Using Natural Language Processing For Analyzing Arabic Poetry Rhythm

Munef Abdullah Ahmed
Faculty of Automatic Control and Computers
University Polytechnic of Bucharest
Bucharest, Romania
yabrahmun@yahoo.com

Stefan Trausan-Matu
Faculty of Automatic Control and Computers
University Polytechnic of Bucharest
Bucharest, Romania
trausan@gmail.com

Abstract— **One from the most difficult tasks for Natural Language Processing (NLP) is to analyze poetry, which uses a different genre of language than that considered by computer-based techniques. Therefore, computational analysis is an interesting task when we use NLP in poetry, but it is also challenging. There are a number of researchers that entered this field from NLP and they got promising results using mathematical analysis for poetry, including rhythm analysis. In this paper we focused on providing solution for automating the rhythm detection of the Arabic poems and finding the number of rhythm for each verse in poem and the total percentage for each rhythm in all verses of poem, additional to other characteristics for the Arabic poem like percentage of mobile letter in Arabic "harf mutaharrik", and The quiescent letter, in Arabic "harf sakin" , In spite of the number of studies in the computational analysis of poetry, we think it needs more, not only to make a better understanding of domain but also in developing applications, considering different literary tastes and the psychological effects, to give a recommendation to the readers and in plagiarism detection[1].**

Keywords—Rhythm; natural language processing; Al arud.

I. INTRODUCTION

In the beginning, we must know what we mean by 'poetry'. There are several definitions: poetry means words or sentences that refer to something beautiful; poetry is a manner of writing which is divided into stanzas or lines and uses a formal organization. From other definition, it is a way of listening to your senses and to understand the life you live by using your sensory channels. Poetry is an innovative expression of language that regularly uses one or more of the crafts of diction, rhythm, symbolism imagery, and sound [1]. Using a computer to process creative writing like poetry is challenging.

Computational analysis of poetry has been studied in many languages. English poems were analyzed in terms of style by Kaplan and Blei [2], who grouped them in clusters. They were classified into three types: syntactic, phonemic, and orthographic. Some studies use besides these kinds the lexical and multi-stylistic features. Kao and Jurafsky [3] classified the poems into ones written by expert and novice poets. Traditional Malay poetry has been classified by using Support Vector Machine (SVM) into various themes [4]. Using the computational analysis, Voigt and Jurafsky [5] show the decrease of the classical nature of the Chinese poetry in the twentieth century. This study was applied on Chinese and English language poems and classified them into three essential categories and nine subcategories [6]. Romanian poetry was analyzed by several mathematicians, which developed formal models [7], [8]. Computer-based analyses were also performed for Romanian [9], [10]. Poetry in the French language was also the subject of automatic analyses [11].

II.STATE OF THE ART

Up to now, the study of rhythm in languages focused on poetry and not on speech, prose or other non-poetic text. Hebert signaled that the rhythm piece can be divided into units that comply with particular patterns of serration and arrangement [12]. Repetition is the main factor for rhythm creation, as which was observed noticed by Tannen. She has focused on the repetition in face to face conversations and in achieving coherence [13]. Repetition was one important concern of Trausan-Matu as a device for artifact generation [14]. A tool for the analysis of rhythm in English texts was implemented by Niculescu and Trausan-Matu [10] starting from the metrical analysis proposed by Marcus and Dinu

[7], [8]. Depending on a broader perspective, Balint and Trausan-Matu found that rhythm and seriation can result from the special arrangement of any linguistic items, such as words, punctuation, syllables, part of speech, phonemes, syntactic structure, stresses, and unit length, by repetition and alternation of linguistic items. The rhythm was achieved by progressive or regressive series of these items, this model

opening the way to some additional devices in the field of rhythm production [15]. The studies that deal with the Arabic poetry analysis or rhythm analysis were very limited and less than the studies that deal with English poetry. There were some works that have been done by Tizhoosh [16], [17]. He used the text classification technique to distinguish between a poem and other texts (non-poem). In this way, there were considered some poetic features such as rhythm, meter, shape, and rhyme. On Arabic poetry meter, Al-Zahra has invented a system based on Al-Khalil bin Ahmed theory. This system used for meter identification was applied on written or spoken Arabic poems [18]. Reddy and Knight proposed a way for rhyme scheme identification [19]

III.FEATURES

In the computational analysis, there are some features used in the classification of the poetry, which can be broadly segmented into the following [1]:

- Orthographic features: this type of features are related to the measurement of multi-units like the number of words, lines, and stanzas also average word length, average line length and so on.
- Phonemic features: the sound of the important factors that have an effect on the poetry, it means the sound devices used in poetry like alliteration and rhyme scheme.
- Lexical features: each word type in this kind is a feature.
- Syntactic features: this kind of features deals with the part of speech and their frequencies in the poetry.

IV.ARABIC POETRY

The Arabic poetry can be classified into two main types. The first type is the measured or rhymed and the second main type is prose. The first type is used in general more than the second type. Al-Khalil bin Ahmed Al Farahidi is one of the most famous scholars that have studied the Arabic poetry and he has concluded that there are fifteen different meters failings under the rhymed poetry, and later he added one more meter. In Arabic, the word "buhur" ("seas" or "oceans" in English) is the name of meters of rhythmical poetry. The measuring unit for Al "buhur" is called "tafilah" (foot). Every "bayt" (verse) in the poem has a certain number of "tafilah"; all the "bayt"-s in the poem end with the same "qufiyah" (rhyme). The "bayt" can be changed from one meter to another meter by adding or removing any letter. Because the Kurdish language uses the same characters as the Arabic language, this feature can be used to make a comparison between the Arabic and Kurdish poem or others languages that is used the same Arabic characteristic. https://ar.wikipedia.org/wiki/%D9%84%D8%BA%D8%A9_%D9%83%D8%B1%D8%AF%D9%8A%D8%A9.

V.RHYME

Classical Arabic poems contain a number of verses. This number is not limited but in general between 20 and 100. In Arabic poetry, the length of the verse varies from one poem to another, and also the style of the verses may be different. Every verse is divided into two parts, which are equivalent in length and called hemistiches and in the same poem, each hemistich must apply to it the same meter. Also, in the classical Arabic poems, the last letter for each verse must be the same [19]. In the Arabic language, there are three main vowel sounds; each of them has two versions, short and long version. For the first version, there are diacritical marks below or above the letter which is before them, but for the second version, it is written as a whole letter. The above-mentioned versions are the same for rhyme purpose. By using these matching rules in Arabic poetry, it helps as in rhyme detection and makes it a simple task compared with English, where a number of different groups of letter combination can refer to the same rhyme [17]. In the modern Arabic poetry, a new challenge adds for the rhyme detection, which can be ignored the short vowels

VI.RHYTHM

Rhythm is important to life in general and for all types of human expression. Rhythm in spoken language is a special arrangement of loud and soft sounds units and that may refer to, pauses, stresses, etc. Rhythm has connections with the emotions and temperament of the speaker and this relationship is understood and explained in a natural manner by humans. Rhythm in the written language is less obvious than rhythm in spoken language. Human nature tends to link the written text with the writer through reasoning and world knowledge, at an unconscious level. Therefore, this kind of usage of common sense for finding a rhythm in the written language by the computer is very limited because of its artificial nature [12]. Rhythm is one of the traits of the poetical texts, in addition to rhyme.

However, it can also be identified in common natural language and epic texts, with the clear intention of the speaker or writer to express a certain state of contradiction, interest, wonder, confusion etc., or to emphasize an interrogation [10].

VII.RHYTHM METER IN ARABIC POETRY

In the Arabic language the science deals with the Arabic poetry called Al Arud "العروض", this science invented by Al-Khalil ibn Ahmad Al-Farahidi and he was the first used the grammatical terminology the classical Arabic poem has sixteen meters, which have the names (Tawil, Madid, Basit, Kamil, Wafir, Hazaj, Rajaz, Hamal, Sari, Munsarih, Khafif, Mudari, Muqtadib, Mujtath, Mutadariq, Mutaqarib) , a certain amount of changed are allowed for each type of this meters, The main terms used in his studies depend on the case of the letters [20]

1. The mobile letter, in Arabic "harf mutaharrik" which is mains that a letter must be followed by a short vowel and it is placed above or below the letter, there are several types of short vowels and this is very important in the Arabic language may change the meaning of word completely like:

   - "fatḥa" which has the sign "ـَ" in the Arabic means opening, a short sound "A" is pronounced after the letter.
   - "ḍammah" which has the sign "ـُ" in Arabic means closing, a short sound "U" is pronounced after the letter.

- "kasra" which has the sign "ِ" in Arabic means breaking, a short sound "I" is pronounced after letter.

The short vowel	Applied to the letter	pronunciation
Fatha	دَ	Da
dammah	دُ	Du
Kasra	دِ	Di
sukoon	دْ	D

TABLE I. THE SHORT VOWELS.

2. The quiescent letter, in Arabic "harf sakin" is a letter which is followed by a short vowel "sukoon" which has the sign "ْ" in Arabic means static see the table (1).

3. There are two markets in the Arabic language represent an extra letter:
- "shadde" which has the sign "ّ".

Example: the word "عدّ" means count, writing "عدّ" but it is pronounced"عدد"
- " tanween " which have three types:
  - The first type "tanween fatha" which has the following sign "ً",
  - The second type "tanween kasra" which has the sign "ٍ",
  - The third type "tanween dammah" which has a sign "ٌ".

The table (2) illustrate the tanween for letter Alf "ا".

The letter	اً	اٌ	اٍ
pronunciation	An	Aon	Ain

TABLE II. EXAMPLE FOR TANWEEN IN ARABIC POETRY

4. Some words in the Arabic language have letters did not write but it is pronounced for example to this case:
The word "لكن" "lken" which mains but in the English language, is writing "لكن" "lken" but it is writing "لاكن" "laken".

5. Some nouns begin with sun letter "harf shamci" after the letters "AL" "ال" and the other nouns begin with moon letter "harf qamarī" after the letters "AL" "ال" for example of these types:
- The sun letters "harf shamci": the word sun "الشّمس" is writing "الشّمس" but it is pronounced "اشّمس" in this case the letter "ل" is deleted from the word.
- The moon letter "harf qamarī":
  - the word moon "القمر" is writing "القمر" it is pronounced "القمر" in this case there is no change between the writing and phonetic.

6. There are two types for the market hamza
- "hamza wasul" which has the sign "~".

Example: for letter Alf "ا" will be written "آ".
- "hamza kateh" which has the sign "ء".

Example: for letter Alf "ا" will be written"أ".

The "hamza kateh" in the Arabic language it is cut the word into two parts on before it and the other after it. It is pronounced whenever it appears. It is the pronunciation of "U" there is no similar case in English language. When it comes at the beginning of the word it usually comes with an "Alf" and it may be over or under the "Alf" depending on the type of short vowel if it "fatḥa" or "ḍammah" it will be written over the letter and if the short vowel "kasra" it will be written under the letter for example:
The first case: The word "أنا" " Ana" which means me, is written "أنا".
The second case: The word "إيمان" "Eman" which means faith, is written "إيمان". (http://www.arabion.net/lesson5.html)

VIII.STYLES IN ARABIC POETRY

In the Arabic poetry the line is called "bayt" which mains in English language verse, this line is divided into two part: the first and second hemistich or two half-verses. One of which is called "sadr" in English language mains chest, which is referred to the first part of "bayt". The second part of the verse is called "ajuz" the meaning belly. The "sadr" has two parts:
- "arud" which is referred to the last word in the "sadr".
- "hashu sadr" meaning filling of the chest, is the rest words of the "sadr".

Also, the "ajuz" has two parts:
- "darb" meaning the hit, which is referred to the last word in the "ajuz".
- "hashu ajuz" meaning filling of the belly, is the rest words of the "ajuz".

The classical Arabic poems can be writing in three styles: the first style is written in a single column and two rows the first row for the "sadr" and the second row for the "ajuz". The second style of the Arabic poem is written in a single column and also two rows, but the "sadr" for each verse is written aligned to the right and the "ajuz" for each "byte" is aligned to the left. The last style of Arabic poem is written both "sadr" and "ajuz" in a single row which is separated by one or more space or punctuation marks the figures (1, 2, 3) illustrate this types. The English means of this verse which is used in the figures:
Let days do what they will and be content when fate treats you ill.

verse	H1 or "sader"	دَع الأيـَّامَ تـَـفـعـَـلُ مـَـا تـَـشـَـــاءْ
	H2 or "ajuz"	و طـبْ نـَـفـْـســـاً إذا حـكـَـمَ الـقـَـضـَـاء

Figure 1: An example of classical Arabic poems with one verse written in Style 1. H1 and H2 are the first and second hemistich [20].

Verse	H1 or "sadr"	دَع الأيَام تَفعلْ مَا تَشَاءْ
	H2 or "ajuz"	وطبْ نفساً إذا حَكَمَ ألقضَاء

Figure.2. An example of classical Arabic poems with one verse written in Style 2 [20].

Verse	H2 or "ajuz"	H1 or "sadr"
	وطبْ نفساً إذا حَكَمَ ألقضَاء	دَع الأيَام تَفعلْ مَا تَشَاءْ

Figure.3. An example of classical Arabic poems with one verse written in Style 3 [20].

IX. Automatic Rhythm Analysis in Arabic Poetry

The general rule for finding a rhythm in the Arabic poem is to limit. The mobile letter, and The quiescent letter, and only the character that is pronounced must be written [21]. The character that has no voice in the reading is deleted, as the bellow steps:

- The symbol "1" used for the moving letters. Which is referred to the letters that have short vowels or tanween or shadde above or under the letter. The symbol "0" used for stationary letters which are referred to the letter that has sukon.

- The shadde" ّ" above the letter means that the letter must be repeated.

Example: "ودّ" must be writing "ودد".

- The Tanween "ًٌٍ" with the letter must remove and add the letter "ن".

Example: "كتاباً" must be writing "كتابن".

- The letter which has sound during speech is written.

Example: "هكذا" must be writing "هاكذا".

- The letter "ل" in some noun pronounce if some letters which are called "Al qamaree" letters come after it. When the other letters which are called" Al shamsee" letters come after it, in this case, the letter isdeleted. (http://www.3raq4all.com/vb/showthread.php?t=54432)

After applying the above rules on the poem and treatment all the cases of the letters by using Python language, we found the percentage of the quiescent and mobile letters for several poems. The table below represents poem with seventeen verses. The case of letter considers the main part that is used for finding the rhythm.

No. of verse	Percentage of quiescent letters	Percentage of mobile letters
1	36.842%	63.157%
2	31.578%	68.421%
3	30%	70%
4	21.052%	78.948%
5	33.333%	66.222%
6	33%	67%
7	30%	70%
8	28.571%	71.428%
9	22.222%	77.777%
10	30%	70%
11	27.777%	22.222%
12	35%	65%
13	20%	80%
14	25%	75%
15	28.571%	71.428%
16	26.315%	73.684
17	23.529	76.470

TABLE III. THE PERCENTAGE OF MOBILE LETTERS

The figure (4) represents the automatic rhythms for each verse in the poem, by depending on the mobile and stationary letters which are found from the above

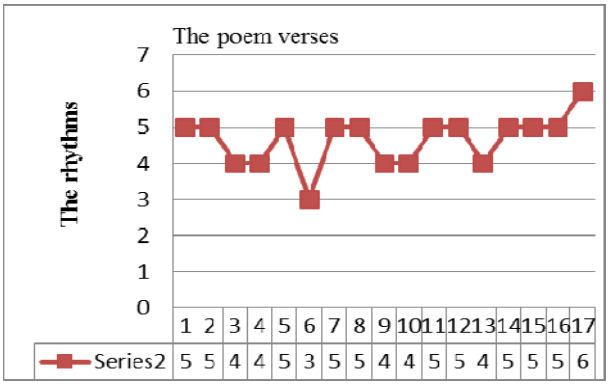

Figure 4: The rhythm in each verse of poem

The figure (5) represent the percentage for every rhythm calculated in the poem, The Arabic poem which is used in the analysis software must have "qufiyah" (rhyme) and all the above-mentioned conditions for the Arabic poem.

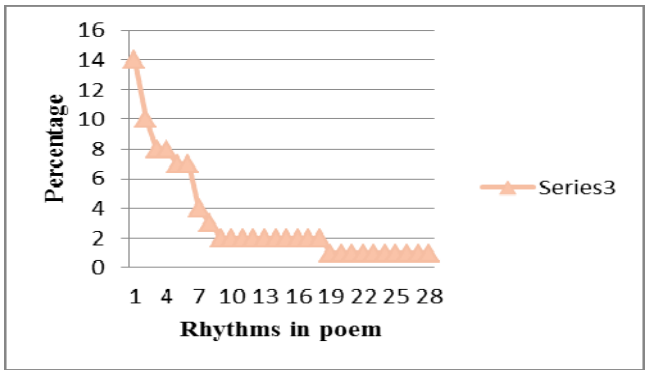

Figure 5: Percentage of each rhythm in poem

XL CONCLUSIONS

Using natural language processing for analyzing the rhythm of the Arabic poetry results to be helpful, in the process of automating the rhythm detection of the Arabic poem. It can, therefore, help to make comparisons between different types of Arabic poems and comparisons with a poetry of another language. Also, it helps the understanding of the power of natural language processing in these fields. The study of the rhythm of Arabic poetry may open the way to study the rhythm of the other kinds of Arabic text. The Arabic poetry which is used in the implementation must have"qufiyah" and all the requirements for the Arabic Poetry.

REFERENCES

[1] G. Rakshit, A. Ghosh, P. Bhattacharyya, and G. Haffari, "Automated Analysis of Bangla Poetry for Classification and Poet Identification." International Conference on Natural Language Processing, Trivandrum, India, 2015, pp. 1 - 7.

[2] D. M. Kaplan, and D. M. Blei, "A computational approach to style in American poetry." In Data Mining, ICDM 2007. Seventh IEEE International Conference on. 553– 558. IEEE.

[3] J. Kao, and D. 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. Montreal, Canada, pp. 8–17.

[4] N. Jamal, M. Mohd, and S. A. Noah, "Poetry classification using support vector machines," Journal of Computer Science, vol. 8, no. 9, pp. 1441, 2012.

[5] R. Voigt, and D. Jurafsky, "Tradition and modernity in 20th-century Chinese poetry". NAACL Workshop on Computational Linguistics for Literature 2013.

[7] S. Marcus, Poetica Matematică, Academy of the Socialist Republic of Romania Publishing House, Bucharest (1970).

[8] M. Dinu, Ritm şi rimă în poezia românească, Cartea Românească Publishing House, Bucharest, (1986).

[9] A.M. Ciobanu, L. P. Dinu. On the Romanian rhyme detection. 2012. Proceedings of COLING 2012: Demonstration Papers, pp. 87–94, COLING 2012, Mumbai, December 2012.

[10] I.D. Niculescu, S. Trausan-Matu, Rhythm analysis of texts using Natural Language Processing. In A. Iftene & J. Vanderdonckt (Eds.), Romanian Conference on Human-Computer Interaction (RoCHI 2016), pp.197- 112, 2016

[11] E. Boychuk, I. Paramonov, N. Kozhemyakin & N. Kasatkina, Automated Approach for Rhythm Analysis in French, Proceeding of the 15th Conference of FRUCT Association (Finnish-Russian University Cooperation in Telecommunications.).

[12] 12. Balint, M., Dascalu, M., & Trausan-Matu, S. (2016). The Rhetorical Nature of Rhythm. In 15th Int. Conf. on Networking in Education and Research (RoEduNet) (pp. 48–53). Bucharest, Romania: IEEE.

[13] D. Tannen, Talking Voices: Repetition, dialogue, and imagery in conversational discourse: Cambridge University Press, Vol. 26, 2007.

[14] Stefan Trausan-Matu, Repetition as Artifact Generation in Polyphonic CSCL Chats, Third International Conference on Emerging Intelligent Data and Web Technologies, IEEE Conference Publications, pp. 194- 198 (2012)

[15] M. Balint, and S. Trausan-Matu, "A critical comparison of rhythm in music and natural language," Annals of the Academy of Romanian Scientists, Series on Science and Technology of Information, vol. 9, no. 1, pp. 43-60, 2016.

[16] H. R. Tizhoosh, and R. A. Dara, "On poem recognition," Pattern analysis and applications, vol. 9, no. 4, pp. 325-338, 2006.

[17] H. R. Tizhoosh, F. Sahba, and R. Dara, "Poetic features for poem recognition: A comparative study," Journal of Pattern Recognition Research, vol. 3, no. 1, pp. 24-39, 2008.

[18] A. Almuhareb, L. A. Ibrahim Alkharashi, and S. H. Altuwaijri, "Recognition of classical Arabic poems," 2013, Computer Research Institute, KACSTRiyadh, Saudi Arabia.

[19] S. Reddy, and K. Knight, "Unsupervised discovery of rhyme schemes." Presented at the 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, pp. 77-82.

[20] R. Drory, Models, and contacts: Arabic literature and its impact on medieval Jewish culture: BRILL, 2000, p.196.

[21] M. A. Mohammed, Z. H. Salih, N. Ţăpuş, and R. A. K. Hasan, "Security and accountability for sharing the data stored in the cloud," in *RoEduNet Conference: Networking in Education and Research, 2016 15th*, 2016, pp. 1-5.