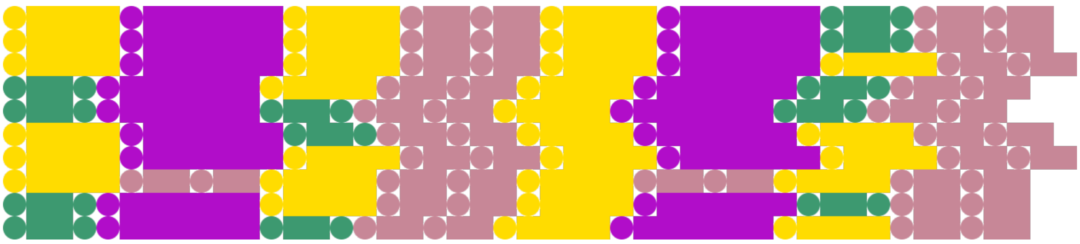


FIGURE 4.14: *Zuhayr (al-Tawīl.C)*

Although only the first ten verses are displayed, examining Figures 4.5 to 4.14 reveals where regularity and irregularity occur. We can also observe how patterns are arranged and how syllables are distributed within these patterns.

This exploratory phase confirms that visualization can indeed provide valuable insights. The main finding from this step is that the simple syllabic quantity metric is a fundamental component of metered Arabic poetry visualization, as it directly affects the lengths of the shapes.

4.4 Materials and methods

The approach adopted in this article generally focused on five main areas: data preparation, data processing, visualization, analysis, and evaluation as shown in Figure 4.15.

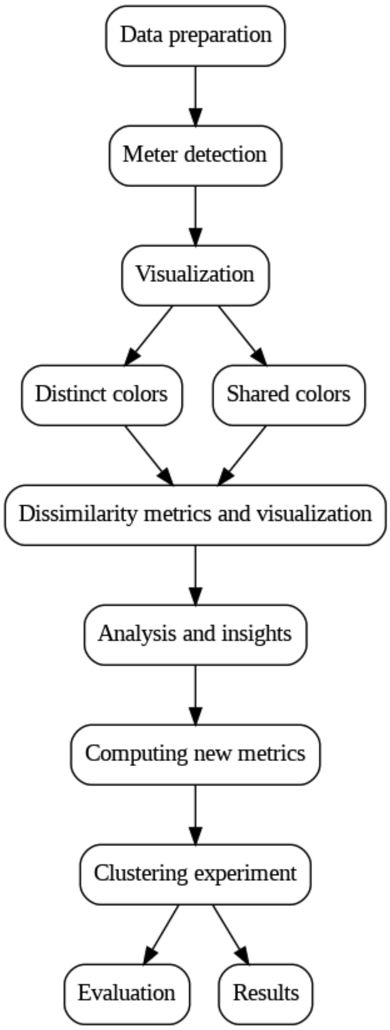


FIGURE 4.15: Research approach overview

Three poetry subsets were selected from the poetry corpus used in Chapter 2. The first subset, used for visualization exploration, consists of ten classical Arabic poems known as *al-Mu'allaqāt*, or *The ten Suspended Odes.* The second subset, chosen for visualization and dissimilarity experiments, includes poems with at least ten verses each. This subset contains 104 poems, 65 distinct poets, and 27 meter variants. To ensure uniform height for all poems converted to PNG format (Portable Network Graphics), only the first 10 verses of each poem were used for visualization, resulting in a subset of 1,040 verses. The third subset is used for clustering experiments and is the same as in Chapter 2 for clustering analysis (see Chapter 2, Subsection 2.5.2).

For each poem, the raw poetry text is converted into syllable sequences using the *AMIS* system [20], which detects the meter and outputs the patterns that compose each verse. Table 4.5 illustrates an example of a poem's structure with 4 patterns. This structure is particularly utilized after converting the patterns into their respective syllabic quantities, ensuring numerical values that are optimal for analysis and experimentation. All poems are structured in 8 columns, which is the maximum number of possible pattern positions.

For visualization, analysis, and evaluation purposes, and in order to have comparable poem dimensions, we supplemented the poems' data with dummy syllables *"N"* as shown in Table 4.6 according to the maximum number of syllables per verse of 30, as found in Arabic metrical system dataset detailed in Chapter 2.

Two visualization options were employed. The first option involved visualizing the patterns with their distinct colors. The second option utilized shared colors based on the syllabic quantity with varying opacity. Differences between images are computed in order to overlay dissimilarities on one of the compared images. This enabled the analysis of dissimilarities and provided insights that led to the development of new metrics and a clustering experiment to evaluate their effectiveness.

Although the Arabic language is written from right to left, the patterns are represented from left to right to make the reading of structures more accessible.

TABLE 4.5: Example of a poem's metrical structure represented as patterns and their syllabic quantities

P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**
LSLL (7)	LSL (5)			LSLL (7)	SSL (4)		
LSLL (7)	LSL (5)			LSLL (7)	SSL (4)		
SSLL (6)	LSL (5)			LSLL (7)	SSL (4)		
LSLL (7)	SSL (4)			LSLL (7)	LSL (5)		

TABLE 4.6: Example of a normalized poem's metrical structure with 16 dummy syllables (N)

P1	**P2**	**P5**	**P6**	**Dummy**
LSLL	LSL	LSLL	SSL	NNNNNNNNNNNNNNNN
LSLL	LSL	LSLL	SSL	NNNNNNNNNNNNNNNN
SSLL	LSL	LSLL	SSL	NNNNNNNNNNNNNNNN
LSLL	SSL	LSLL	LSL	NNNNNNNNNNNNNNNN

4.5 Visualization

Patterns are represented by colored rectangles whose lengths correspond to their respective syllabic quantities, with a mapping between the syllabic quantity and the number of pixels to draw. The difference between the verse's syllabic quantity and the maximum width is shown as white rectangles to the right of each verse, as illustrated in Figures 4.1, 4.16, 4.17, 4.18, 4.19, 4.20, and 4.21. These white rectangles represent the added dummy syllables. The patterns that do not have values are not represented, e.g., patterns 3, 4, 7, and 8 in the example of Tables 4.5 and 4.6.

4.5.1 Patterns with distinct colors

In this visualization, each pattern is assigned a distinct color. The sole criterion guiding the color selection is the need to visually differentiate one pattern from another [21]. Figure 4.16 provides an example visualization of the poem *al-Mu'allaqa*, composed by *Imru' al-Qays*, on the meter *al-Tawīl.C*.

FIGURE 4.16: Distinct colors option

4.5.2 Patterns with shared colors

In this visualization, color selection is determined by the syllabic quantity. A unique root color is assigned to each syllabic quantity, and patterns sharing the same syllabic quantity are given colors with reduced opacity, distributed linearly among the patterns. The *viridis* color palette [22] was used for its ability to provide perceptual uniformity, ensuring that changes in data are represented uniformly by changes in color intensity. Additionally, this palette enhances accessibility for individuals with color vision deficiencies. Figure 4.17 provides an example visualization of the poem *al-Mu'allaqa*, composed by *Imru' al-Qays*, on the meter *al-Tawīl.C*.

FIGURE 4.17: Shared colors option

4.5.3 Dissimilarities

To highlight differences and to quantify the disparities between pairs of images in both structure and color, a balanced dissimilarity approach is used to capture global and perceptual differences. Visualization masks are computed to indicate the regions of significant change, assisting in the interpretation and analysis of the differences between a pair of images. Both histogram and SSIM (Structural Similarity Index Measure) metrics are employed in the mask calculation process.

Histogram

Histogram Similarity measures the distribution of pixel intensities in an image [23]. Comparing the histograms of two images can reveal differences in color distribution, which may indicate changes in the scene or variations in image capture conditions. This metric is sensitive to global changes in color and intensity distribution, making it suitable for detecting significant variations in the overall appearance between two images.

The histogram mask highlights specific regions where the color or intensity differs significantly between the images. A pixel-wise mask is generated by computing the absolute difference between the two images and applying a threshold, marking pixels with differences greater than the threshold.

Structural Similarity Index Measure (SSIM)

This measure is a perceptual metric that considers changes in structural information, luminance, and contrast [24]. SSIM is calculated between the grayscale versions of the images, and the SSIM score is inverted to express dissimilarity. SSIM focuses on structural changes that are more relevant to human visual perception, making it a robust measure for detecting perceptible differences.

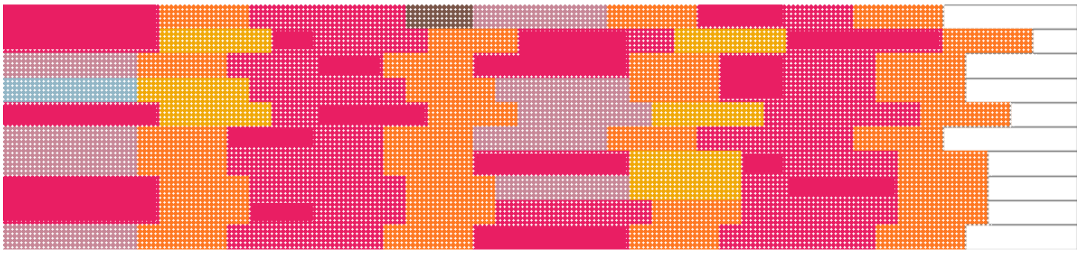
The SSIM mask highlights regions where structural changes are significant. It is generated from the SSIM difference map by marking pixels where the SSIM index is below a certain threshold. By focusing on structural dissimilarities, this mask complements the histogram mask, offering a different perspective on image differences that are more aligned with human visual perception.

Highlighting differences

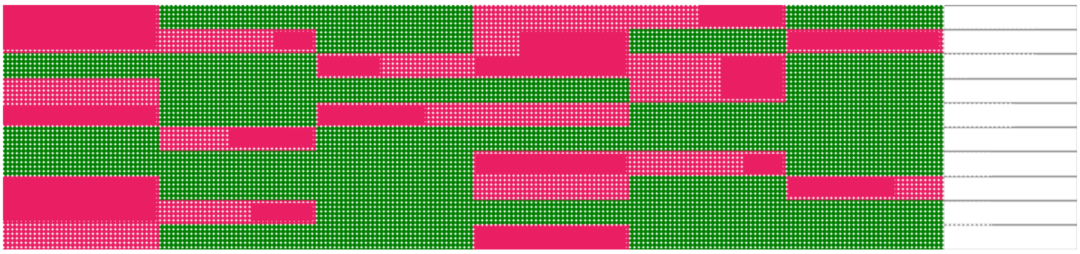
A white and a contrasted hatching pattern with specific spacing and thickness is used to overlay the areas of difference between the images. Each mask, which identifies the regions of change, is converted to a 3-channel BGR (Blue, Green, Red) image if necessary. A temporary image is created by applying the hatched pattern to the designated areas, and this is then blended with a copy of the original image, preserving its details while highlighting the differences. This approach can accommodate multiple masks, highlighting diverse types of differences within a single image.

Poem 1: Al-A'shā, al-Mu'allaqa, al-Basīt.C (poet, poem, meter).

Poem 2: 'Antarah ibn Shaddād, al-Mu'allaqa, al-Kāmil.C, (poet, poem, meter).



Highlighted differences on poem 1

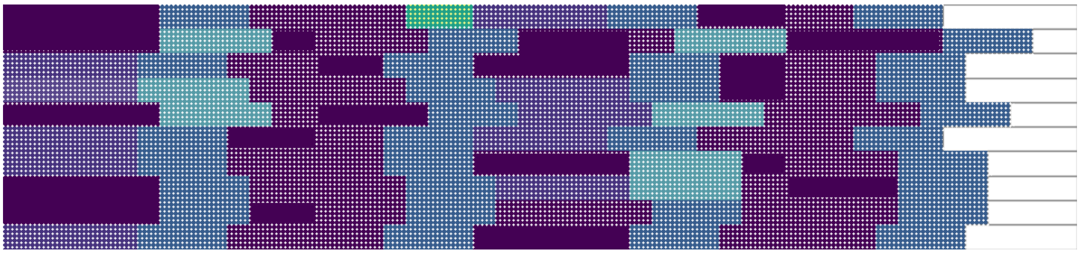


Highlighted differences on poem 2

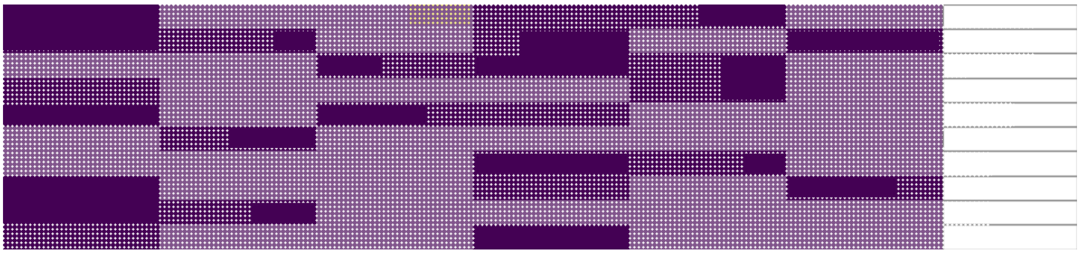
FIGURE 4.18: Dissimilarities: different meters, patterns with distinct colors

Poem 1: Al-A'shā, al-Mu'allaqa, al-Basīt.C, (poet, poem, meter).

Poem 2: 'Antarah ibn Shaddād, al-Mu'allaqa, al-Kāmil.C, (poet, poem, meter).



Highlighted differences on poem 1

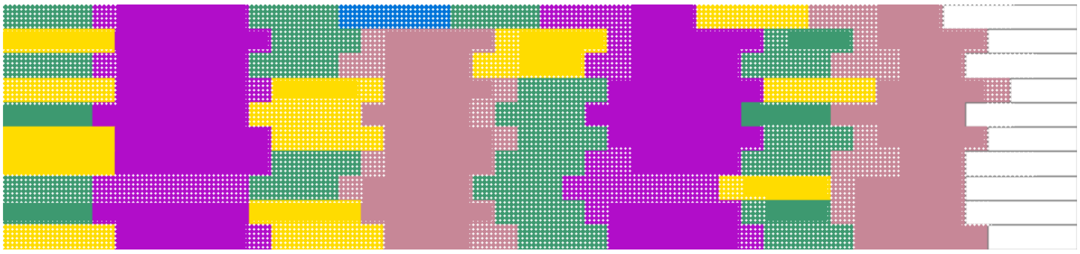


Highlighted differences on poem 2

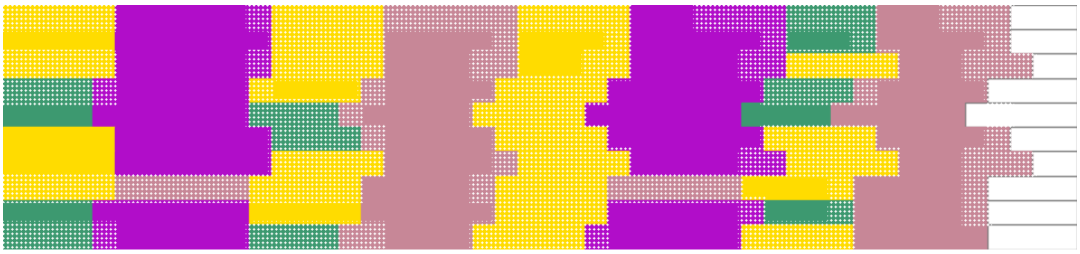
FIGURE 4.19: Dissimilarities: different meters, patterns sharing colors

Poem 1: Tarafa ibn al-'Abd, al-Mu'allaqa, al-Tawīl.C (poet, poem, meter).

Poem 2: Zuhayr ibn Abī Sulmā, al-Mu'allaqa, al-Tawīl.C, (poet, poem, meter).



Highlighted differences on poem 1



Highlighted differences on poem 2

FIGURE 4.20: Dissimilarities: same meters, patterns with distinct colors

Poem 1: Tarafa ibn al-'Abd, al-Mu'allaqa, al-Tawīl.C (poet, poem, meter).

Poem 2: Zuhayr ibn Abī Sulmā, al-Mu'allaqa, al-Tawīl.C, (poet, poem, meter).

Highlighted differences on poem 1

Highlighted differences on poem 2

FIGURE 4.21: Dissimilarities: same meters, patterns sharing colors

4.6 Experiment

When visualizing image dissimilarities, we observed that differences are highlighted not only when patterns change but also when their positions shift completely or partially. In general, dissimilarity is common, even when comparing poems with the same meter. However, in a few exceptional cases, dissimilarity disappears when the compared poems use identical patterns in the same positions.

We computed five new metrics related to the distribution of patterns across the poem to evaluate whether clustering results using these metrics outperform those obtained in the prior research detailed in Chapter 3. The experiment was conducted on the same dataset as the previous study.

In the context of metered Arabic poetry, the columns correspond to the 8 positions of the patterns. The rows correspond to the verses.

4.6.1 Distribution across columns

This metric represents the average quantity per column. It is computed by dividing the total syllabic quantity by the number of columns. This gives an indication of how much quantity, on average, is distributed across each pattern position.

$$
\begin{aligned}
D_{\text{cols}} &= \frac{Q_{\text{total}}}{N_{\text{cols}}} \\
Q_{\text{total}} &: \text{Total syllabic quantity of the poem.} \\
N_{\text{cols}} &: \text{Number of columns.}
\end{aligned}
\tag{4.2}
$$

4.6.2 Distribution across rows

This metric represents the average quantity per row. It is computed by dividing the total syllabic quantity by the number of rows. This gives an indication of how much quantity, on average, is distributed across each verse.

$$
\begin{aligned}
D_{\text{rows}} &= \frac{Q_{\text{total}}}{N_{\text{rows}}} \\
Q_{\text{total}} &: \text{Total syllabic quantity of the poem.} \\
N_{\text{rows}} &: \text{Number of rows.}
\end{aligned}
\tag{4.3}
$$

4.6.3 Distribution across cells

This metric represents the average syllabic quantity per intersection of verse and pattern position. It is computed by dividing the total syllabic quantity by the total number of intersections. This provides an indication of the average distribution of syllabic quantity across the poem's dataset.

$$
\begin{aligned}
D_{\text{cells}} &= \frac{Q_{\text{total}}}{N_{\text{rows}} \times N_{\text{cols}}} \\
Q_{\text{total}} &: \text{Total syllabic quantity of the poem.} \\
N_{\text{rows}} &: \text{Number of rows.} \\
N_{\text{cols}} &: \text{Number of columns.}
\end{aligned} \tag{4.4}
$$

4.6.4 Positions across columns

This metric represents the mean of the column position ratios that are computed as the sum of each column divided by the difference between the number of rows and the sum of that column. The mean of these ratios provides an average measure of how the quantities are positioned across the columns relative to the number of rows.

$$
\begin{aligned}
P_{\text{cols}} &= \frac{1}{N_{\text{cols}}} \sum_{j=1}^{N_{\text{cols}}} \frac{Q_{\text{col},j}}{Q_{\text{col},j} - N_{\text{rows}} + \epsilon} \\
N_{\text{cols}} &: \text{Number of columns.} \\
Q_{\text{col},j} &: \text{Sum of syllabic quantities in the } j\text{-th column.} \\
\epsilon &: \text{Small positive number to prevent division by zero.}
\end{aligned} \tag{4.5}
$$

4.6.5 Positions across rows

This metric represents the mean of the row position ratios that are computed as the sum of each row divided by the difference between the number of columns and the sum of that row. The mean of these ratios provides an average measure of how the quantities are positioned across the rows relative to the number of columns.

$$P_{\text{rows}} = \frac{1}{N_{\text{rows}}} \sum_{i=1}^{N_{\text{rows}}} \frac{Q_{\text{row},i}}{Q_{\text{row},i} - N_{\text{cols}} + \epsilon} \tag{4.6}$$

N_{rows} : Number of rows.

$Q_{\text{row},i}$: Sum of syllabic quantities in the i-th row.

ϵ : Small positive number to prevent division by zero.

4.6.6 Clustering

Selecting the optimal set of features for clustering is crucial for achieving meaningful clusters. The quality of clustering results heavily depends on the chosen features, as irrelevant or redundant features can obscure the underlying data structure, leading to poor clustering performance. In this approach, we employed a systematic method to identify the best combination of metrics among the five options (D_{cols}, D_{rows}, D_{cells}, P_{cols}, P_{rows}) to optimize performance.

We started by iterating through all possible feature combinations. Each subset of features is standardized to ensure that all contribute equally to the clustering process, minimizing the effects of varying scales.

For each feature combination, we applied the K-means clustering algorithm. Initially, to allow for comparison with previous composition clustering results [4] detailed in Chapter 3, we used the same number of clusters ($k = 3$). After analyzing this option, we also tested a configuration with $k = 4$ clusters. The clustering process is repeated 10 times to ensure stability and reliability.

Clustering performance is evaluated for each feature combination using three metrics: the Silhouette coefficient, the Davies-Bouldin index, and the Calinski-Harabasz index. Additionally, we assess clustering tendency using the Hopkins statistic. These metrics, detailed in Chapter 3, guide us in identifying the optimal feature combination based on the highest Silhouette score, the lowest Davies-Bouldin index, and the highest Calinski-Harabasz score.

The evaluation process indicates that the feature combination containing only the single metric "Positions across columns" yields the best clustering performance for the poem data.

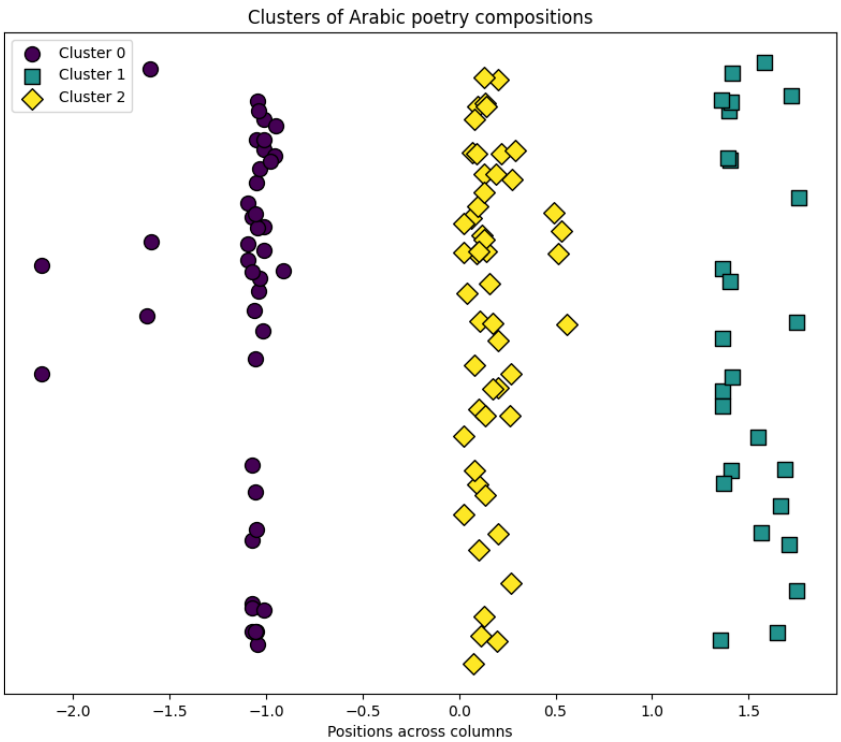


FIGURE 4.22: Clustering by positions across columns: option k=3

TABLE 4.7: Cluster membership and average syllabic quantity (Q) for k=3 and k=4 options

Cluster	Members	Average (Q)
	Option k=3	
C0	42	24.90
C1	26	40.12
C2	51	35.32
	119	
	Option k=4	
C0	37	25.97
C1	26	40.12
C2	51	35.32
C3	05	17.01
	119	

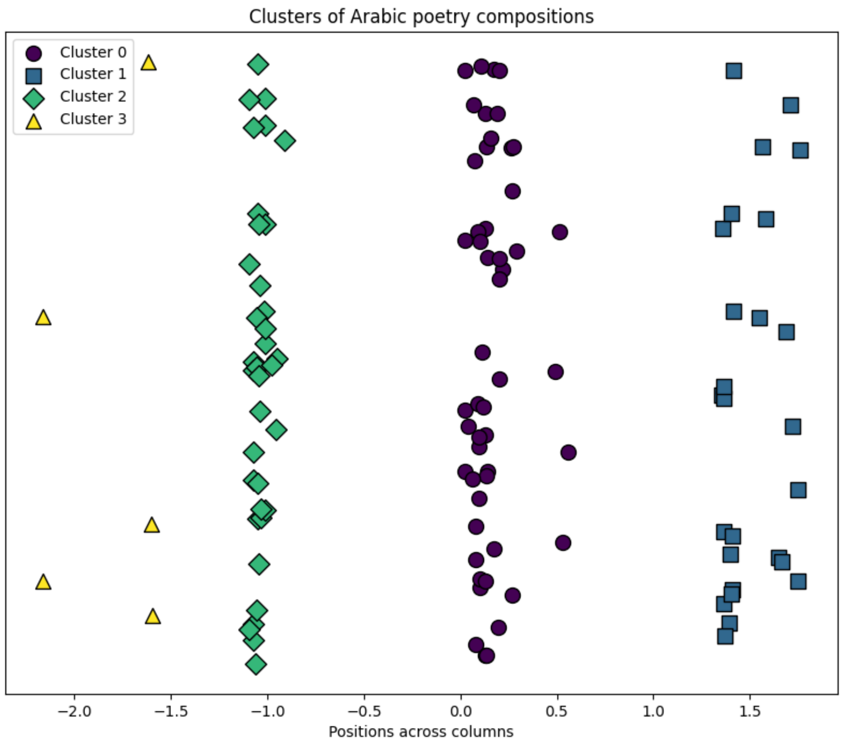


FIGURE 4.23: Clustering by positions across columns: option k=4

4.6.7 Evaluation

Table 4.8 presents a comparison of clustering scores using the "Positions across columns" metric, evaluating clustering solutions with 4-cluster *(k=4)* and 3-cluster *(k=3)* options against a previously used 3-cluster solution [4].

TABLE 4.8: Comparison of clustering scores using "Positions across columns" metric with past values

Score	**k=4**	**k=3**	**Past value (k=3)**
Hopkins	0.99	0.99	0.88
Silhouette	0.89	0.88	0.58
Davies-Bouldin	0.28	0.19	0.70
Calinski-Harabasz	2491.95	1463.30	116.57

The high Hopkins statistics values of 0.99 for both *(k=4)* and *(k=3)* indicate strong clustering tendencies in these solutions, suggesting that the data has a high likelihood of cluster structure. The improvement over the past value of 0.88 further confirms enhanced cluster validity.

The Silhouette scores of 0.89 for *(k=4)* and 0.88 for *(k=3)* indicate that both solutions produce well-defined clusters with high intra-cluster cohesion and inter-cluster separation. These scores are significantly higher than the past value of 0.58, suggesting that the updated clustering solutions offer clearer and more distinct groupings.

Regarding Davies-Bouldin Index, lower values indicate better-defined clusters. Index value for *(k=4)* (0.28) is slightly higher than for *(k=3)* (0.19), but both are substantially lower than the past value of 0.70. This indicates that both updated solutions yield more compact and well-separated clusters compared to the previous solution, though the 3-cluster configuration slightly outperforms the 4-cluster option in compactness.

For Calinski-Harabasz index, higher values indicate better-defined clusters. The scores for the solutions *(k=4)* (2491.95) and *(k=3)* (1463.30) are significantly higher than the past score of 116.57, indicating that the new clustering solutions provide greater separation between clusters. The 4-cluster configuration is particularly effective in this metric, suggesting better-defined clusters in this setup.

Overall, the updated clustering solutions with *(k=4)* and *(k=3)* clusters demonstrate substantial improvements across all metrics compared to the past 3-cluster configuration. The high Hopkins and Silhouette scores suggest that both new solutions create distinct, well-separated clusters. Although the Davies-Bouldin index slightly favors *(k=3)*, the significantly higher Calinski-Harabasz score for *(k=4)* suggests that the 4-cluster option may offer the most clearly defined separation.

4.7 Discussion

Knowing that the syllabic quantity in the Arabic metrical system, detailed in Chapter 2, ranges from 10 to 48, we observe the following:

In the 3-cluster option shown in Figure 4.22, cluster C0 contains 42 members with an average syllabic quantity of 24.90, likely representing simpler compositions within the metrical range. This cluster includes all poems that employ meters with patterns arranged in 2, 3, or 4 positions. Cluster C2 is the largest, with 51 members and an average syllabic quantity of 35.32, suggesting more elaborate compositions and placing it in the mid-range of syllabic quantity. This cluster includes poems with 6-position meters. Cluster C1 contains 26 members with the highest average syllabic quantity of 40.12, indicating the most complete compositions and primarily including poems that use meters with 8 positions.

An additional observation in cluster C0 reveals that five poems, positioned somewhat distantly from other members on the left of Figure 4.22, employ rare meters where all patterns occur in the first part of the verse, resulting in reduced syllabic quantities. The presence of these five poems is the primary distinction between the 3-cluster and the 4-cluster options summarized in Table 4.7. In the 4-cluster configuration shown in Figure 4.23, these five poems are assigned their own cluster (C3). The former C0 cluster from the 3-cluster option is divided into two clusters (C0 and C3) in the 4-cluster option, providing finer granularity by isolating a subset of simpler poems. In the 4-cluster option, C1 and C2 remain consistent across both configurations.

Standardizing all poems to 8 positions and padding with dummy syllables to reach a maximum of 30 syllables per verse enabled the comparison of poems despite differing meter patterns. Thus, empty positions in poems with fewer than 8 positions offer meaningful information, just as filled positions do. This standardization provides a consistent framework for analyzing and comparing the structural characteristics of each poem, highlighting the presence and absence of poetic elements across different compositions.

4.8 Conclusion

This research has made three significant contributions. First, it demonstrated the potential of visualization to alert analysts of Arabic poetry to insights that may warrant further exploration through new metrics. By leveraging visual representations, previously unseen patterns can be uncovered, fostering a more interactive and dynamic approach to the study and appreciation of Arabic poetry. The visualizations presented in this paper are the result of numerous explorations and attempts, as detailed in Appendix E, aimed at identifying the most effective ways to visualize Arabic poetry.

Second, the study found that the metric related to the positioning of patterns across columns is a key feature for clustering poetic compositions and highlighting the degree of similarity or dissimilarity among poetic structures.

Finally, the creation and availability of a corpus of Arabic poetry in the form of images facilitates the visual exploration of poetic compositions. This study has addressed the research question and contributed to enhancing the understanding of the intricate patterns inherent in Arabic poetry.

References

[1] Lawrence Venuti. "Introduction: Poetry and translation". In: *Translation studies* 4.2 (2011), pp. 127–132.

[2] Daniela Oelke, Dimitrios Kokkinakis, and Mats Malm. "Advanced visual analytics methods for literature analysis". In: *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. 2012, pp. 35–44.

[3] Riccardo Mazza. *Introduction to information visualization*. Springer Science & Business Media, 2009.

[4] Abdelmalek Berkani and Adrian Holzer. "Clustering analysis of metered Arabic poetry compositions". In: *2023 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA)*. IEEE. 2023, pp. 1–8.

[5] Mirella De Sisto et al. "Understanding poetry using natural language processing tools: a survey". In: *Digital Scholarship in the Humanities* 39.2 (2024), pp. 500–521.

[6] Nitin Madnani. "Emily: A tool for visual poetry analysis". In: *University of Maryland, Tech. Rep* (2005).

[7] Luis Meneses and Richard Furuta. “Visualizing poetry: creating Tools for critical analysis”. In: *paj: The Journal of the Initiative for Digital Humanities, Media, and Culture* 3 (2012).

[8] Jason R Robinson. “Colors of poetry: Computational deconstruction”. MA thesis. University of Georgia, 2006.

[9] Alfie Abdul-Rahman et al. “Rule-based Visual Mappings - with a Case Study on Poetry Visualization”. In: *Computer Graphics Forum*. The Eurographics Association; 2013. DOI: 10.1111/cgf.12125.

[10] Nina McCurdy et al. “Poemage: Visualizing the sonic topology of a poem”. In: *IEEE transactions on visualization and computer graphics* 22.1 (2015), pp. 439–448.

[11] Rodolfo Delmonte. “Visualizing poetry with SPARSAR Visual maps from poetic content”. In: *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*. 2015, pp. 68–78.

[12] Adiel Mittmann, Aldo von Wangenheim, and Alckmar Luiz dos Santos. “A multi-level visualization scheme for poetry”. In: *2016 20th International Conference Information Visualisation (IV)*. IEEE. 2016, pp. 312–317.

[13] Onur Musaoglu et al. “A Generic Tool for Visualizing Patterns in Poetry.” In: *DH*. 2017.

[14] Zak Risha, Deniz Sonmez Unal, and Erin Walker. “Soliloquy: Fostering Poetry Comprehension Using an Interactive Think-aloud Visualization”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–16.

[15] Ya-Juan Gao et al. “A Study of Communication in Turning Poetry into Painting”. In: *Cross-Cultural Design* (2017).

[16] Fred Lerdahl. "The sounds of poetry viewed as music". In: *Annals of the New York Academy of Sciences* 930.1 (2001), pp. 337–354.

[17] Gregor Schoeler. "The Genres of Classical Arabic Poetry Classifications of Poetic Themes and Poems by Pre-Modern Critics and Redactors of diwans". In: *Quaderni di studi arabi* (2010), pp. 1–48.

[18] Al-Zawzani. *Explanation of the Ten Mu'allaqat*. Arabic. Beirut: Dar Maktabat Al Hayat, 1983.

[19] James T Monroe. "Oral composition in pre-Islamic poetry". In: *Journal of Arabic Literature* 3.1 (1972), pp. 1–53.

[20] Abdelmalek Berkani, Adrian Holzer, and Kilian Stoffel. "Pattern matching in meter detection of Arabic classical poetry". In: *2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA)*. IEEE. 2020, pp. 1–8.

[21] Maureen Stone. "Choosing colors for data visualization". In: *Business Intelligence Network* 2 (2006).

[22] J. D. Hunter. "Matplotlib: A 2D graphics environment". In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55.

[23] Md Al-Amin Bhuiyan and Abdul Raouf Khan. "Image quality assessment employing RMS contrast and histogram similarity." In: *Int. Arab J. Inf. Technol.* 15.6 (2018), pp. 983–989.

[24] Zhou Wang et al. "Image quality assessment: from error visibility to structural similarity". In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612.

Chapter 5

Conclusion

5.1 Research assessment

The primary objective of this thesis was to address the issue of insufficiency of data and natural language processing tools related to the metrical system of Arabic poetry. To alleviate this issue, we developed the *Arabic Meters Identification System (AMIS)*, a comprehensive framework that integrates metrical reference data and natural language processing tools for the structural analysis of metered Arabic poetry.

Metrical system data. We designed the metrical system data component to incorporate all reference data related to known meter variants and patterns in Arabic literature. As detailed in Chapter 2, we collected, and structured all metrical combinations documented in the literature, adapting the data to handle both typical cases and exceptional instances, such as single-verse detection. This component ensures the preservation of verse concordance and was evaluated using a diverse poetry corpus. The resulting Arabic metrical system dataset contains 29 meter variants, 43 distinct patterns, and 23,518 pattern combinations.

Meter identification. Besides the metrical data, we developed several natural language processing tools for meter identification, covering all stages from text preparation to final meter detection. Various methods were employed, including exact and similarity matching, both of which rely on the reference data from the metrical system. As discussed in Chapter 2, we evaluated these methods across different versions of *AMIS* and the poetry corpus, assessing their performance in detecting meters with a high accuracy of 99.97%. The verses for which the meter was not detected are those where there is a voluntary or involuntary deviation by the poet in word choice, resulting in too many or too few syllables and, consequently, a lack of correspondence with any occurrence in the metrical system data.

Structure analysis. Once meters were detected, the next step was to analyze the structures of these poetic compositions. The aim was to identify metrics that could facilitate the comparison of different poems, independent of semantics, poets, eras, or specific meters. In Chapter 3, we identified five key composition metrics: *Pattern Usage*, which evaluates the extent of pattern use in relation to the number of verses; *Pattern Potential*, which reflects unused patterns; *Repetition*, measuring the maximum recurrence of syllabic sequences; *Variability*, indicating the variation in syllabic quantity; and *Evenness*, which measures how evenly patterns are distributed throughout the poem. After conducting a clustering experiment, three metrics, *Repetition*, *Variability*, and *Pattern Usage*, emerged as the most significant factors in clustering metered Arabic poetry. The results also suggest that it is more effective to analyze these metrics in pairs rather than all three simultaneously. Specifically, the combinations of *Pattern Usage* with *Repetition*, or *Variability* with *Repetition*, produced meaningful clustering into three distinct groups.

Visualization. To complement traditional analytical methods, we explored the use of visualization to uncover new metrics and insights within Arabic poetry. After organizing the poetry data in a suitable format to be converted into images of uniform size, we adopted an exploratory approach that allowed us to visualize entire poems in a summarized form and depict detailed pattern arrangements within subsets of verses. As detailed in Chapter 4, these visualizations revealed five additional key metrics for analyzing the structure of metered Arabic poetry. The first three metrics focused on distribution: *Distribution across columns*, *Distribution across rows*, and *Distribution across cells*. These metrics measure the average syllabic quantity allocated to columns, rows, and individual cells, respectively, offering insights into the overall syllabic distribution within a poem. The next two metrics, *Positions across columns* and *Positions across rows*, capture the relative positioning of syllabic quantities, considering the alignment of patterns in relation to both columns and rows. We found that the metric related to the positioning of patterns across columns is a key feature for clustering poetic compositions and for highlighting the degree of similarity or dissimilarity among poetic structures. This metric improved clustering outcomes, confirming that visualization is a valuable tool for the structural analysis of metered Arabic poetry.

Poetry corpus. A critical component of this research was the poetry corpus, compiled with the aim of ensuring diversity in poetic structure, and used, fully or partially, throughout the thesis. We manually collected and vowelized the poems, a labor-intensive but necessary task. The final version of the poetry corpus consists of 3,450 verses, 126 poems, 69 distinct poets, and 28 distinct announced meters, spanning 9 historical eras from the pre-Islamic period to the present day. This comprehensive dataset provided a robust foundation for evaluating the metrical system data, conducting meter detection, and performing structural analysis and visualization. Notably, the visualization process led to the creation of an image corpus, which complements the traditional text corpus and further expands the analytical possibilities of the research.

5.2 Future research

The research conducted in this thesis has highlighted several open directions for future investigation. We outline three key areas below.

Automated multilingual corpus. In this thesis, the corpus was manually compiled and vowelized, a time-consuming effort that underscores the need for automation in future research, as such a manual approach is not easily scalable. A significant area of future research is the development of a large-scale, multilingual Arabic poetry corpus to serve both Arabic and non-Arabic audiences. Ideally, this corpus would include versions in at least Arabic, French, and English to maximize accessibility and cultural exchange. Achieving this would require automation techniques for vowelization, potentially through rule-based, statistical, machine learning, or hybrid methods [1]. Furthermore, the promising capabilities of large language models could streamline diacritization, translation, summarization, and other NLP tasks such as part-of-speech tagging (POS) [2].

Poet skill classification. While this thesis focused on the structural aspects of Arabic poetry, future research could expand this work by incorporating semantic features to classify poets' skills through metrics that quantify their proficiency in structural, semantic, and stylistic elements. Such metrics could assess mastery over metrical patterns, semantic richness, thematic complexity, and the use of rhetorical and stylistic figures. The existing literature includes various critiques and evaluations [3], as well as classifications of poets [4], which are primarily linked to the semantic level. It would be insightful to investigate whether the assessments provided by critics could be confirmed or refined through these metrics, using both structural and semantic dimensions.

Computational aesthetics and visualization. Another area of future research is the computational evaluation of the aesthetics of Arabic poetry. It is essential to assess whether visual representations enhance users' understanding of Arabic poetry or evoke a sense of aesthetic beauty similar to that experienced through reading or listening. Future research could focus on three main dimensions: first, a semantic dimension [5] that explores word choice [6] and style [7]; second, a structural dimension that examines poetic composition, rhythm, and rhyme [8]; and third, the interaction between semantics and structure through recurrent formulaic styles [9]. Achieving this would require developing new types of visualizations capable of representing not only the structural aspects of poetry but also conveying semantic content [10], enabling a more comprehensive aesthetic experience. Such visualizations would also be particularly valuable for non-Arabic speakers, allowing them to appreciate the poetic structure and thematic richness without needing language proficiency. Beauty appreciation could be evaluated through the development of quantified metrics for objective measurement, or through empirical studies using participant evaluations. This multidimensional approach, combining semantics and structure through computational metrics and human evaluations, would contribute to a more holistic understanding of how visualizations impact the appreciation of Arabic poetry.

References

[1] Manar M Almanea. “Automatic methods and neural networks in Arabic texts diacritization: a comprehensive survey”. In: *IEEE Access* 9 (2021), pp. 145012–145032.

[2] Zaid Alyafeai et al. “Taqyim: Evaluating arabic nlp tasks using chatgpt models”. In: *arXiv preprint arXiv:2306.16322* (2023).

[3] Abu Said Al Asma’i. *Eminent Poets*. Arabic. Beirut: Dar Al Kitab Al Jadid, 1980.

[4] Muhamed Ben Salam Al Jomahi. *Classes of Eminent Poets*. Arabic. Beirut: Dar Al Kotob Al ’Ilmya, 1988.

[5] Amy M Belfi, Edward A Vessel, and G Gabrielle Starr. “Individual ratings of vividness predict aesthetic appeal in poetry.” In: *Psychology of Aesthetics, Creativity, and the Arts* 12.3 (2018), p. 341.

[6] Arthur M Jacobs. “Quantifying the beauty of words: a neurocognitive poetics perspective”. In: *Frontiers in Human Neuroscience* 11 (2017), p. 622.

[7] Justine T Kao and Dan Jurafsky. “A computational analysis of poetic style: Imagism and its influence on modern professional and amateur poetry”. In: *Linguistic Issues in Language Technology* 12 (2015).

[8] Christian Obermeier et al. “Aesthetic and emotional effects of meter and rhyme in poetry”. In: *Frontiers in psychology* 4 (2013), p. 10.

[9] Bruno Paoli. “Métrique arabe: au-delà du cercle vicieux des théories (Arabic Metrics: beyond the vicious circle of theories)”. French. In: *Bulletin d’études orientales* 65 (2017), pp. 177–216.

[10] Winfried Menninghaus et al. “The emotional and aesthetic powers of parallelistic diction”. In: *Poetics* 63 (2017), pp. 47–59.

Part IV

Appendices

Appendix A

Meter ambiguities: pruning code

al-Kāmil.C and *al-Rajaz.C*: Pattern *LLSL (mustaf'ilun)* is more present in *al-Rajaz.C* than *al-Kāmil.C*.

```
DELETE data
WHERE meter = "al-Kamil.C"
AND fp1s = "LLSL"
AND fp2s = "LLSL"
AND fp3s IN ("LLSL","LLSS")
AND sp1s = "LLSL"
AND sp2s = "LLSL"
AND sp3s = "LLSL";

DELETE data
WHERE meter ="al-Rajaz.C"
AND fp1s = "LLSL"
AND fp2s ="LLSL"
AND fp3s IN ("LLSL","LLSS")
AND sp1s = "LLSL"
AND sp2s ="LLSL"
AND sp3s = "LLL";
```

al-Kāmil.C and *al-Sarī'.C*: Pattern *LLSL (mustaf'ilun)* is more present in *in al-Sarī'.C* than *al-Kāmil.C*.

```
DELETE data
```

```
WHERE meter = "al-Kamil.C"
AND fp1s = "LLSL"
AND fp2s = "LLSL"
AND fp3s in ("SSL","SSS")
AND sp1s = "LLSL"
AND sp2s = "LLSL"
AND sp3s IN ("SSL","LL");
```

al-Kāmil.P and *al-Rajaz.P1.*: Pattern *LLSL (mustaf'ilun)* is more present in *al-Rajaz.P1.* than *al-Kāmil.P.*

```
DELETE data
WHERE meter ="al-Kamil.P"
AND fp1s = "LLSL"
AND fp2s IN ("LLSL","LLSS")
AND sp1s = "LLSL"
AND sp2p ="LLSL";
```

al-Wāfir.P and *al-Hazaj.C*: Pattern *SLLL (mafā'īlun)* is more present in *al-Hazaj.C* than *al-Wāfir.P.*

```
DELETE data
WHERE meter ="al-Wafir.P"
AND fp1s = "SLLL"
and fp2s IN ("SLLL","SLLS")
and sp1s = "SLLL"
and sp2s = "SLLL";
```

al-Rajaz.P1 and *al-Hazaj.C*: Pattern *SLSL (mafā'ilun)* is more present in *al-Hazaj.C* than *al-Rajaz.P1.*

```
DELETE data
WHERE meter = "al-Rajaz.P1."
AND fp1s = "SLSL"
AND fp2s IN ("SLSL","SLSS", "SLL", "SLS")
AND sp1s ="SLSL"
AND sp2s IN ("SLSL","SLL");
```

al-Ramal.P and *al-Madīd.P*: Pattern *LSLL (fā'ilātun)* is more present in *al-Ramal.P* than *al-Madīd.P*.

```
DELETE data
WHERE meter ="al-Madid.P"
AND fp1s IN ("LSLL","SSLL")
AND fp2s IN ("LSL", "LSS", "SSL","SSS")
AND sp1s IN ("LSLL", "SSLL")
AND sp2s IN ("LSL","SSL","LSLA");
```

Appendix B

Metrical system data: view by meter

Example: in meter view for *al-Tawīl.C*, the indication *fa'ūlu*(1,3,4,5,7) means that the pattern *fa'ūlu (SLS)* is present in positions 1,3,4,5 and 7 but not in 2, 6 and 8. Meter *al-Tawīl.C* has 8 pattern positions, 4 in each hemistich.

Meter	**Pattern**	**Positions**
al-Tawīl.C	*fa'ūlun (SLL)*	1, 3, 4, 5, 7, 8
	mafā'īlun (SLLL)	2, 4, 6, 8
	mafā'īlu (SLLS)	4
	fa'ūlu (SLS)	1, 3, 4, 5, 7
	mafā'ilun (SLSL)	2, 4, 6, 8
	mafā'ilu (SLSS)	4

Meter	**Pattern**	**Positions**
al-Madīd.C	*fi'lun (LL)*	6
	fā'ilun (LSL)	2, 3, 5, 6
	fā'ilān (LSLA)	6
	fā'ilātun (LSLL)	1, 3, 4, 6
	fā'ilātu (LSLS)	3
	fā'ilu (LSS)	3
	fa'ilun (SSL)	2, 3, 5, 6
	fa'ilātun (SSLL)	1, 3, 4, 6
	fa'ilātu (SSLS)	3
	fa'ilu (SSS)	3

Meter	Pattern	Positions
al-Madīd.P	*fi'lun (LL)*	4
	fā'ilun (LSL)	2
	fā'ilātun (LSLL)	1, 3
	fā'ilu (LSS)	2
	fa'ilun (SSL)	2
	fa'ilātun (SSLL)	1, 3
	fa'ilu (SSS)	2

Meter	Pattern	Positions
al-Basīt.C	*fi'lun (LL)*	4, 8
	mustaf'ilun (LLSL)	1, 3, 5, 7
	fi'lu (LS)	4
	fā'ilun (LSL)	2, 6
	mufta'ilun (LSSL)	1, 3, 5, 7
	mafā'ilun (SLSL)	1, 3, 5, 7
	fa'ilun (SSL)	2, 4, 6, 8
	fa'ilu (SSS)	4

Meter	Pattern	Positions
al-Basīt.P1	*maf'ūlun (LLL)*	6
	mustaf'ilun (LLSL)	1, 3, 4, 6
	mustaf'ilān (LLSLA)	3, 6
	mustaf'ilu (LLSS)	3
	fā'ilun (LSL)	2, 5
	mufta'ilun (LSSL)	1, 3, 4
	mufta'ilu (LSSS)	3
	fa'ūlun (SLL)	6
	mafā'ilun (SLSL)	1, 3, 4
	mafā'ilu (SLSS)	3
	fa'ilun (SSL)	2, 5

Meter	Pattern	Positions
al-Basīt.P2	*maf'ūlun (LLL)*	2, 3, 5, 6
	maf'ūlu (LLS)	3
	mustaf'ilun (LLSL)	1, 4
	fā'ilun (LSL)	2, 5
	mufta'ilun (LSSL)	1, 4
	fa'ūlun (SLL)	3, 6
	fa'ūlu (SLS)	3
	mafā'ilun (SLSL)	1, 4
	fa'ilun (SSL)	2, 5

Meter	**Pattern**	**Positions**
al-Wāfir.C	*fa'ūlun (SLL)*	3, 6
	mafā'īlun (SLLL)	1, 2, 4, 5
	fa'ūlu (SLS)	3
	mufā'alatun (SLSSL)	1, 2, 4, 5

Meter	**Pattern**	**Positions**
al-Wāfir.P	*mafā'īlun (SLLL)*	1, 2, 3, 4
	mafā'īlu (SLLS)	2
	mufā'alatun (SLSSL)	1, 2, 3, 4
	mufā'alatu (SLSSS)	2

Meter	**Pattern**	**Positions**
al-Kāmil.C	*fi'lun (LL)*	6
	maf'ūlun (LLL)	6
	mustaf'ilun (LLSL)	1, 2, 3, 4, 5, 6
	mustaf'ilu (LLSS)	3
	fa'ilun (SSL)	3, 6
	fa'ilātun (SSLL)	3, 6
	fa'ilātu (SSLS)	3
	mutafā'ilun (SSLSL)	1, 2, 3, 4, 5, 6
	mutafā'ilu (SSLSS)	3
	fa'ilu (SSS)	3

Meter	**Pattern**	**Positions**
al-Kāmil.P	*mustaf'ilun (LLSL)*	1, 2, 3
	mustaf'ilātun (LLSLL)	2, 4
	mustaf'ilātu (LLSLS)	2
	mustaf'ilu (LLSS)	2
	mutafā'ilun (SSLSL)	1, 2, 3, 4
	mutafā'ilān (SSLSLA)	4
	mutafā'ilātun (SSLSLL)	2, 4
	mutafā'ilātu (SSLSLS)	2
	mutafā'ilu (SSLSS)	2

Meter	Pattern	Positions
al-Hazaj.C	*fa'ūlun (SLL)*	2, 4
	mafā'īlun (SLLL)	1, 2, 3, 4
	mafā'īlu (SLLS)	1, 2, 3
	fa'ūlu (SLS)	2
	mafā'ilun (SLSL)	1, 2, 3, 4
	mafā'ilu (SLSS)	1, 2, 3

Meter	Pattern	Positions
al-Rajaz.C	*maf'ūlun (LLL)*	3, 6
	maf'ūlān (LLLA)	3, 6
	maf'ūlu (LLS)	3
	mustaf'ilun (LLSL)	1, 2, 3, 4, 5, 6
	mustaf'ilu (LLSS)	3
	mufta'ilun (LSSL)	1, 2, 3, 4, 5, 6
	mufta'ilu (LSSS)	3
	fa'ūlun (SLL)	3, 6
	fa'ūlān (SLLA)	3, 6
	fa'ūlu (SLS)	3
	mafā'ilun (SLSL)	1, 2, 3, 4, 5, 6
	mafā'ilu (SLSS)	3
	fa'ilatun (SSSL)	1, 2, 4, 5

Meter	Pattern	Positions
al-Rajaz.P1	*maf'ūlun (LLL)*	2, 4
	maf'ūlān (LLLA)	2, 4
	maf'ūlu (LLS)	2
	mustaf'ilun (LLSL)	1, 2, 3, 4
	mustaf'ilu (LLSS)	2
	mufta'ilun (LSSL)	1, 2, 3, 4
	mufta'ilu (LSSS)	2
	fa'ūlun (SLL)	2, 4
	fa'ūlān (SLLA)	2, 4
	fa'ūlu (SLS)	2
	mafā'ilun (SLSL)	1, 2, 3, 4
	mafā'ilu (SLSS)	2
	fa'ilatun (SSSL)	1, 3

Meter	**Pattern**	**Positions**
al-Rajaz.P2	*maf'ūlun (LLL)*	3
	maf'ūlān (LLLA)	3
	mustaf'ilun (LLSL)	1, 2, 3
	mufta'ilun (LSSL)	1, 2, 3
	fa'ūlun (SLL)	3
	fa'ūlān (SLLA)	3
	mafā'ilun (SLSL)	1, 2, 3
	fa'ilatun (SSSL)	1, 2

Meter	**Pattern**	**Positions**
al-Rajaz.P3	*maf'ūlun (LLL)*	2
	mustaf'ilun (LLSL)	1, 2
	mufta'ilun (LSSL)	1, 2
	fa'ūlun (SLL)	2
	mafā'ilun (SLSL)	1, 2
	fa'ilatun (SSSL)	1

Meter	**Pattern**	**Positions**
al-Ramal.C	*fā'ilun (LSL)*	3, 6
	fā'ilān (LSLA)	3, 6
	fā'ilātun (LSLL)	1, 2, 3, 4, 5, 6
	fā'ilātān (LSLLA)	3, 6
	fā'ilātu (LSLS)	3
	fā'ilu (LSS)	3
	fa'ilun (SSL)	3, 6
	fa'ilān (SSLA)	3, 6
	fa'ilātun (SSLL)	1, 2, 3, 4, 5, 6
	fa'ilātān (SSLLA)	3, 6
	fa'ilātu (SSLS)	3
	fa'ilu (SSS)	3

Meter	Pattern	Positions
al-Ramal.P	*fā'ilun (LSL)*	2, 4
	fā'ilān (LSLA)	2, 4
	fā'ilātun (LSLL)	1, 2, 3, 4
	fā'ilātān (LSLLA)	2, 4
	fā'ilātu (LSLS)	2
	fā'ilu (LSS)	2
	fa'ilun (SSL)	2, 4
	fa'ilān (SSLA)	2, 4
	fa'ilātun (SSLL)	1, 2, 3, 4
	fa'ilātān (SSLLA)	2, 4
	fa'ilātu (SSLS)	2
	fa'ilu (SSS)	2

Meter	Pattern	Positions
al-Sarī'.C	*fi'lun (LL)*	3, 6
	mustaf'ilun (LLSL)	1, 2, 4, 5
	fi'lu (LS)	3
	fā'ilun (LSL)	3, 6
	fā'ilān (LSLA)	3, 6
	fā'ilu (LSS)	3
	mufta'ilun (LSSL)	1, 2, 4, 5
	mafā'ilun (SLSL)	1, 2, 4, 5
	fa'ilun (SSL)	3, 6
	fa'ilu (SSS)	3
	fa'ilatun (SSSL)	1, 2, 4, 5

Meter	Pattern	Positions
al-Munsarih.C	*maf'ūlun (LLL)*	6
	maf'ūlātu (LLLS)	2, 5
	mustaf'ilun (LLSL)	1, 3, 4
	mustaf'ilu (LLSS)	3
	fā'ilātu (LSLS)	2, 5
	mufta'ilun (LSSL)	1, 3, 4, 6
	mufta'ilu (LSSS)	3
	mafā'īlu (SLLS)	2, 5
	mafā'ilun (SLSL)	1, 4

Meter	Pattern	Positions
al-Khafīf.C	*maf'ūlun (LLL)*	3, 6
	maf'ūlu (LLS)	3
	mustaf'ilun (LLSL)	2, 5
	fā'ilun (LSL)	3, 6
	fā'ilātun (LSLL)	1, 3, 4, 6
	fā'ilātu (LSLS)	3
	fā'ilu (LSS)	3
	mafā'ilun (SLSL)	2, 5
	fa'ilātun (SSLL)	1, 3, 4, 6
	fa'ilātu (SSLS)	3

Meter	Pattern	Positions
al-Khafīf.P	*mustaf'ilun (LLSL)*	2, 4
	mustaf'ilu (LLSS)	2
	fā'ilātun (LSLL)	1, 3
	fa'ūlun (SLL)	4
	mafā'ilun (SLSL)	2, 4
	mafā'ilu (SLSS)	2
	fa'ilātun (SSLL)	1, 3

Meter	Pattern	Positions
al-Mudāri'.C	*fā'ilātun (LSLL)*	2, 4
	fā'ilātu (LSLS)	2
	mafā'īlun (SLLL)	1, 3
	mafā'īlu (SLLS)	1, 3

Meter	Pattern	Positions
al-Muqtadab.C	*maf'ūlātu (LLLS)*	1, 3
	fā'ilātu (LSLS)	1, 3
	mufta'ilun (LSSL)	2, 4
	mufta'ilu (LSSS)	2
	mafā'īlu (SLLS)	1, 3

Meter	Pattern	Positions
al-Mujtath.C	*maf'ūlun (LLL)*	4
	mustaf'ilun (LLSL)	1, 3
	fā'ilātun (LSLL)	2, 4
	fā'ilātu (LSLS)	2
	mafā'ilun (SLSL)	1, 3
	fa'ilātun (SSLL)	2, 4
	fa'ilātu (SSLS)	2

Meter	Pattern	Positions
al-Mutaqārib.C	*fa' (L)*	4, 8
	fa'al (SL)	4, 8
	fa'ūl (SLA)	4, 8
	fa'ūlun (SLL)	1, 2, 3, 4, 5, 6, 7, 8
	fa'ūlu (SLS)	1, 2, 3, 4, 5, 6, 7

Meter	Pattern	Positions
al-Mutaqārib.P	*fa' (L)*	3, 6
	fa'al (SL)	3, 6
	fa'ūl (SLA)	3, 6
	fa'ūlun (SLL)	1, 2, 3, 4, 5, 6
	fa'ūlu (SLS)	1, 2, 3, 4, 5

Meter	Pattern	Positions
al-Mutadārak.C	*fi'lun (LL)*	1, 2, 3, 4, 5, 6, 7, 8
	fi'lu (LS)	4
	fā'ilun (LSL)	1, 2, 3, 4, 5, 6, 7, 8
	fā'ilu (LSS)	4
	fa'ilun (SSL)	1, 2, 3, 4, 5, 6, 7, 8
	fa'ilu (SSS)	4

Meter	Pattern	Positions
al-Mutadārak.P1	*fi'lun (LL)*	1, 2, 3, 4, 5, 6
	fi'lu (LS)	3
	fā'ilun (LSL)	1, 2, 3, 4, 5, 6
	fā'ilān (LSLA)	3, 6
	fā'ilu (LSS)	3
	fa'ilun (SSL)	1, 2, 3, 4, 5, 6
	fa'ilu (SSS)	3

Meter	**Pattern**	**Positions**
al-Mutadārak.P2	*fi'lun (LL)*	1, 2, 3, 4
	fi'lu (LS)	2
	fā'ilun (LSL)	1, 2, 3, 4
	fā'ilān (LSLA)	2, 4
	fā'ilu (LSS)	2
	fa'ilun (SSL)	1, 2, 3, 4
	fa'ilu (SSS)	2

Appendix C

Metrical system data: view by pattern

Pattern)	Meter	Positions
fa' (L)	*al-Mutaqārib.P*	3, 6
	al-Mutaqārib.C	4, 8

Pattern)	Meter	Positions
fi'lun (LL)	*al-Madīd.C*	6
	al-Madīd.P	4
	al-Basīt.C	4, 8
	al-Kāmil.C	6
	al-Sarī'.C	3, 6
	al-Mutadārak.C	1, 2, 3, 4, 5, 6, 7, 8
	al-Mutadārak.P1	1, 2, 3, 4, 5, 6
	al-Mutadārak.P2	1, 2, 3, 4

Pattern)	Meter	Positions
maf'ūlun (LLL)	*al-Basīt.P1*	6
	al-Basīt.P2	2, 3, 5, 6
	al-Kāmil.C	6
	al-Rajaz.C	3, 6
	al-Rajaz.P1	2, 4
	al-Rajaz.P2	3
	al-Rajaz.P3	2
	al-Munsarih.C	6
	al-Khafīf.C	3, 6
	al-Mujtath.C	4

Pattern)	Meter	Positions
maf'ūlān (LLLA)	*al-Rajaz.C*	3, 6
	al-Rajaz.P1	2, 4
	al-Rajaz.P2	3

Pattern)	Meter	Positions
maf'ūlātu (LLLS)	*al-Munsarih.C*	2, 5
	al-Muqtadab.C	1, 3

Pattern)	Meter	Positions
maf'ūlu (LLS)	*al-Basīt.P2*	3
	al-Rajaz.C	3
	al-Rajaz.P1	2
	al-Khafīf.C	3

Pattern)	Meter	Positions
mustaf'ilun (LLSL)	*al-Basīt.C*	1, 3, 5, 7
	al-Basīt.P1	1, 3, 4, 6
	al-Basīt.P2	1, 4
	al-Kāmil.P	1, 2, 3
	al-Kāmil.C	1, 2, 3, 4, 5, 6
	al-Rajaz.C	1, 2, 3, 4, 5, 6
	al-Rajaz.P1	1, 2, 3, 4
	al-Rajaz.P2	1, 2, 3
	al-Rajaz.P3	1, 2
	al-Sarī'.C	1, 2, 4, 5
	al-Munsarih.C	1, 3, 4
	al-Khafīf.C	2, 5
	al-Khafīf.P	2, 4
	al-Mujtath.C	1, 3

Pattern)	Meter	Positions
mustaf'ilān (LLSLA)	*al-Basīt.P1*	3, 6

Pattern)	Meter	Positions
mustaf'ilātun (LLSLL)	*al-Kāmil.P*	2, 4

Pattern)	Meter	Positions
mustaf'ilātu (LLSLS)	*al-Kāmil.P*	2

Pattern)	Meter	Positions
mustaf'ilu (LLSS)	*al-Basīt.P1*	3
	al-Kāmil.P	2
	al-Kāmil.C	3
	al-Rajaz.C	3
	al-Rajaz.P1	2
	al-Munsarih.C	3
	al-Khafīf.P	2

Pattern)	Meter	Positions
fi'lu (LS)	*al-Basīt.C*	4
	al-Sarī'.C	3
	al-Mutadārak.C	4
	al-Mutadārak.P1	3
	al-Mutadārak.P2	2

Pattern)	Meter	Positions
fā'ilun (LSL)	*al-Madīd.C*	2, 3, 5, 6
	al-Madīd.P	2
	al-Basīt.C	2, 6
	al-Basīt.P1	2, 5
	al-Basīt.P2	2, 5
	al-Ramal.P	2, 4
	al-Ramal.C	3, 6
	al-Sarī'.C	3, 6
	al-Khafīf.C	3, 6
	al-Mutadārak.C	1, 2, 3, 4, 5, 6, 7, 8
	al-Mutadārak.P1	1, 2, 3, 4, 5, 6
	al-Mutadārak.P2	1, 2, 3, 4

Pattern)	Meter	Positions
fā'ilān (LSLA)	*al-Madīd.C*	6
	al-Ramal.P	2, 4
	al-Ramal.C	3, 6
	al-Sarī'.C	3, 6
	al-Mutadārak.P1	3, 6
	al-Mutadārak.P2	2, 4

Pattern)	Meter	Positions
fā'ilātun (LSLL)	*al-Madīd.C*	1, 3, 4, 6
	al-Madīd.P	1, 3
	al-Ramal.P	1, 2, 3, 4
	al-Ramal.C	1, 2, 3, 4, 5, 6
	al-Khafīf.C	1, 3, 4, 6
	al-Khafīf.P	1, 3
	al-Mudāri'.C	2, 4
	al-Mujtath.C	2, 4

Pattern)	Meter	Positions
fā'ilātān (LSLLA)	*al-Ramal.P*	2, 4
	al-Ramal.C	3, 6

Pattern)	Meter	Positions
fā'ilātu (LSLS)	*al-Madīd.C*	3
	al-Ramal.P	2
	al-Ramal.C	3
	al-Munsarih.C	2, 5
	al-Khafīf.C	3
	al-Mudāri'.C	2
	al-Muqtadab.C	1, 3
	al-Mujtath.C	2

Pattern)	Meter	Positions
fā'ilu (LSS)	*al-Madīd.C*	3
	al-Madīd.P	2
	al-Ramal.P	2
	al-Ramal.C	3
	al-Sarī'.C	3
	al-Khafīf.C	3
	al-Mutadārak.C	4
	al-Mutadārak.P1	3
	al-Mutadārak.P2	2

Pattern)	Meter	Positions
mufta'ilun (LSSL)	*al-Basīt.C*	1, 3, 5, 7
	al-Basīt.P1	1, 3, 4
	al-Basīt.P2	1, 4
	al-Rajaz.C	1, 2, 3, 4, 5, 6
	al-Rajaz.P1	1, 2, 3, 4
	al-Rajaz.P2	1, 2, 3
	al-Rajaz.P3	1, 2
	al-Sarī'.C	1, 2, 4, 5
	al-Munsarih.C	1, 3, 4, 6
	al-Muqtadab.C	2, 4

Pattern)	Meter	Positions
mufta'ilu (LSSS)	*al-Basīt.P1*	3
	al-Rajaz.C	3
	al-Rajaz.P1	2
	al-Munsarih.C	3
	al-Muqtadab.C	2

Pattern)	Meter	Positions
fa'al (SL)	*al-Mutaqārib.P*	3, 6
	al-Mutaqārib.C	4, 8

Pattern)	Meter	Positions
fa'ūl (SLA)	*al-Mutaqārib.P*	3, 6
	al-Mutaqārib.C	4, 8

Pattern)	Meter	Positions
fa'ūlun (SLL)	*al-Tawīl.C*	1, 3, 4, 5, 7, 8
	al-Basīt.P1	6
	al-Basīt.P2	3, 6
	al-Wāfir.C	3, 6
	al-Hazaj.C	2, 4
	al-Rajaz.C	3, 6
	al-Rajaz.P1	2, 4
	al-Rajaz.P2	3
	al-Rajaz.P3	2
	al-Khafīf.P	4
	al-Mutaqārib.P	1, 2, 3, 4, 5, 6
	al-Mutaqārib.C	1, 2, 3, 4, 5, 6, 7, 8

Pattern)	Meter	Positions
fa'ūlān (SLLA)	*al-Rajaz.C*	3, 6
	al-Rajaz.P1	2, 4
	al-Rajaz.P2	3

Pattern)	Meter	Positions
mafā'īlun (SLLL)	*al-Tawīl.C*	2, 4, 6, 8
	al-Wāfir.C	1, 2, 4, 5
	al-Wāfir.P	1, 2, 3, 4
	al-Hazaj.C	1, 2, 3, 4
	al-Mudāri'.C	1, 3

Pattern)	Meter	Positions
mafā'īlu (SLLS)	*al-Tawīl.C*	4
	al-Wāfir.P	2
	al-Hazaj.C	1, 2, 3
	al-Munsarih.C	2, 5
	al-Mudāri'.C	1, 3
	al-Muqtadab.C	1, 3

Pattern)	Meter	Positions
fa'ūlu (SLS)	*al-Tawīl.C*	1, 3, 4, 5, 7
	al-Basīt.P2	3
	al-Wāfir.C	3
	al-Hazaj.C	2
	al-Rajaz.C	3
	al-Rajaz.P1	2
	al-Mutaqārib.P	1, 2, 3, 4, 5
	al-Mutaqārib.C	1, 2, 3, 4, 5, 6, 7

Pattern)	**Meter**	**Positions**
mafā'ilun (SLSL)	*al-Tawīl.C*	2, 4, 6, 8
	al-Basīt.C	1, 3, 5, 7
	al-Basīt.P1	1, 3, 4
	al-Basīt.P2	1, 4
	al-Hazaj.C	1, 2, 3, 4
	al-Rajaz.C	1, 2, 3, 4, 5, 6
	al-Rajaz.P1	1, 2, 3, 4
	al-Rajaz.P2	1, 2, 3
	al-Rajaz.P3	1, 2
	al-Sarī'.C	1, 2, 4, 5
	al-Munsarih.C	1, 4
	al-Khafīf.C	2, 5
	al-Khafīf.P	2, 4
	al-Mujtath.C	1, 3

Pattern)	**Meter**	**Positions**
mafā'ilu (SLSS)	*al-Tawīl.C*	4
	al-Basīt.P1	3
	al-Hazaj.C	1, 2, 3
	al-Rajaz.C	3
	al-Rajaz.P1	2
	al-Khafīf.P	2

Pattern)	**Meter**	**Positions**
mufā'alatun (SLSSL)	*al-Wāfir.C*	1, 2, 4, 5
	al-Wāfir.P	1, 2, 3, 4

Pattern)	**Meter**	**Positions**
mufā'alatu (SLSSS)	*al-Wāfir.P*	2

Pattern)	Meter	Positions
fa'ilun (SSL)	*al-Madīd.C*	2, 3, 5, 6
	al-Madīd.P	2
	al-Basīt.C	2, 4, 6, 8
	al-Basīt.P1	2, 5
	al-Basīt.P2	2, 5
	al-Kāmil.C	3, 6
	al-Ramal.P	2, 4
	al-Ramal.C	3, 6
	al-Sarī'.C	3, 6
	al-Mutadārak.C	1, 2, 3, 4, 5, 6, 7, 8
	al-Mutadārak.P1	1, 2, 3, 4, 5, 6
	al-Mutadārak.P2	1, 2, 3, 4

Pattern)	Meter	Positions
fa'ilān (SSLA)	*al-Ramal.P*	2, 4
	al-Ramal.C	3, 6

Pattern)	Meter	Positions
fa'ilātun (SSLL)	*al-Madīd.C*	1, 3, 4, 6
	al-Madīd.P	1, 3
	al-Kāmil.C	3, 6
	al-Ramal.P	1, 2, 3, 4
	al-Ramal.C	1, 2, 3, 4, 5, 6
	al-Khafīf.C	1, 3, 4, 6
	al-Khafīf.P	1, 3
	al-Mujtath.C	2, 4

Pattern)	Meter	Positions
fa'ilātān (SSLLA)	*al-Ramal.P*	2, 4
	al-Ramal.C	3, 6

Pattern)	Meter	Positions
fa'ilātu (SSLS)	*al-Madīd.C*	3
	al-Kāmil.C	3
	al-Ramal.P	2
	al-Ramal.C	3
	al-Khafīf.C	3
	al-Mujtath.C	2

Pattern)	Meter	Positions
mutafā'ilun (SSLSL)	*al-Kāmil.P*	1, 2, 3, 4
	al-Kāmil.C	1, 2, 3, 4, 5, 6

Pattern)	Meter	Positions
mutafā'ilān (SSLSLA)	*al-Kāmil.P*	4

Pattern)	Meter	Positions
mutafā'ilātun (SSLSLL)	*al-Kāmil.P*	2, 4

Pattern)	Meter	Positions
mutafā'ilātu (SSLSLS)	*al-Kāmil.P*	2

Pattern)	Meter	Positions
mutafā'ilu (SSLSS)	*al-Kāmil.P*	2
	al-Kāmil.C	3

Pattern)	Meter	Positions
fa'ilu (SSS)	*al-Madīd.C*	3
	al-Madīd.P	2
	al-Basīt.C	4
	al-Kāmil.C	3
	al-Ramal.P	2
	al-Ramal.C	3
	al-Sarī'.C	3
	al-Mutadārak.C	4
	al-Mutadārak.P1	3
	al-Mutadārak.P2	2

Pattern)	Meter	Positions
fa'ilatun (SSSL)	*al-Rajaz.C*	1, 2, 4, 5
	al-Rajaz.P1	1, 3
	al-Rajaz.P2	1, 2
	al-Rajaz.P3	1
	al-Sarī'.C	1, 2, 4, 5

Appendix D

Visualization samples

This appendix presents visualizations of poems that are most representative of their meters. Each illustration includes an upper section showing a visualization where each pattern is assigned a unique color, and a lower section where patterns share the same color based on syllabic quantity. Each image is labeled with the poet's name, the poem's title, and its meter.

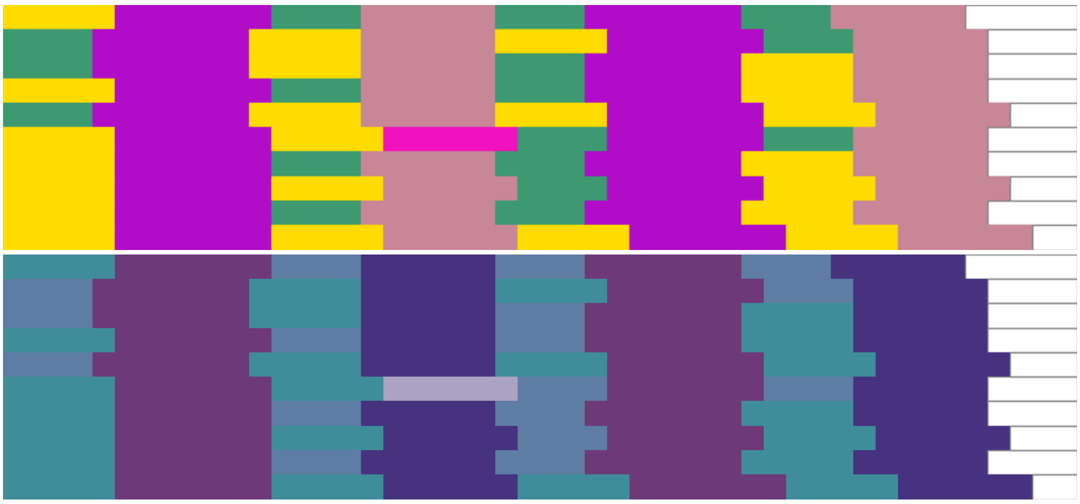


FIGURE D.1: *Abū al-Fadl al-Kināni̅, idhā mā waradnā mā'a Madyan, al-Tawīl.C*

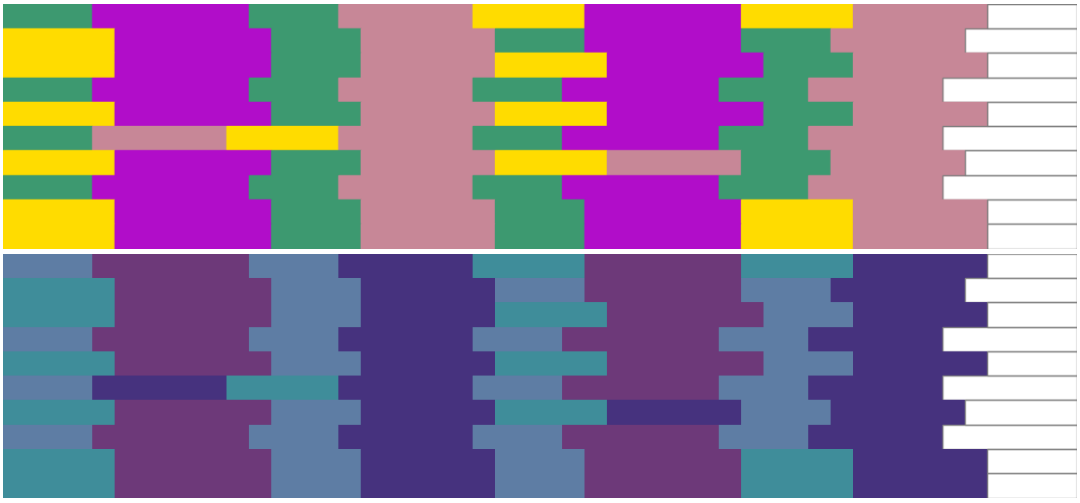


Figure D.2: *Imru' al-Qays, liman talalun bayna al-judayyati wa al-jabal, al-Tawīl.C*

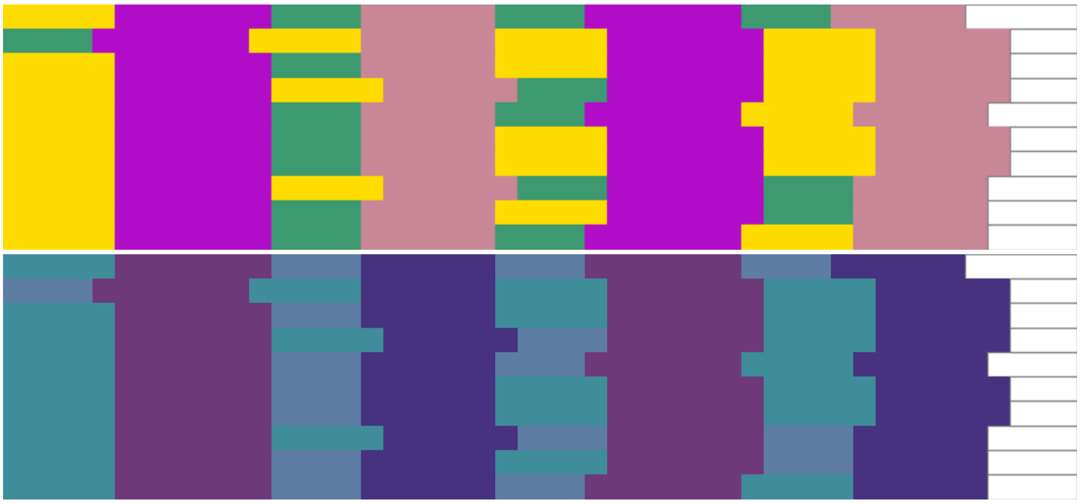


Figure D.3: *Qays ibn al-Mulawwah, alā fas'ali al-rukkbāna, al-Tawīl.C*

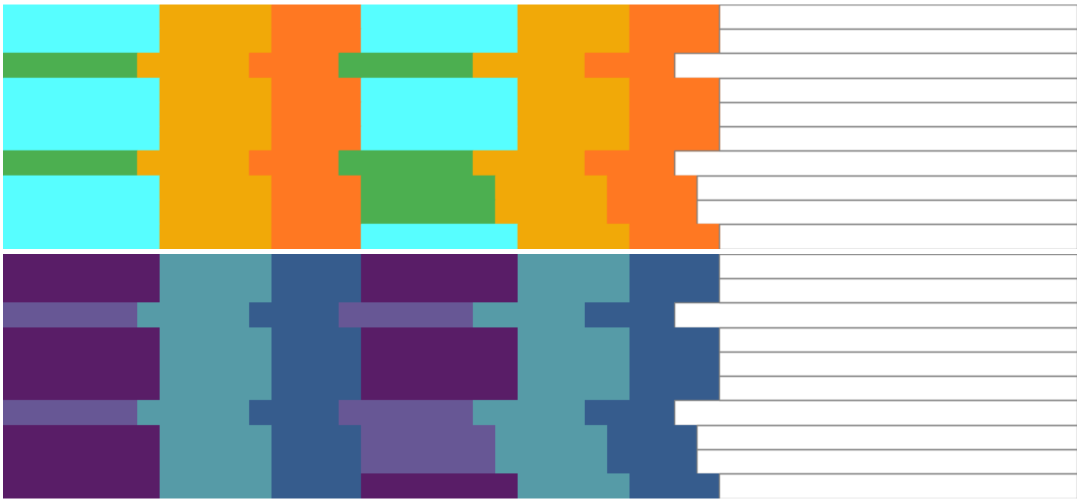


FIGURE D.4: *Ahmad Shawqī, bayna sam'i Allāhi wa al-basari, al-Madīd.C*

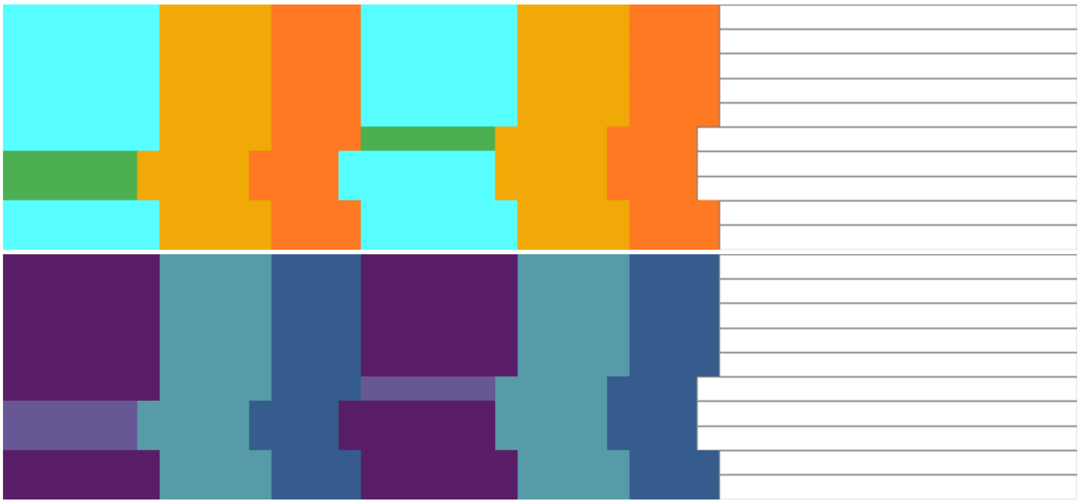


FIGURE D.5: *Hāfiz Ibrāhīm, mā li-hādhā al-najmi fī al-sahari, al-Madīd.C*

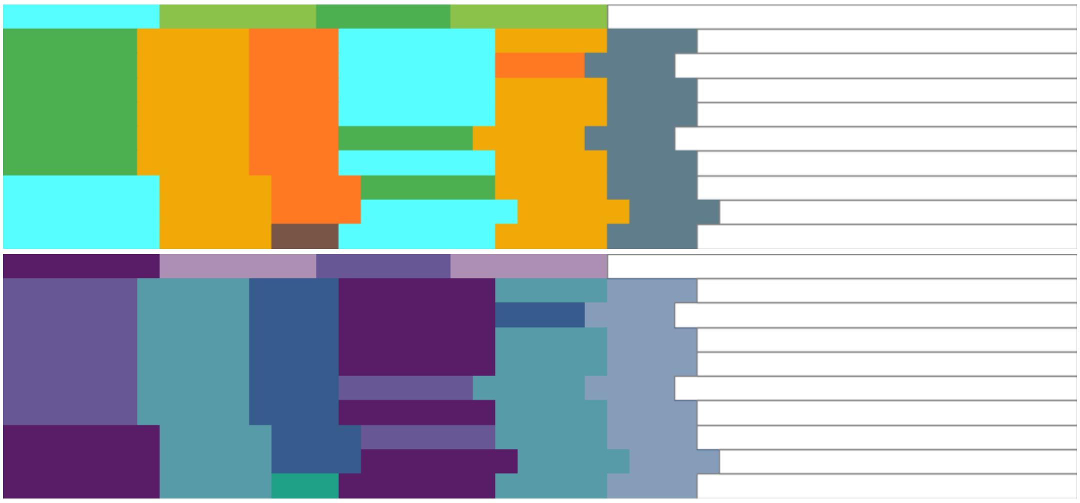


FIGURE D.6: *Ibn al-Mu'tazz, jāra hādhā ad-dahru aw ābā, al-Madīd.C*

FIGURE D.7: *Ta'abbata sharran, inna bi-sh-sha'bi alladhī dūna sal'in, al-Madīd.C*

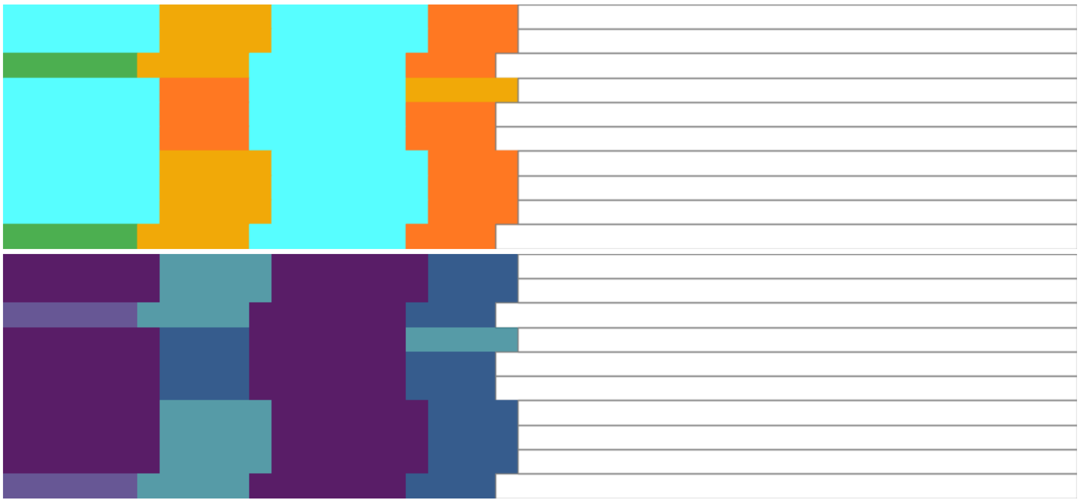


FIGURE D.8: *As-Sulakah Ummu as-Sulayk, tāfa yabghī najwatan, al-Madīd.P*

FIGURE D.9: *Ahmad Rāmī, ilā Ummu Kulthūm, al-Basīt.C*

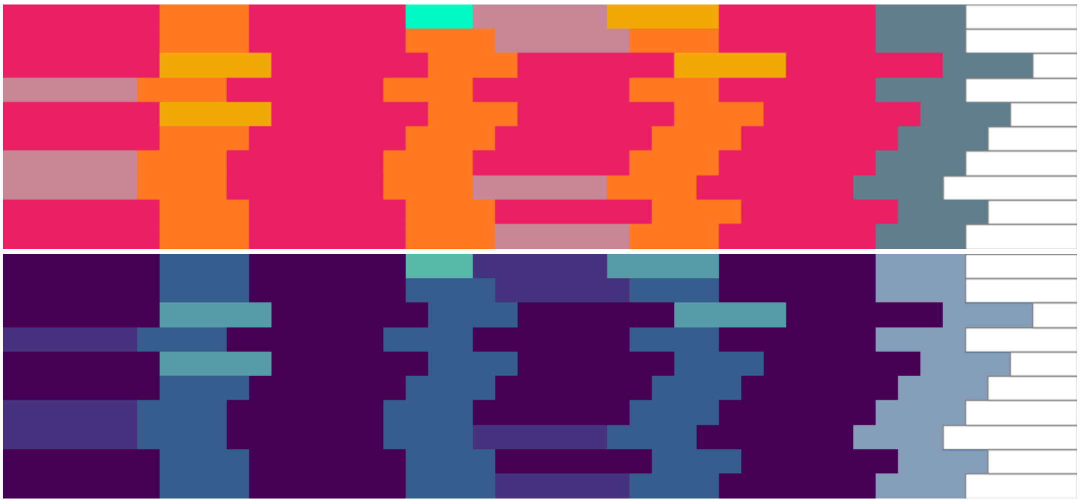


FIGURE D.10: *al-Mutanabbī, 'idun bi-ayyati hālin 'udta yā 'idu, al-Basīt.C*

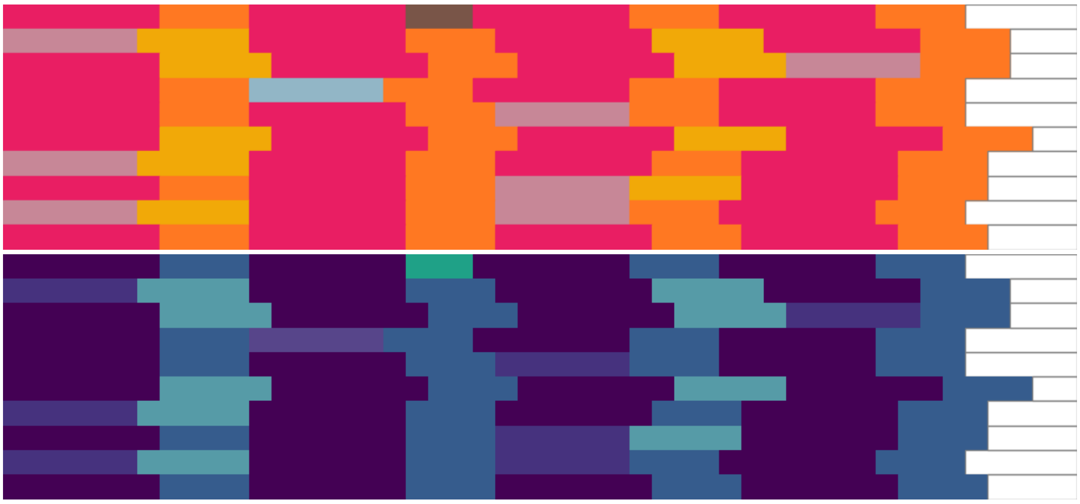


FIGURE D.11: *Al-Nābighah al-Dhubyānī, al-Mu'allaqa, al-Basīt.C*

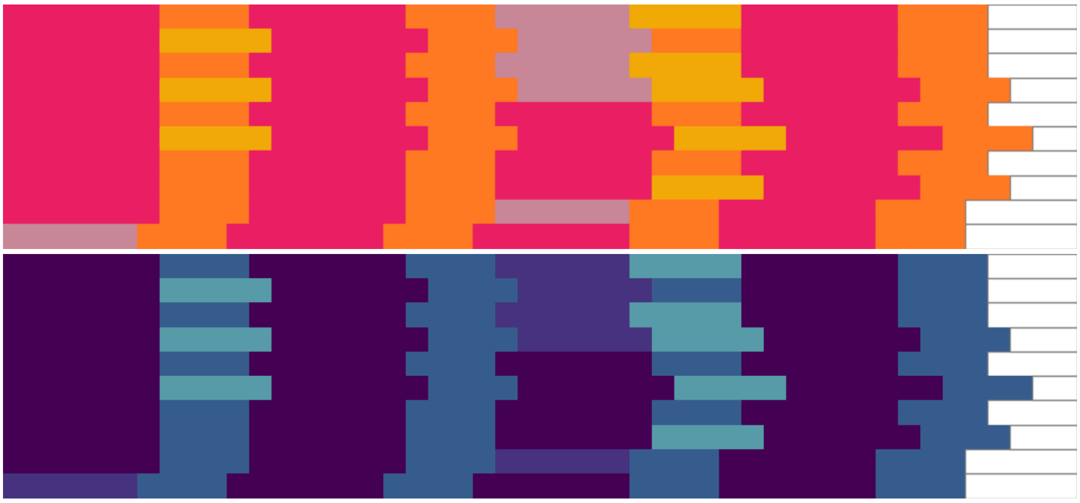


Figure D.12: *Zuhayr ibn Abī Sulmā, bāna al-khalītu wa lam ya'wū liman tarakū, al-Basīt.C*

Figure D.13: *Abū Firās al-Hamadhānī, al-lawmu lil-'āshiqīna lūmu, al-Basīt.P2*

Figure D.14: *Ibn al-Rūmī, ghusnun mina al-bāni fī wishāhin, al-Basīt.P2*

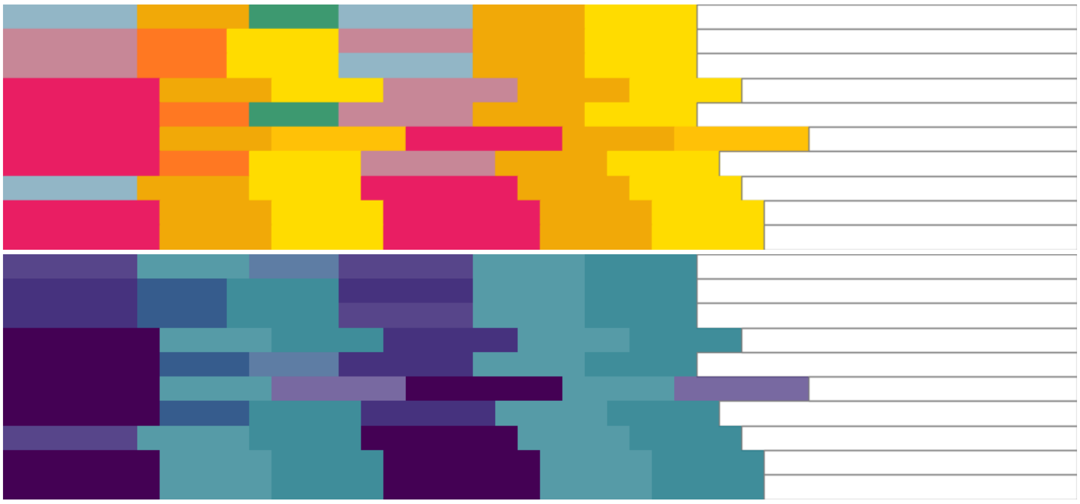


Figure D.15: *'Ubayd ibn al-Abras, al-Mu'allaqa, al-Basīt.P2*

Figure D.16: *'Amr ibn Kulthūm, al-Mu'allaqa, al-Wāfir.C*

Figure D.17: *Bashār ibn Burd, tajahhaza tāla fī an-nasabi ath-thawā'u, al-Wāfir.C*

Figure D.18: *Khalīl Matrān, 'alāma yuqāmu timthālī, al-Wāfir.P*

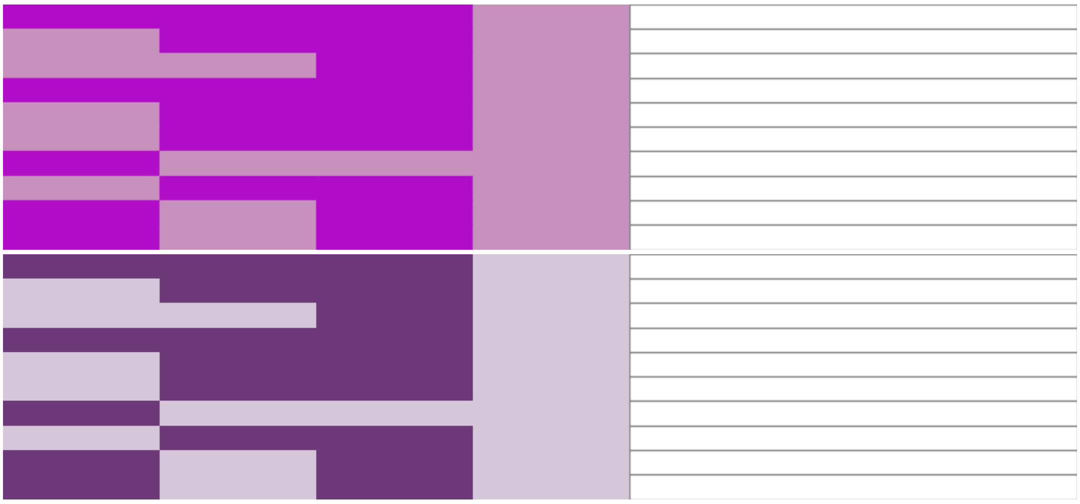


Figure D.19: *'Umar ibn Abī Rabī'a, taribta warda man tahwā, al-Wāfir.P*

Figure D.20: *Abī al-'Atāhiya, inna al-fanā'a mina al-baqā'i qarību, al-Kāmil.C*

Figure D.21: *'Antarah ibn Shaddād, yā 'Ablu khallī 'anki qawla al-muftarī, al-Kāmil.C*

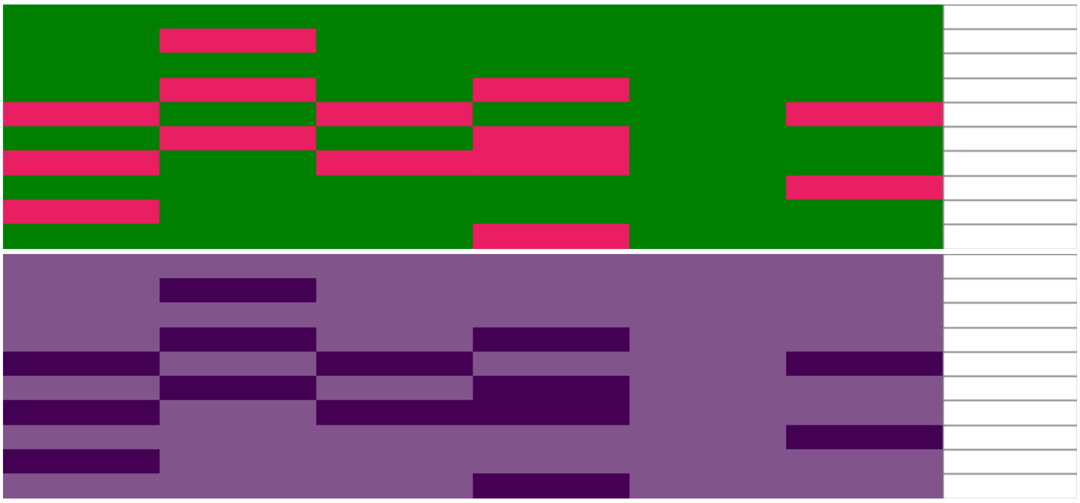


Figure D.22: *Labīd ibn Rabī'a, al-Mu'allaqa, al-Kāmil.C*

Figure D.23: *Hāfiz Ibrāhīm, ahlan bi-awwali muslimin, al-Kāmil.P*

FIGURE D.24: *Mustafā Sādiq al-Rāfi'ī, qāsūki yā shamsu al-duhā, al-Kāmil.P*

FIGURE D.25: *Ahmad Shawqī, 'afīfu al-jahri wa al-hamsi, al-Hazaj.C*

Figure D.26: *Badī' az-Zamān al-Hamadhānī, la'in ahrazaka ad-dā'ī, al-Hazaj.C*

Figure D.27: *Umaymah bint Umayyah, abā Layliyya an yad-hhab, al-Hazaj.C*

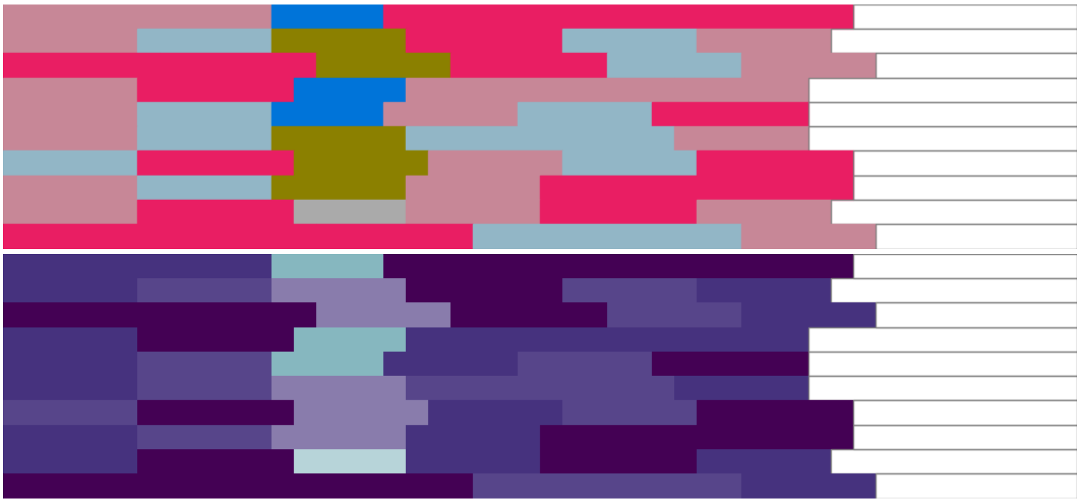


FIGURE D.28: *al-Mutanabbī, wa shāmikhin mina al-jibāli aqwadi, al-Rajaz.C*

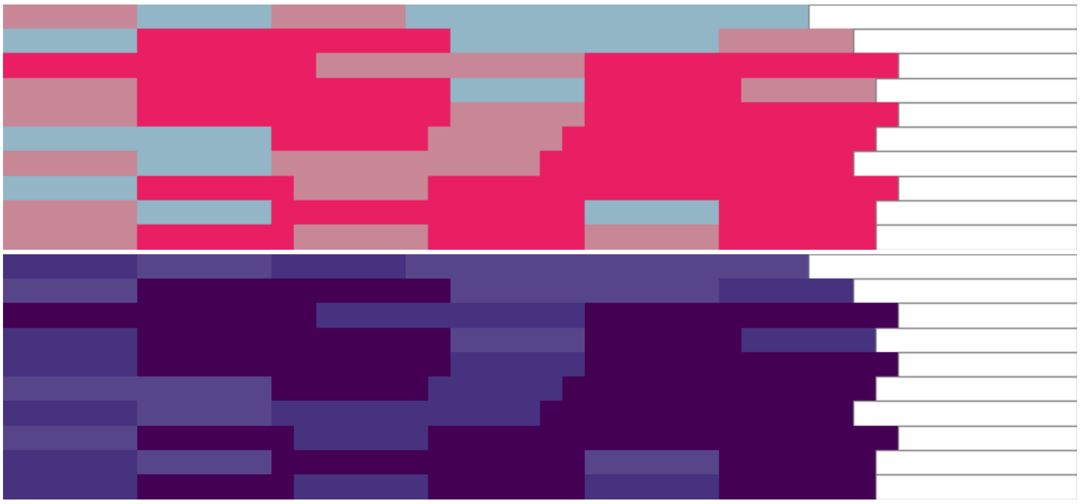


FIGURE D.29: *Ibn al-Rūmī, laqad ra'aynā 'ajaban mina al-'ajab, al-Rajaz.C*

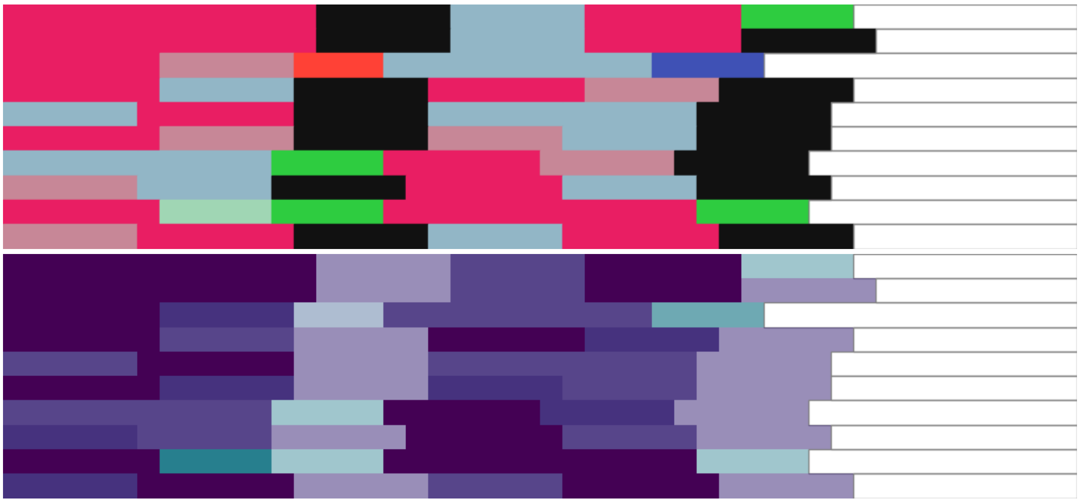


FIGURE D.30: *Jarīr, innī imru'un yabnī liya al-majda al-bān, al-Rajaz.C*

FIGURE D.31: *Ibn Zaydūn, yā dam'u sub mā shi'ta an tasūbā, al-Rajaz.C*

Figure D.32: *Ahmad Shawqī, fashawdatun riwāyatun, al-Rajaz.P1*

Figure D.33: *Ibn Sinā' al-Mulk, laqad laqītu nasaban, al-Rajaz.P1*

FIGURE D.34: *Al-Akhtal, qad gharrahum minnī la'īm jambah, al-Rajaz.P2*

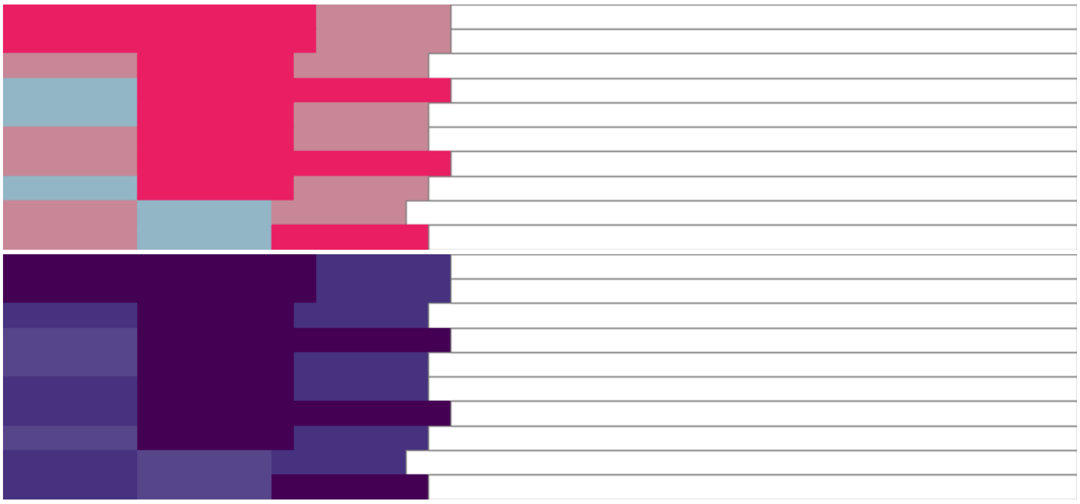


FIGURE D.35: *al-Mutanabbī, mā lil-murūji al-khudri wa al-hadā'iqi, al-Rajaz.P2*

Figure D.36: *Abū Nuwās, ilāhanā, al-Rajaz.P3*

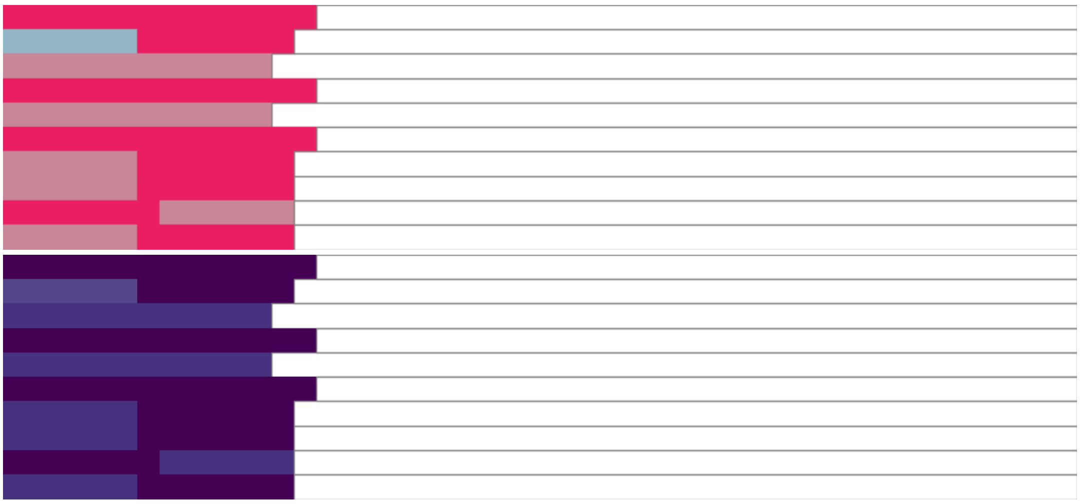


Figure D.37: *Ibn Nabaatah Al Misri, afdī qamar 'aqlī qamar, al-Rajaz.P3*

FIGURE D.38: *Ahmad Shawqī, akhadhat na'shaki Misru bi al-yamīni, al-Ramal.C*

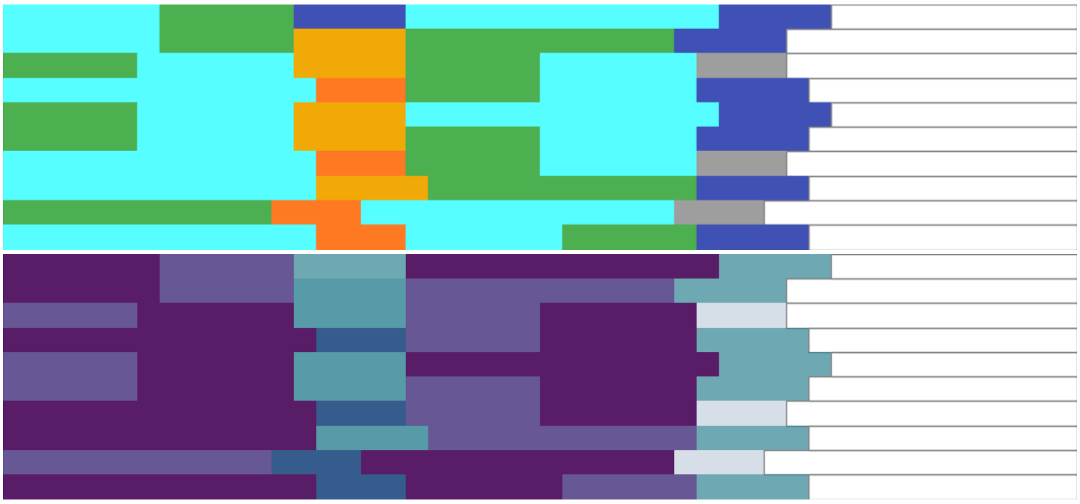


FIGURE D.39: *Jubrān Khalīl Jubrān, rabbata ad-dawlati wa al-jāhi al-makīn, al-Ramal.C*

FIGURE D.40: *Maḥmūd Sāmī al-Bārūdī, bādir al-fursata wa ihd-har fawtahā, al-Ramal.C*

FIGURE D.41: *Mu'īn Basīsū, al-Mīlād, al-Ramal.C*

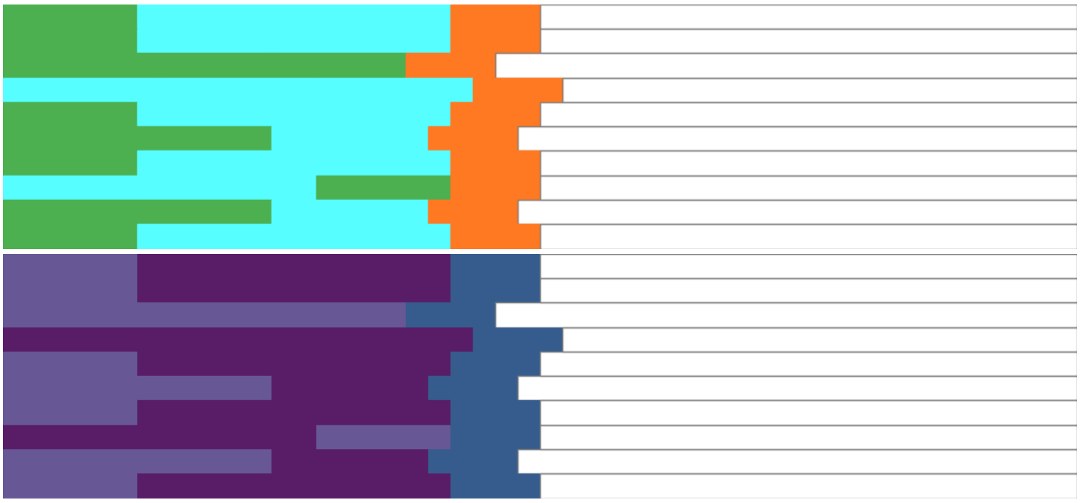


Figure D.42: *al-Khansā, marihat 'aynī, al-Ramal.P*

Figure D.43: *Bashār ibn Burd, tāla fī Hindin 'itābī, al-Ramal.P*

FIGURE D.44: *Abū al-'Alae al-Ma'arrī, ijtanib an-nāsa wa 'ish wāḥidan, al-Sarī'.C*

FIGURE D.45: *Ahmad Rāmī, rubā'iyāt al-Khayyām, al-Sarī'.C*

FIGURE D.46: *Ibn Al Mu'taz, lā 'udhra lil-'ādhili fī al-ka'si, al-Sarī'.C*

FIGURE D.47: *Imru' al-Qays, yā dāra Māwiyata bi al-hā'il, al-Sarī'.C*

Figure D.48: *Abū Firās al-Hamadhānī, yā hasratan mā akādu ahmiluhā, al-Munsarih.C*

Figure D.49: *al-Mutanabbī, ahaqqun 'āfin bi-dam'ika al-himamu, al-Munsarih.C*

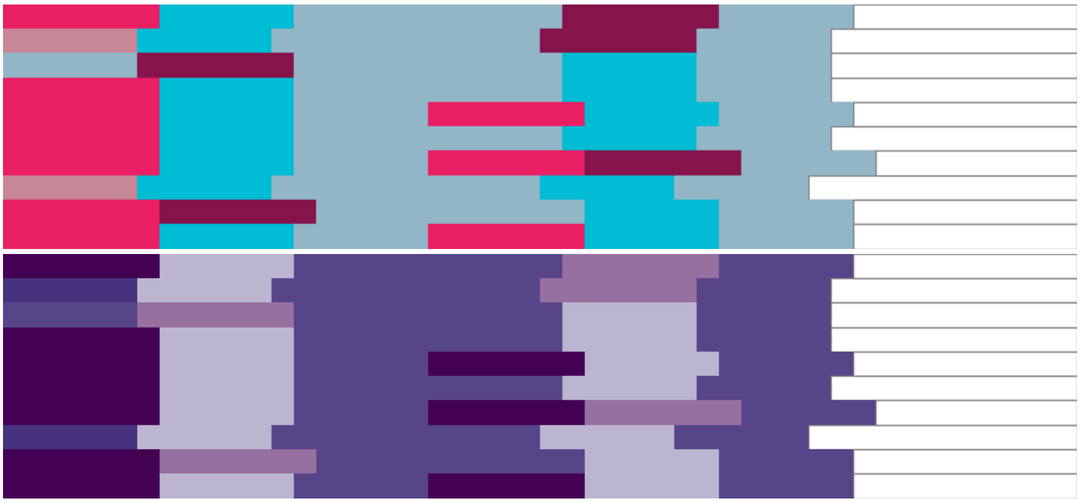


FIGURE D.50: *Hassān ibn Thābit, unzur khalīlī, al-Munsarih.C*

FIGURE D.51: *Abū al-Qāsim al-Shābbī, yā rafīqī, al-Khafīf.C*

FIGURE D.52: *Al-Hārith ibn Hillizah, al-Mu'allaqa, al-Khafīf.C*

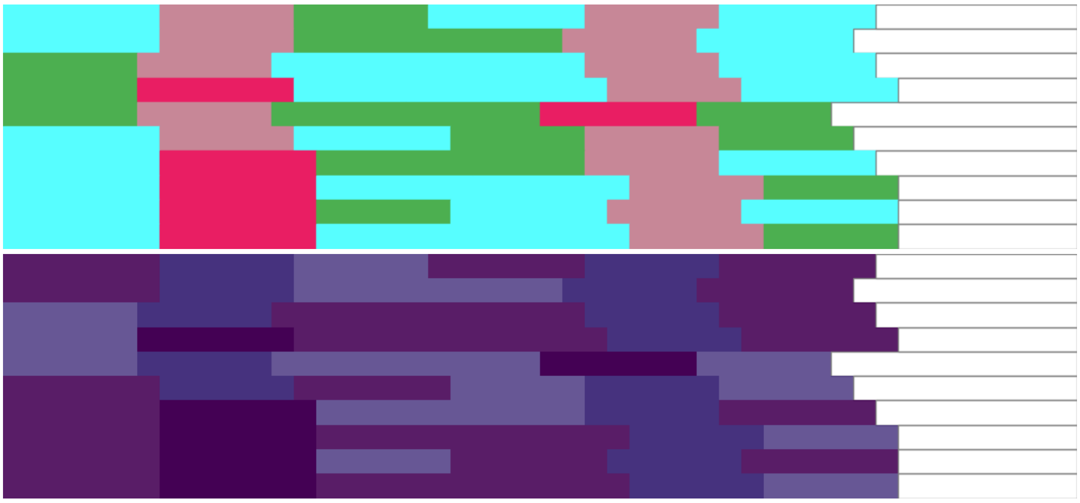


FIGURE D.53: *al-Manfalūtī, ayyuhā al-fātiku al-athīmu ruwaydan, al-Khafīf.C*

Figure D.54: *Ibrāhīm Nāgī, qultu lil-bahri idh waqaftu masā'an, al-Khafīf.C*

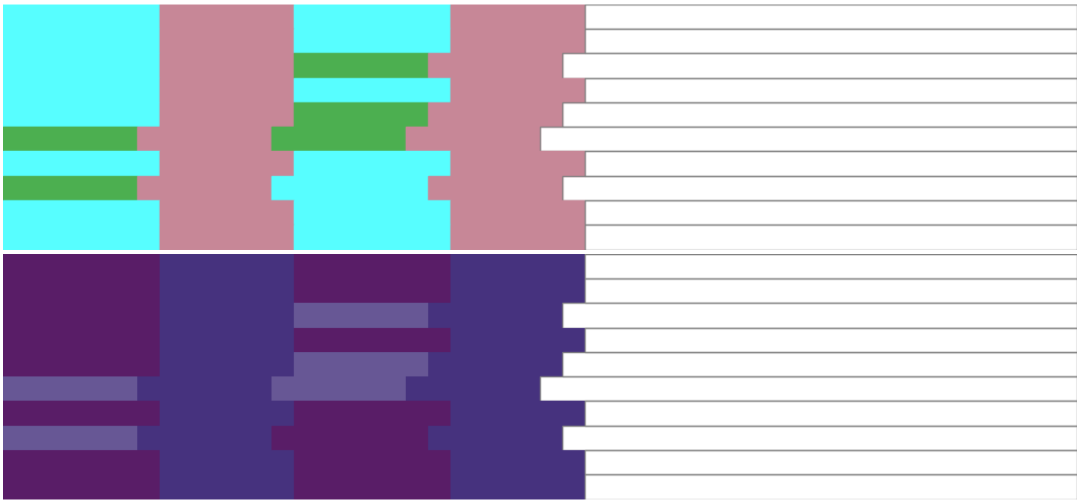


Figure D.55: *Ahmad Shawqī, yā ibn Zaydūn marhabā, al-Khafīf.P*

Figure D.56: *al-Sharīf al-Murtaḍā, ilā kam aqūdu qawman, al-Muḍāri'.C*

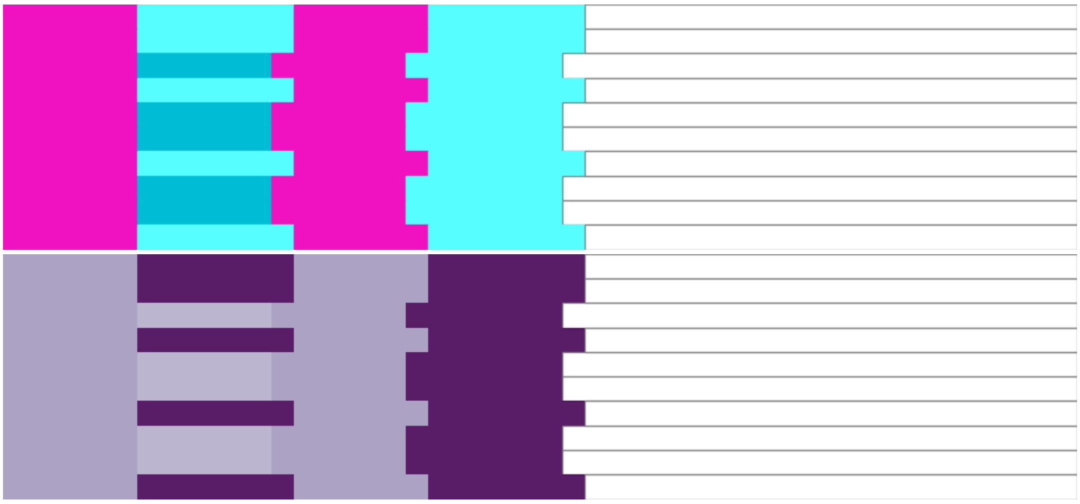


Figure D.57: *al-Hays Bays, wa bil-qasri aryīhiyun, al-Muḍāri'.C*

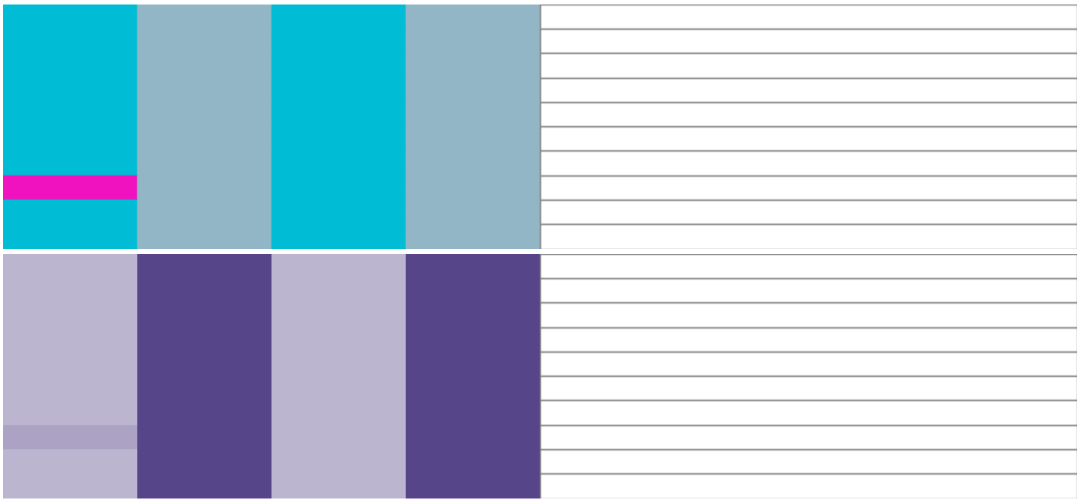


FIGURE D.58: *al-Shādhlī Khaznadār, rāhatu al-nuhā al-tarabu, al-Muqtadab.C*

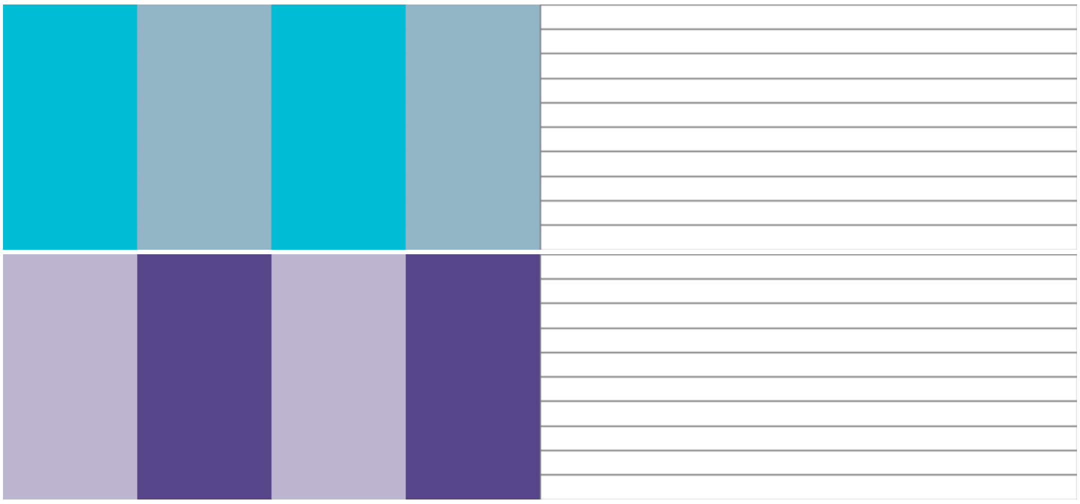


FIGURE D.59: *Ilyās Abū Shabaka, lā yashuqunā 'arabun, al-Muqtadab.C*

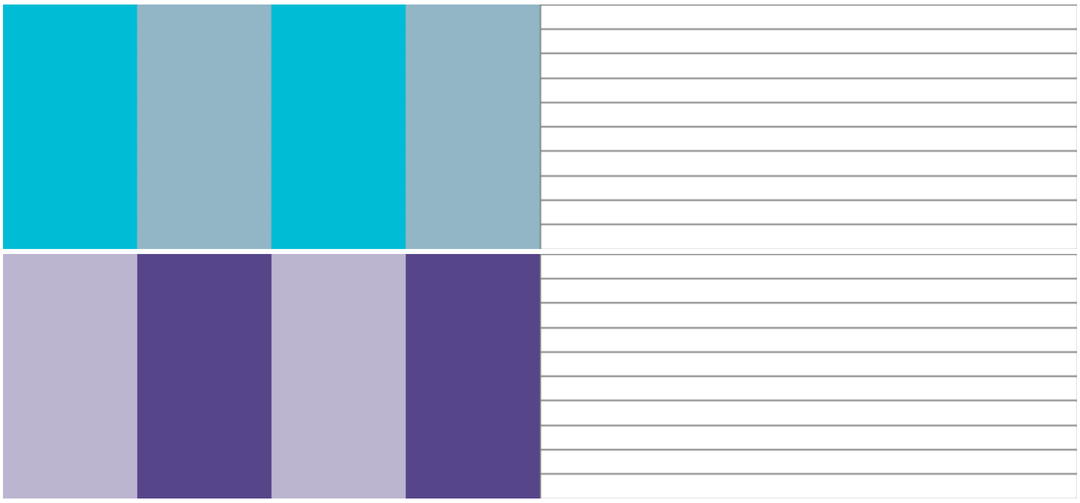


FIGURE D.60: *Khalīl Matrān, al-qulūbu wa al-muqalu, al-Muqtadab.C*

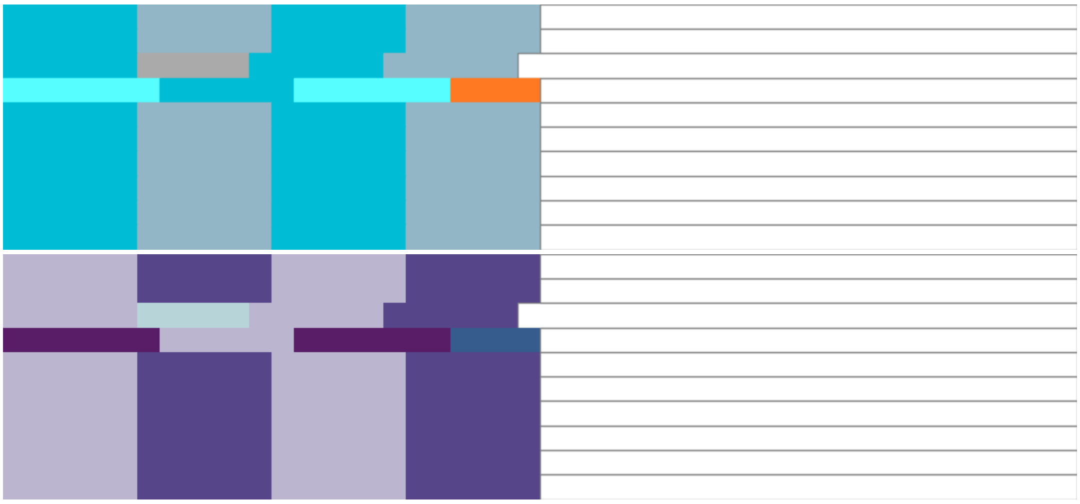


FIGURE D.61: *Safiy al-Dīn al-Hillī, ayna fī al-himā 'Arabun, al-Muqtadab.C*

Figure D.62: *Abū al-Qāsim al-Shābbī, yā labtisāmati qalbin, al-Mujtath.C*

Figure D.63: *Abū Nuwās, 'azzū akhillāya qalbī, al-Mujtath.C*

FIGURE D.64: *Khalīl Matrān, hakū lanā 'an himārin, al-Mujtath.C*

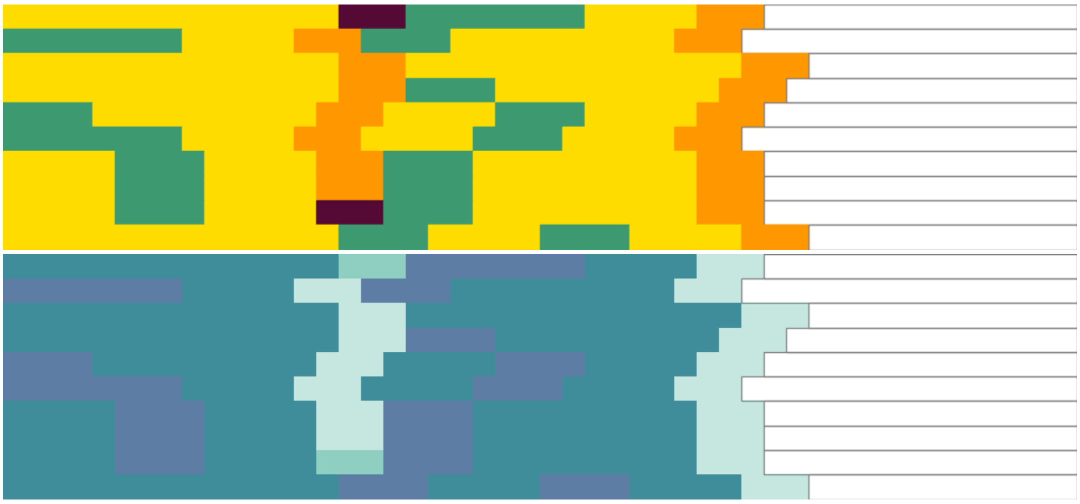


FIGURE D.65: *Aws ibn Hajr, alam tuksafi al-shamsu wa al-badru, al-Mutaqārib.C*

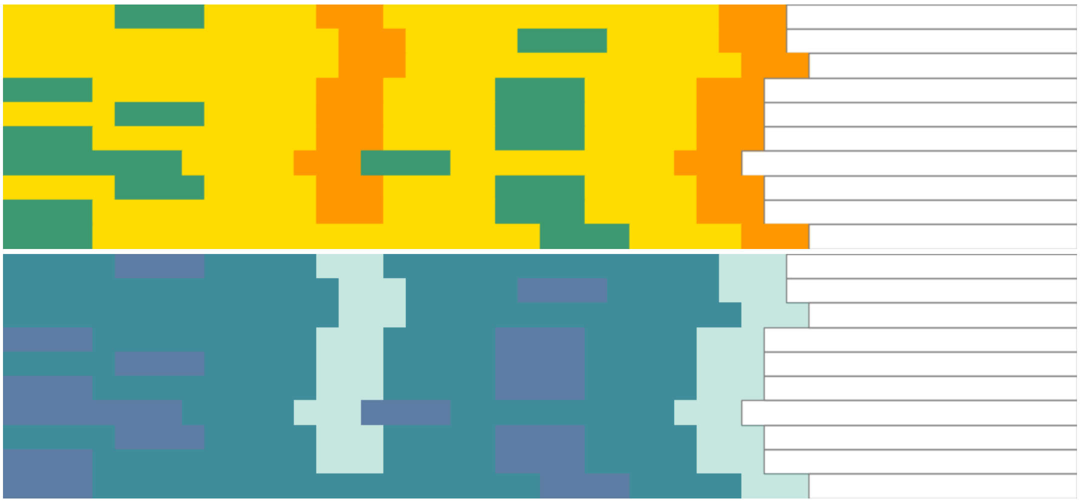


FIGURE D.66: *Ḥājib ibn Ḥabīb, wa bātat talūmu 'alā thādiqin, al-Mutaqārib.C*

FIGURE D.67: *Ibn Sinā' al-Mulk, a yaddfa'unī ad-dahru 'an matlabī, al-Mutaqārib.C*

Figure D.68: *Jarīr, ba'anna al-khalīta ghadāta al-jinābi, al-Mutaqārib.C*

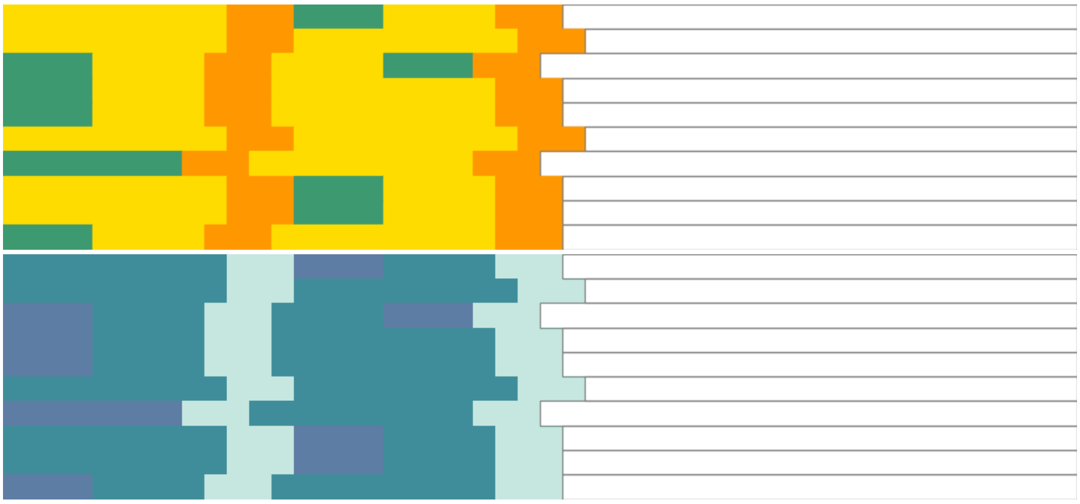


Figure D.69: *Abū al-'Alae al-Ma'arrī, qifī waqfatan ta'lamī, al-Mutaqārib.P*

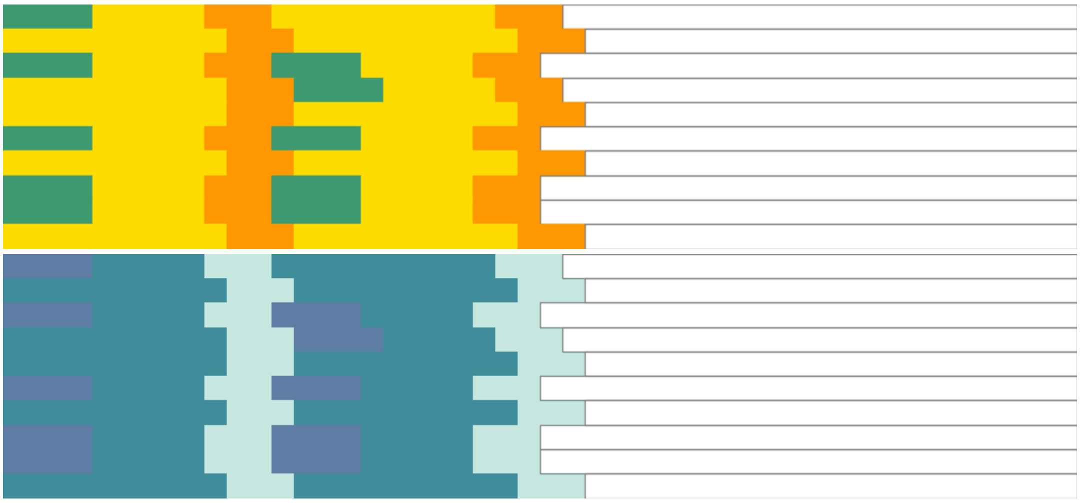


FIGURE D.70: *Abū Firās al-Hamadhānī, li-'ayyikum adhkuru, al-Mutaqārib.P*

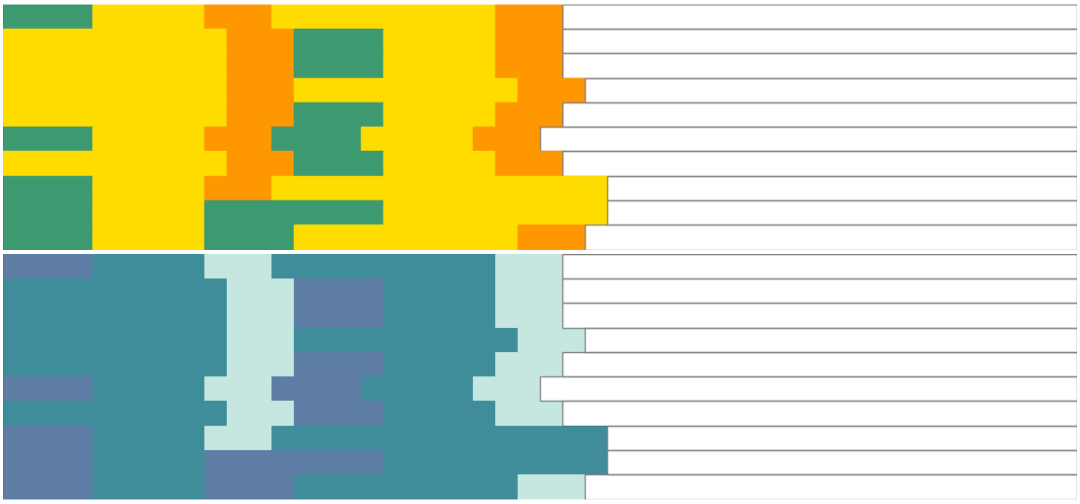


FIGURE D.71: *Safiy al-Dīn al-Hillī, shakawtu ilayki al-jawā, al-Mutaqārib.P*

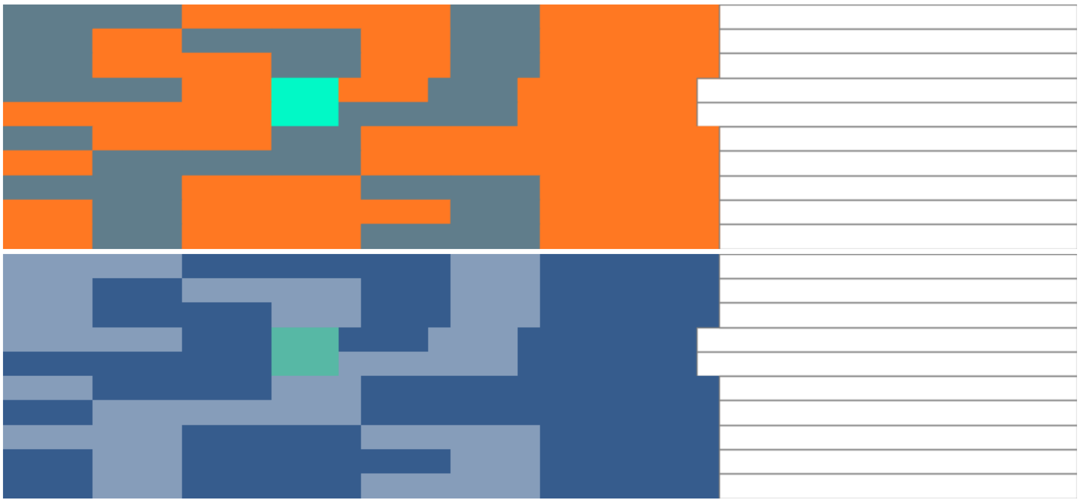


Figure D.72: *Abū al-Qāsim al-Shābbī, ghannāhu al-'amsu, al-Mutadārak.C*

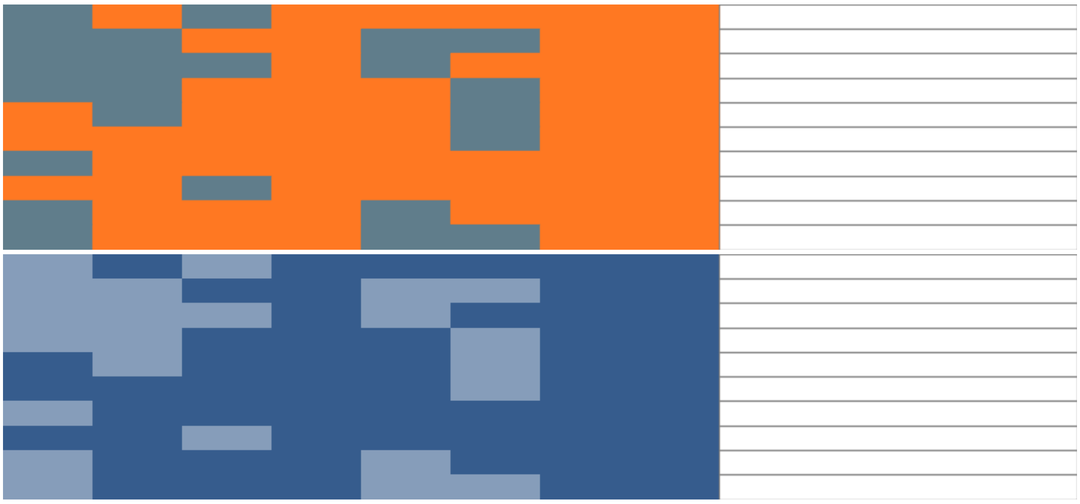


Figure D.73: *Ahmad Shawqī, mudnāka jafāhu marqaduhu, al-Mutadārak.C*

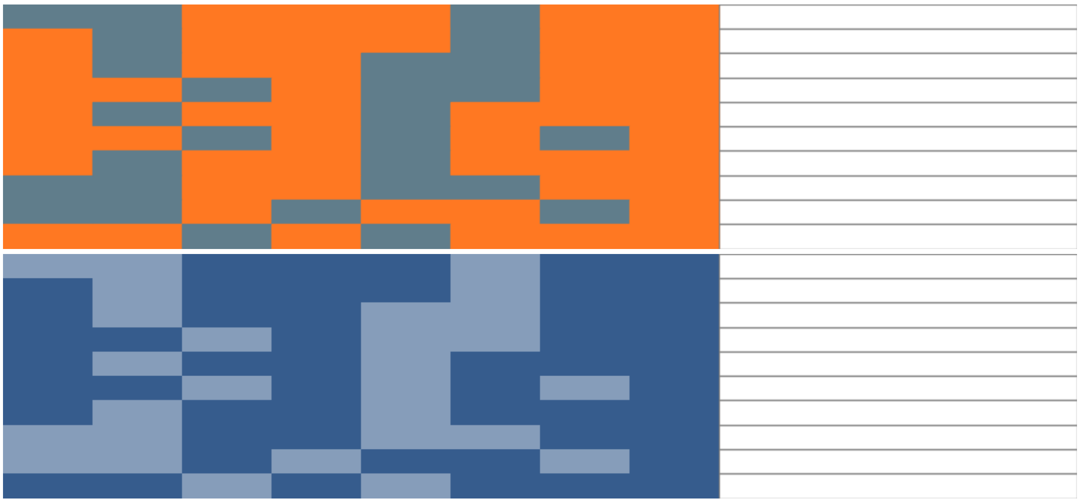


FIGURE D.74: *Alī al-Husarī al-Qayrawānī, yā laylu al-sabbu, al-Mutadārak.C*

FIGURE D.75: *Mīkhā'īl Nu'ayma, saqfu baytī hadīd, al-Mutadārak.P2*

Appendix E

Visualization experiments

This appendix provides an overview of various visualization explorations, including textual aspects such as verses and letters, phonetic aspects related to articulation points, as well as syllabic segmentation, summary, and detailed pattern visualizations. The example given features the poem *alā 'abkī*, composed in the *al-Tawīl.C* meter by the pre-Islamic poetess *al-Khansā'*.

Verse	Part 2	Part 1
1	إِذَا الحَرْبُ هَرَّتْ وَاسْتَمَرَّ مَرِيرُهَا	أَلَا ابْكِي عَلَى صَخْرٍ وَصَخْرٌ ثِمَالُنَا
2	عَلَى صَعْبِهَا حَتَّى اسْتَقَامَ عَسِيرُهَا	أَقَامَ جَنَاحَيْ رَبْعِهَا وَتَرَافَدُوا
3	مَنَاكِبُهَا مَسْمُومَةٌ وَنُحُورُهَا	بِبَارِقَةٍ لِلْمَوْتِ فِيهَا عَجَاجَةٌ
4	هَمَاهِمُ أَبْطَالٍ قَلِيلٌ فُتُورُهَا	أَهَلَّ بِهَا وَكُفُ الدِّمَاءِ وَرَعْدُهَا
5	وَصَخْرٌ إِذَا خَانَ الرِّجَالُ يُطِيرُهَا	فَصَخْرٌ لَدَيْهَا مِدْرَهُ الحَرْبِ كُلِّهَا
6	صَفَاهَا وَمَا إِنْ كَالصُّخُورِ صُخُورُهَا	مِنَ الهَضْبَةِ العُلْيَا الَّتِي لَيْسَ كَالصَّفَا
7	مَنِيعُ الذُّرَى عَالٍ عَلَى مَنْ يُثِيرُهَا	لَهَا شَرَفَاتٌ لَا تُنَالُ وَمَنْكِبٌ
8	وَأُخْرَى بِأَطْرَافِ القَنَاةِ شُقُورُهَا	لَهُ بَسْطَتَا مَجْدٍ فَكَفٌّ مُفِيدَةٌ
9	إِذَا مَلَّ عَنْهَا ذَاتَ يَوْمٍ ضَجُورُهَا	مِنَ الحَرْبُ رَبَّتْهُ فَلَيْسَ بِسَائِمٍ
10	بِهِ عَنْ حِيَالٍ مُلْقِحٍ مَنْ يَبُورُهَا	إِذَا مَا اقْمَطَرَّتْ لِلْمَغَارِ وَأَيْقَنَتْ

FIGURE E.1: Poem

TABLE E.1: Patterns

Verse	FP1S	FP2S	FP3S	FP4S	SP1S	SP2S	SP3S	SP4S
1	*SLL*	*SLLL*	*SLL*	*SLSL*	*SLL*	*SLLL*	*SLS*	*SLSL*
2	*SLS*	*SLLL*	*SLS*	*SLSL*	*SLL*	*SLLL*	*SLS*	*SLSL*
3	*SLS*	*SLLL*	*SLL*	*SLSL*	*SLS*	*SLLL*	*SLS*	*SLSL*
4	*SLS*	*SLLL*	*SLS*	*SLSL*	*SLS*	*SLLL*	*SLL*	*SLSL*
5	*SLL*	*SLLL*	*SLL*	*SLSL*	*SLL*	*SLLL*	*SLS*	*SLSL*
6	*SLL*	*SLLL*	*SLL*	*SLSL*	*SLL*	*SLLL*	*SLS*	*SLSL*
7	*SLS*	*SLLL*	*SLS*	*SLSL*	*SLL*	*SLLL*	*SLL*	*SLSL*
8	*SLL*	*SLLL*	*SLL*	*SLSL*	*SLL*	*SLLL*	*SLS*	*SLSL*
9	*SLL*	*SLLL*	*SLS*	*SLSL*	*SLL*	*SLLL*	*SLL*	*SLSL*
10	*SLL*	*SLLL*	*SLS*	*SLSL*	*SLL*	*SLLL*	*SLL*	*SLSL*

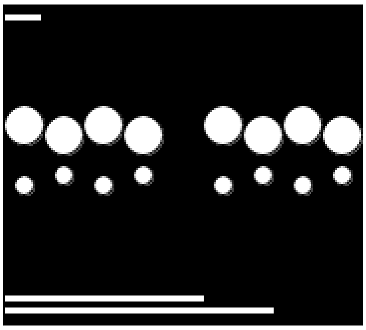


FIGURE E.2: Summary

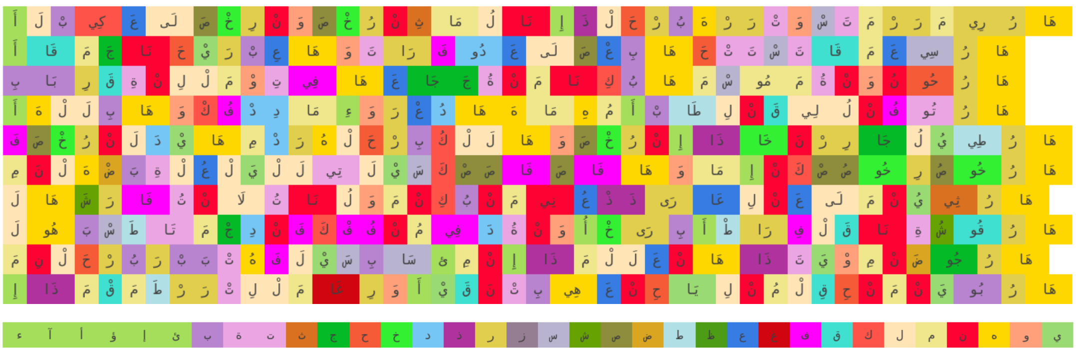


FIGURE E.3: Pronounced letters

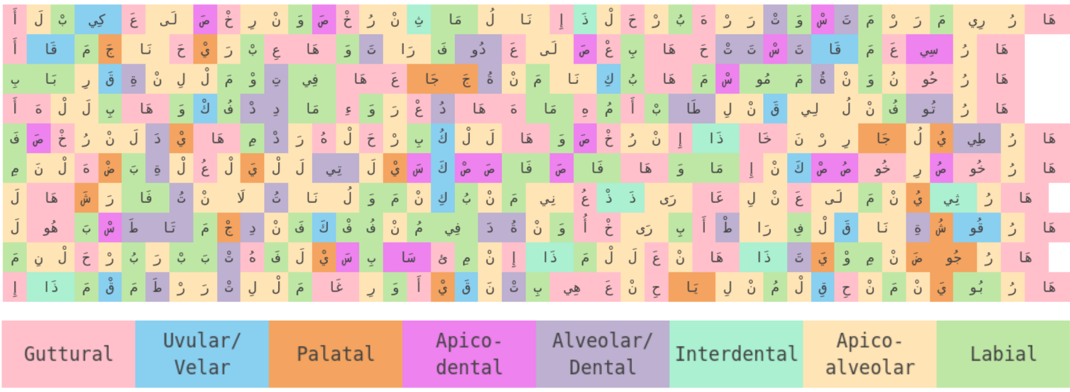


FIGURE E.4: Exit points

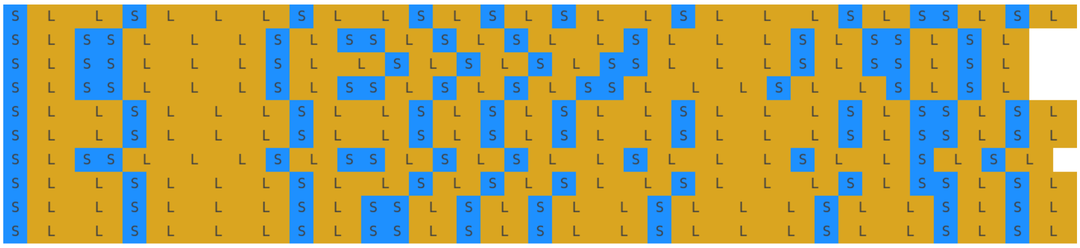


FIGURE E.5: Syllable option 1

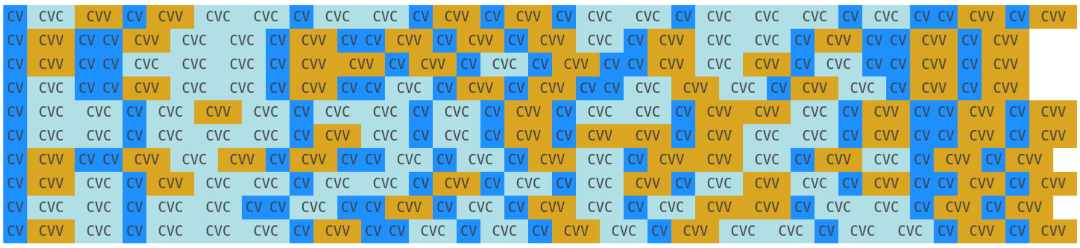


FIGURE E.6: Syllable option 2

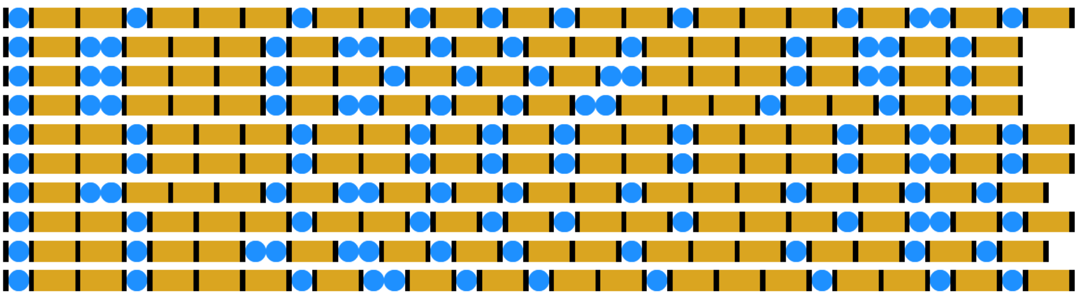


FIGURE E.7: Syllable option 3

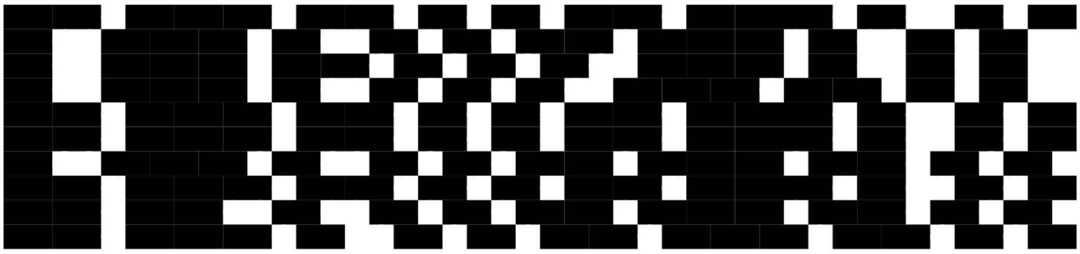


Figure E.8: Syllable option 4

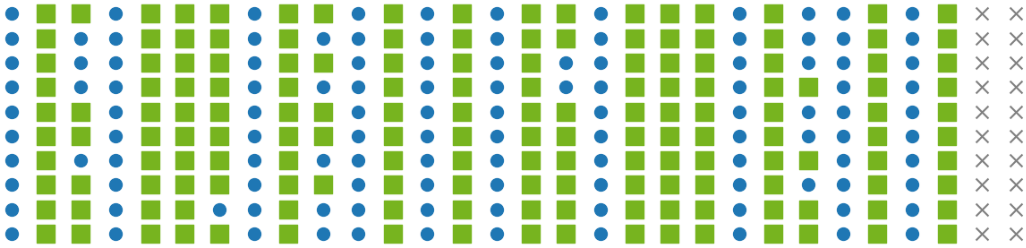


Figure E.9: Syllable option 5

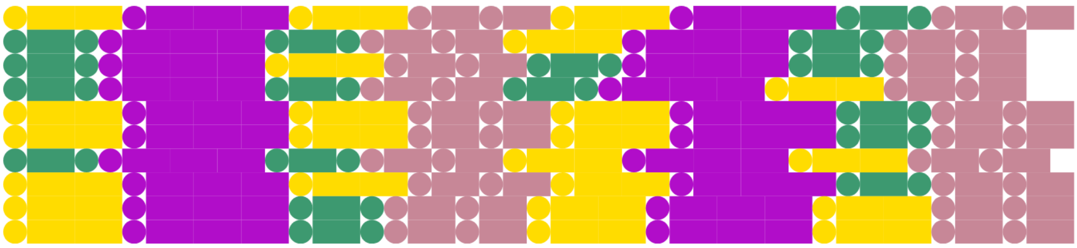


Figure E.10: Patterns option 1, own colors

Figure E.11: Patterns option 2, own colors

Figure E.12: Patterns option 3, shared colors