# Rhythm analysis of texts using Natural Language Processing

**Irina-Diana Niculescu**
Politehnica University of Bucharest
313 Splaiul Independentei,
Bucharest, Romania
irinaniculescu93@gmail.com

**Stefan Trausan-Matu**
University Politehnica of Bucharest
313 Splaiul Independentei,
Bucharest, Romania
and
Research Institute for Artificial Intelligence
and
Academy of Romanian Scientists
stefan.trausan@cs.pub.ro

**ABSTRACT**
The paper is introducing a research aiming to analyze rhythm in various genres of texts. After a review of the literature on rhythm formalization in texts, a Natural Language Processing application was developed for analyzing the rhythmicity in three cases: poem, prose, and political speech. The application implemented three formal approaches for rhythm analysis, proposed for the Romanian and French languages. Factors such as: rhythmic unit, rhythmic structure, and rhythmic index were considered. Comparisons are made between different genres, approaches, and languages.

**Author Keywords**
Rhythm; natural language processing; rhythmic unit; rhythmic structure; rhythmic index

**ACM Classification Keywords**
I.2.7 Natural Language Processing: Text analysis.

## INTRODUCTION

The paper presents the first steps of a research aiming to analyze rhythm in all genres of texts. After performing a state of the art, we have developed a tool for analyzing, using Natural Language Processing (NLP), the rhythmicity of texts, belonging to both the lyric (poems) and epic genre (prose, political speeches). This tool was used for analyzing three texts: a poem, a fragment of a short story and a speech. Some first conclusions were driven.

Rhythm is one of the characteristics of the poetical texts, together with measure and rhyme. However, rhythm is present in everyday speech and in epic texts as well, with the role of emphasizing certain words or syllables, with the explicit intention of the speaker or writer.

Rhythm can be defined as an alternation of stressed with unstressed syllables, while maintaining a certain homogeneity and symmetry.

From the point of view of NLP, rhythm analysis is an interesting research problem, as rhythmicity is of utmost importance both in daily conversation, as well as in the scientific or fictional texts. It is desirable to understand the use of rhythm and to find patterns that can be applied to the written texts in order to identify in an automatic way the rhythmicity of a specific text.

In the natural language used on a daily basis, emphasizing words is related to the speaker's intention to express a certain state (confusion, puzzlement, contradiction, interest, wonder) or to find information (interrogative sentences). In contrast with the everyday speech, the automation of this process is difficult because the computer is not able to assume human emotions. Therefore, in order to analyze a text's rhythmicity, algorithms based on formulas and observations shall be used. Thus, by analyzing texts for each genre, it can be created a model applicable to all the texts from a particular literary genre.

## STATE OF THE ART

An important amount of research has been done in connection with words stress and speech intonation. They differ from language to language. For example, for French always the last syllable is stressed.

Rhythm is obviously influenced by the stresses. Solomon Marcus introduced the concepts of rhythmic structure, rhythmic length and rhythmic index [1]. A text span (verse, sentence, paragraph, etc.) is analyzed by the following factors:

- the length of the considered text (number of words from the span);
- the rhythmic structure, i.e. a string of elements that represent the distance between two stressed syllables;
- the rhythmic length (the length of the rhythmic structure);
- the rhythmic index of a text span, i.e. the smallest natural number k which solves the following inequalities, whatever is the span belonging to the considered language [1]:

  span_length <= rhythmic_length <= span_length * k (1)

The inequalities (1) assume a relationship between the length of a rhythmic unit and the length of the corresponding rhythmic structure [1]. The uniformity of the language is given by the equality between the rhythm indexes for each considered text span. In the framework

of Solomon Marcus, the rhythmic unit used is the phrase or verse, in the case of a poem.

Mihai Dinu analyzed the poetries rhythm, from the point of view of the rhythmic structure. He proposes to "operate only with single stressed units" [2], so every rhythmic unit contains a single stressed word and the unstressed words that follow a stressed word belong to the next rhythmic unit. Therefore, a verse will be formed of several rhythmic units, unlike the approach used by Solomon Marcus, which considers the entire verse as a rhythmic unit.

Starting from this idea, Mihai Dinu develops what he calls a "succession law of the rhythmic units in a verse" [2]: two consecutive rhythmic units comply with the succession law if the distance between the stressed syllables is divided by 2, 3 or 4, depending on the verse's rhythm [2].

## IMPLEMENTATION DETAILS

We have developed an application that facilitates an analysis and a comparison of the rhythm between texts belonging to three literary genres: poetry, prose, political speech. A first experiment using the application has been performed with the following texts:

- The poem “A dream within a dream”, written by Edgar Allan Poe [3]. The poem is divided into two stanzas, the first one of 11 lines, and the second one of 13 lines.
- 2 paragraphs from the epic text “A Descent into the Maelstrom”, published in 1845 and written by the same Edgar Allan Poe [4].
- A political speech of Jesse Jackson from 1988 [5].

The application is developed for the texts written in English, because there are numerous materials, research articles, and software in this language. For example, for words hyphenation in English, an open-source project based on the Liang algorithm was used [7], and for the words stress, the CMU dictionary was very helpful (http://www.speech.cs.cmu.edu/cgi-bin/cmudict).

The application has been written in Java and it has a graphical interface made by using Java applets and Java Swing. In Figure 1 is a snapshot of the application interface.

The processing has the following stages:

- word hyphenation;
- finding the stressed and unstressed syllables;
- applying several approaches of dividing texts into rhythmic units: the Solomon Marcus approach, the Mihai Dinu approach (only for the poetic texts), and the Boychuk et al. approach [6].

A description of each stage will be made in the following paragraphs.

Writing a program of word hyphenation in English can be considered a difficult process, taking into account the numerous exceptions from the English language.

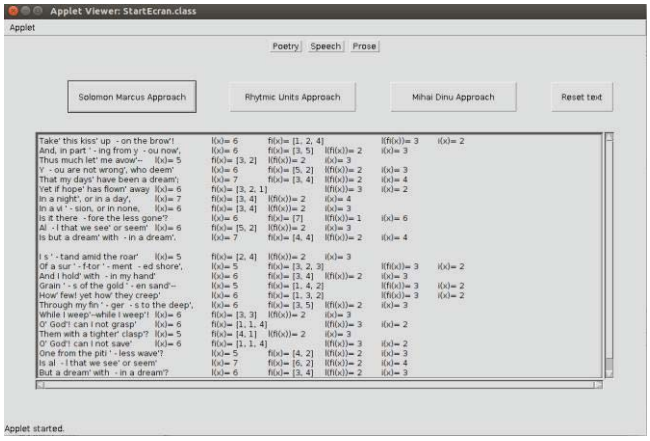


**Figure 1. A screenshot of the analysis of Poe’s poem using the approach of Solomon Marcus.**

For this step, we have used an open-source program, written in Java (https://github.com/joeha480/texhyphj). This program implements the word hyphenation algorithm of Franklin Mark Liang [7].

1. As mentioned, rhythm represents a symmetric alternation of stressed syllables with unstressed syllables. Thus, finding the stressed and unstressed syllables from a word appears as a necessity in a research dedicated to the rhythm analysis. For this step, we have used an open-source project developed at the Carnegie Mellon University: the CMU dictionary. The aim of this dictionary is to provide the words pronunciation from the North American English (http://www.speech.cs.cmu.edu/cgi-bin/cmudict). The dictionary itself contains a list of words and their hyphenation in 39 phonemes. In order to mark the primary stress of a word, it has been used the notation 1 for the stressed phoneme, the notation 2 for secondary stress and 0 for unstressed phonemes.
2. The problem encountered in the use of this dictionary was the fact that phonemes separation does not correspond always to the syllables separation, because a syllable may contain one or several phonemes. Therefore, in order to solve this issue we used a modified CMU dictionary (https://webdocs.cs.ualberta.ca/~kondrak/cmudict.html).
3. Unlike the standard one, the modified CMU dictionary contains the syllables separation of the words. In other words, phonemes are delimited, depending on the words hyphenation.

From the point of view of the implementation, every word from the text is searched in the CMU dictionary. If it is found, it will be verified in which syllable the primary stress can be found. After that, the word is separated in syllables according to the first stage and the corresponding syllable is stressed.

One of the problems encountered was that separation in syllables from the first stage does not correspond with the separation in syllables from the second stage, for some words. This is caused by the fact that the open-source

program used at the first step does not deliver 100% accurate results. For example, in the English language, the word "panorama" is hyphenated, according to the English dictionary, in: "pan - o- ra - ma". In the modified CMU dictionary, the phonemes separation is: "P AE2 - N ER0 - AE1 - M AH0". It should be noted that the notation "1" appears in the third syllable (corresponding phoneme "AE"), so it results that the third syllable is the stressed one. However, the word hyphenation according to the program used in the first step is: "panora-ma". A contradiction appears at the stressed syllables, and it leads to errors in the outlined approaches from the following steps. In the software implementation, when the position of the stressed syllable (found with the help of the CMU dictionary) is greater than the total number of syllables (computed by open-source program), the last syllable is considered stressed. Other examples are some nouns at plural: "streets" is hyphenated by the program in "street-s", which is wrong, because the word is monosyllabic.

Another problem is that certain words are not found in the CMU dictionary (e.g. the word "promontory"), so it cannot be found the stressed syllable in this way. From the point of view of the software implementation, the unfound words are considered unstressed.

As a result of the research we have performed, three approaches (of Solomon Marcus, Mihai Dinu, and Boychuck et al. approach) have been chosen in order to make a first comparison and an analysis of the rhythm in the three texts we considered, written in English, belonging to both the lyrical and epic genre (prose and speech). For the Solomon Marcus approach, the rhythmic unit is considered to be a phrase or a verse (in the case of poetry) [1]. For the Mihai Dinu approach (applied only on poetry), we delimited every verse into several metric units, each of them containing a single stressed word and the unstressed words that precede it [2]. For the Boychuck et al. approach, each phrase/verse is delimited by punctuation marks, coordinating conjunctions and subordinating conjunctions [6]. The three approaches differ by the rhythmic units division of the chosen texts, but the applied formulas are the same and the analysis is done depending on the same factors In other words, for every rhythmic unit of the texts and approaches chosen, the following rhythmic factors are computed: rhythmic structure, rhythmic length, rhythmic index. The algorithm for computing the rhythmic factors is based on the mathematical formalism described by Solomon Marcus [1] and mentioned in the *State of the art* section of this paper.The problem encountered in the implementation of these approaches is finding the stressed and the unstressed words from the English language. The research made by Solomon Marcus and Mihai Dinu have been done for the Romanian language, where the unstressed words are the prepositions, conjunctions, pronouns (in some cases), etc. A classification of words in stressed and unstressed is difficult to be done in the Romanian language, because there are several exceptions (for example, cases when the stress is on a pronoun), determined by the writer's intention to emphasize certain words [2].

According to Mihai Dinu, there are both unstressed monosyllabic and polysyllabic words, most of them being supporting words [2].

1. In the English language, words are divided into content words and function words. In the first category, there are nouns, verbs, adjectives, adverbs, etc i.e. the parts of speech that are essential for the transmission of important information in the communication. In the second category, there are pronouns, prepositions, conjunctions, articles, supporting words (http://pronuncian.com/content-and-function-words).

Daniel Jurafsky & James H. Martin mentioned that function words are generally unstressed and content words are stressed [8]. However, this is not a general rule.

Initially, the CMU dictionary was used to find out if a word has accent or not. We noticed that in the CMU dictionary, the majority of the words (with small exceptions) – monosyllabic, polysyllabic, function words - are stressed (the notation "1" at the corresponding phoneme). If all the words would have been considered stressed, the rhythmic index obtained for every sentence would have been 1 (number of stressed words = number of words from the rhythmic unit). Another idea was to consider the monosyllabic words as unstressed, but this would be wrong because there are many monosyllabic content words in English (e.g.: the noun "dream").

Elisa Hansen reiterates the fact that the stress of the monosyllabic words in poems depends on the context and the intention of the poet to suggest a certain sense [9]. Finally, a list of function words was considered. In the implementation, for each word from the text it is checked if it is in the list of the function words. If it is found in the list, it is not searched in the CMU dictionary and it is considered unstressed. Otherwise, the word is considered stressed. As mentioned, this approach does not provide the correct results in all cases, because it may happen that a function word is stressed in a specific verse or context.

## RESULTS OF A FIRST EXPERIMENT

### Solomon Marcus Approach

For the Solomon Marcus approach, a rhythmic unit was considered a sentence in the case of political speech and prose or a verse in the case of the poem.

In the case of the chosen poem [3], the number of the rhythmic units is equal to the number of the verses, so in this case it will be 24. All the verses are made of 5 - 7 words and the rhythmic index for each verse varies between 2, 3 (obtained for half of the rhythmic units) and 4, with the exception of the verse "Is it therefore the less gone", where the rhythmic index has resulted 6, because the only stressed word found was "gone". The maximal index – equal to 4 - has been obtained for the verses with the maximum length 7.

In the case of the prose [4], the number of the rhythmic units is equal to 9 (the number of the phrases from the two paragraphs chosen).

| Rhythmic index | % |
|---|---|
| 2 | 29.16% |
| 3 | 50% |
| 4 | 16.66% |
| 6 | 4.16% |

**Table 1. Rhythmic indices obtained for poetry using Solomon Marcus approach.**

The length of the phrases varies greatly (minimum length is 10, and the maximum 66), but it has been noticed that the length of the rhythmic structure is approximately half of the length of the sentence. The rhythmic index corresponding to each text span varies between 2 and 3. For each phrase, an unique rhythmic structure and obviously an unique rhythmic length are obtained. The prose chosen is formed by heterogeneous phrases, in contrast with the poem, where the verses keep mostly the same number of words and syllables.

| Rhythmic index | % |
|---|---|
| 2 | 44.44% |
| 3 | 55.55% |

**Table 2. Rhythmic indices obtained for prose using Solomon Marcus approach**

The political speech [5], unlike the paragraphs from the prose containing descriptive passages, consists mainly of short sentences. The number of rhythmic units (phrases) obtained is 239. In contrast with the poem, where the verses are formed by 5-7 words, and the prose, where the phrases have between 10 and 66 words, in the chosen speech there are sentences formed by a single word ("Dream.", "Leadership.", "Why?"), mostly stressed. The most frequent rhythmic index is 2 (obtained for 68% of the total rhythmic units).When the speaker wishes to underline certain aspects, sentences were repeated. Obviously, the sentences which are repeated have the same rhythmic structure, rhythmic length and rhythmic index (ex: "They work every day" - the length of the sentence is 4, equal to the number of words, the rhythmic structure is the string of length 2 composed of “[2, 3]”, corresponding to the words "work" and "day", and the rhythmic index is 2). It has been noticed that for sentences with similar structure (e.g.: "Call it pain", "Call agony it", and "Call it agony"), the same rhythmic structure and rhythmic index is obtained.

| Rhythmic index | % |
|---|---|
| 1 | 11.29% |
| 2 | 68.20% |
| 3 | 20.08% |
| 4 | 0.41% |

**Table 3. Rhythmic indices obtained for political speech using Solomon Marcus approach.**

### Mihai Dinu Approach

In the case of the Mihai Dinu approach, the formulas from the Solomon Marcus approach are applied, but each rhythmic unit contains a single stressed word. This approach has been applied only to the poem [3]. The lyrics are divided into 2, maximum 3 rhythmic units and similar structures such as "In a night", "or in a day", "But a dream", "within a dream" can be observed. In the Mihai Dinu approach 52 rhythmic units were obtained. The rhythmic units without a stressed word were not considered, because Mihai Dinu proposes "operating only with single stressed units" [2]. For example, the verse "In a vision, or in none" [3], the only stressed word found is "vision", so it was divided into two rhythmic units "In a vision," and "or in none". The last one is not a stressed unit, so it was not considered in the computation.

| Rhythmic index | % |
|---|---|
| 1 | 21.15% |
| 2 | 28.84% |
| 3 | 30.76% |
| 4 | 15.38% |
| 5 | 1.92% |
| 6 | 1.92% |

**Table 4. Rhythmic indices obtained for poetry using Mihai Dinu approach**

As in the Solomon Marcus approach for poetry, the most frequent rhythmic index obtained is 3. The rhythmic index 6 appears only once for both approaches, in the case of the verse "Is it therefore the less gone". The rhythmic lengths are obviously 1 (in 82% of the cases) or 2. In the Mihai Dinu approach, the last word in the rhythmic unit is the one stressed, so there are only two cases. If the last syllable is the one stressed, the rhythmic structure has a single element, so the rhythmic length is equal to 1. Otherwise, the rhythmic structure has two elements: the first one is equal to the position of the stressed syllable in the rhythmic unit, and the second one is the difference between the total number of syllables and the position of the stressed syllable. In the selected poem [3], the stressed words are monosyllabic (according to the open-source project used for hyphenation) for all the cases in which the rhythmic length is 1.

### Boychuck et al. Approach

In the third considered approach [6], the authors perform a delimitation of the rhythmic units depending on the punctuation marks, coordinating conjunctions and subordinating conjunctions. Although the paper refers to the French language, this approach was used for English and it was applied on the three texts. From the point of view of the implementation, a file for the punctuation marks [10] and for coordinating and subordinating conjunctions from the English language was made [11]. Each verse/phrase was delimited in rhythmic units and the same algorithm from the previous approaches was applied.

In the case of the poem, a similarity of the rhythmic units in this case with the rhythmic units from the Mihai Dinu approach can be noticed. A few examples of identical rhythmic units that appear in both approaches are: "can I not save", "or seem", "can I not grasp", "In a vision". In the Boychuck et al. approach, the number of the rhythmic units is equal to 39. The number is greater than the number of rhythmic units for Solomon Marcus approach, but smaller than the number of rhythmic units for Mihai Dinu approach. In contrast with the Mihai Dinu approach, in the implementation, the rhythmic units without stressed words were considered for the computation of the rhythmic factors, due to the fact that Boychuck et al. approach does not delimit the rhythmic units by stressed words. The most frequent rhythmic index obtained is 3, as for the Solomon Marcus and Mihai Dinu approach for poetry. The rhythmic index 6 appears only once as well, in the verse "Is it therefore the less gone". The rhythmic lengths vary between 1 (19 cases out of 39 rhythmic units), 2 (17 cases) and 3 (3 cases).

| Rhythmic index | % |
|---|---|
| 1 | 20.51% |
| 2 | 25.64% |
| 3 | 38.46% |
| 4 | 12.82% |
| 6 | 2.56% |

**Table 5. Rhythmic indices obtained for poetry using Solomon Boychuck et al. approach**

In the case of the prose, obviously the rhythmic units have a smaller length than in the case of Solomon Marcus approach, but the observation that the length of the rhythmic units is approximately half of the length of the corresponding sentence is kept. The number of the rhythmic units is 51, much bigger than the number of the phrases (9), due to the numerous coordinating and subordinating conjunctions and commas existing in the descriptive paragraphs chosen for the experiment. There are also rhythmic units without stressed words, such as: “Although”, “or”. The rhythmic index varies between 1, 2, 3 and 4, the biggest percentage is obtained the rhythmic index 2.

| Rhythmic index | % |
|---|---|
| 1 | 17.64% |
| 2 | 58.82% |
| 3 | 19.60% |
| 4 | 3.92% |

**Table 6. Rhythmic indices obtained for prose using Boychuck et al. approach**

In the case of the political speech, the number of rhythmic units obtained is equal to 561, much bigger than the number of the rhythmic units obtained in the Solomon Marcus approach (239). There are observed similar syntagms, repetitions or enumerations, which leads to similar rhythmic structures (ex: enumeration of nouns "suicide, cynicism, pessimism" is divided into three rhythmic units, each containing a stressed noun). Repetitions of rhythmic units such as "Common Ground!", "keep hope alive", "They work every day" are noticed. The examples of repetitions appear also in the Solomon Marcus approach. The most frequent rhythmic index is 2, as for the Solomon Marcus approach for political speech.

| Rhythmic index | % |
|---|---|
| 1 | 25.31% |
| 2 | 59.89% |
| 3 | 13.19% |
| 4 | 1.24% |
| 5 | 0.17% |
| 6 | 0.17% |

**Table 7. Rhythmic indices obtained for political speech using Boychuck et al. approach**

## CONCLUSIONS

All the three considered approaches offer a view of the rhythm in literary texts. For small, alike rhythmic units, similarities in the rhythmic structure and rhythmic index obtained are noted. In the case of the prose it has not been noticed an uniformity of the rhythmic units with similar structure, but in the case of the poem and the political speech, the repetition of syntagms/sentences leads to identical rhythmic factors.

However, the implementation of the three approaches have errors due to the mistaken word hyphenation provided by the used software. We also generalized that the emphasis does not fall on the supporting words (function words), but this is not a rule neither in the Romanian language, nor in the English language.

The research and implementation emphasized differences between the considered languages: English, French, and Romanian. However, the Solomon Marcus and Mihai Dinu approaches, even proposed for the Romanian language, proved to be usable also for English. Of course that further investigations should be done in the direction of comparing the constant aspects of rhythmicity across different languages and genres.

In conclusion, the natural language processing for the rhythm analysis of texts results to be helpful, as by automating the process for the rhythm determination, comparisons between various texts and genres can be made.

## REFERENCES

1. Marcus, S. *Poetica matematică*, Editura Academiei Republicii Socialiste Romania, 1970.
2. Dinu, M. *Ritm si rimă în poezia românească*, Cartea românească, 1986.
3. http://famouspoetsandpoems.com/poets/edgar_allan_poe/poems/18847, last accessed on 8 July 2016

4. http://poestories.com/read/descent, last accessed on 8 July 2016
5. http://www.famous-speeches-and-speech-topics.info/famous-speeches/jesse-jackson-speech-common-ground -and-common-sense.htm, last accessed on 8 July 2016
6. Boychuk, E., Paramonov, I., Kozhemyakin, N., Kasatkina, N. and Demidov, P.G. Automated approach for rhythm analysis of French literary texts. *Proceeding of the 15th conference of FRUCT association.* (Finnish-Russian University Cooperation in Telecommunications.), 15–23.
7. Liang, F. M. *Word Hy-phen-a-tion by Com-put-er*, written by, Report No. STAN-CS-83-977, 1983.
8. Jurafsky, D. and Martin, J. H. *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition*, Second edition, John Wiley, 2009.
9. http://classroom.synonym.com/use-monosyllables-poems-3516.html, last accessed on 8 July 2016
10. http://grammar.ccc.commnet.edu/grammar/marks/marks.htm, last accessed on 8 July 2016
11. http://grammar.ccc.commnet.edu/grammar/conjunctions.htm, last accessed on 8 July 2016