The Rhetorical Nature of Rhythm

Mihaela Balint, Mihai Dascalu, Stefan Trausan-Matu
University Politehnica of Bucharest, Computer Science Department
Bucharest, Romania
mihaela.balint@cs.pub.ro, mihai.dascalu@cs.pub.ro, stefan.trausan@cs.pub.ro

***Abstract*—Up to date, linguistic rhythm has been studied for speech, but the rhythm of written texts has been merely recognized, and not analyzed or interpreted in connection to natural language tasks. We provide an extension of the textual rhythmic features we proposed in previous work, and demonstrate its benefits for the task of text categorization. Rhythmic features require that the text be segmented in rhythmic units, the sentence being the default unit. We compute our features using three kinds of rhythmic units, provide a comparison between their relevance, and use the results to outline the rhetorical nature of rhythm.**

Keywords—rhythm; discourse segmentation; Rhetorical Structure Theory; RST Discourse Treebank; text categorization; Natural Language Processing; discourse analysis.

I. INTRODUCTION

Rhythm is essential to life in general, and to all forms of human expression, in particular. Rhythm manifests as a particular arrangement of items, which, for spoken language, may refer to stresses, pauses, low and high pitches, soft and loud sounds, etc. A particular rhythm correlates with the speaker's mood and emotions, and the correlation is naturally understood and interpreted by humans. The rhythmic signals of written language are less obvious, and people tend to relate to the writer and the text by using common sense and knowledge of the world, without becoming aware of subtler cues, which do, however, operate at an unconscious level. On the other hand, these kind of cues are extremely appropriate to be employed by computers, whose use of common sense and world knowledge is severely limited, and of an artificial nature.

Our first hypothesis is that rhythmic properties are indicative features in a large number of natural language processing (NLP) tasks, including text categorization (the task of assigning a text to a class from a set of predefined classes, e.g. to a genre), sentiment analysis, and ultimately natural language understanding. In this paper we present a refined and extended set of rhythmic features compared to the set we proposed in previous work [1], which improves the accuracy of the text categorization experiment presented in [1] by 5% (15% if stricter statistical criteria are enforced). The main addition to the model is the usage of elementary discourse units (EDUs), obtained through segmentation in the framework of Rhetorical Structure Theory (RST) [2], as rhythmic units.

Our second hypothesis is that rhythm correlates with the rhetorical structure of a text, as it does with its syntactical and grammatical structures. We test this hypothesis by observing the distribution of rhythmic features over EDUs, sentences, and punctuation units (textual spans separated by punctuation), and the efficiency of each subset (one for every choice of rhythmic unit) of features in the task of text categorization.

The remainder of this paper is structured as follows. Section II overviews other work related to this study. Section III describes our implementation of a discourse segmenter, which was used to extract the EDU structure of the analyzed texts. In section IV we present the refined and extended feature set, together with the corpora on which we run our experiments. A statistically-driven feature selection, together with the results of classification, are presented in section V. Section VI is devoted to a comparison between rhythmic units, and section VII presents our final considerations and directions for future work.

II. RELATED WORK

Up to date, the rhythmic study of language has focused on the prosodic study of speech, and solely on the metrical properties of written text. But rhythm, as observed by Hebert [3], is achieved whenever the rhythmic piece can be segmented into units that follow specific patterns of arrangement and seriation. The main device for rhythm creation is repetition – which was signaled by Tannen [4], but she focused on the role of repetition in achieving fluency and coherence in face to face conversations. Trausan-Matu et al. also studied repetition as a means for artifact generation [5] and, in general, for discourse building in online chat conversations [6].

Boychuk et al. [7] implemented a tool for the analysis of rhythm in French literary texts, which highlights text that meets the following three criteria: *"(1) presence of a number of repeated elements; (2) the smallest possible distance between the original element and the repetitions; (3) high frequency of occurrence in the text*" [7]. Niculescu and Trausan-Matu [8] implemented a similar tool, enhanced with metrical analysis, for the study of English texts.

In our previous work [1, 9], we adopted a broad view of rhythm that can result from a particular arrangement and seriation of any linguistic items, such as punctuation, phonemes, stresses, syllables, words, parts of speech, n-grams, syntactic structure, or lengths of rhythmic units. We consider

that rhythm is achieved through repetition, alternation, or progressive/regressive sequencing of linguistic items, and the model is open to the addition of other devices for rhythm production.

Our metrical features are partly inspired by the work of Solomon Marcus [10], who proposed several measures for the metrical analysis of a textual span: the *rhythmic structure* (the string of inter-stress distances), the *rhythmic length* (the length of the rhythmic structure), and the *rhythmic index* (which roughly defines the ratio between the length of the span and the rhythmic length).

Rhetorical Structure Theory (RST) [2] is a discourse theory based on the assumption that texts are hierarchically structured, and adjacent text parts are rhetorically related. RST-style discourse segmentation represents the segmentation of a text into the smallest units between which rhetorical relations (such as *elaboration, cause, contrast*) hold. So far, this task was most successfully completed using statistical methods, which modeled discourse segmentation as either a problem of classification [11-14], or of sequence labeling [15-17]. The methodology of discourse segmentation differs in whether it is centered on individual tokens (words), or rather on pairs of tokens. Traditionally, feature vectors are computed for every token inside a sentence, and used to decide whether a boundary should be inserted before that token.

The state-of-the-art segmenter was designed by Feng and Hirst [17], who successfully implemented pair-centered segmentation. Their approach considers every pair of adjacent words in a sentence, computes individual features for both tokens in the pair, and uses them to decide whether a segment boundary belongs between the two tokens. Their method passes twice through the text. The first pass produces a preliminary segmentation based on traditional lexico-syntactic features. The second pass uses all the features of the first pass, plus features derived from the previous segmentation decisions, achieving an F1-score of 92.6% for the class of actual segment boundaries, when tested on the RST Discourse Treebank (RST-DT) corpus [18].

III. Discourse Segmentation

Accurate segmentation into EDUs is crucial for obtaining accurate EDU-based rhythmic features and for testing the adequacy of EDUs as rhythmic units. Therefore, our implementation follows closely the state-of-the-art model of Feng and Hirst [17], opting for: (1) segmentation as classification (we use a Linear Support Vector Machine), (2) pair-centered features, and (3) passing twice through the text, with the results of the first pass entering the computation of features for the second pass. The model was trained and tested on the RST-DT corpus [18], a collection of Wall Street Journal newspaper articles for which complete RST analyses are provided.

We used three classes of features: lexico-syntactic, contextual, and global. Each document was first subjected to POS tagging, word lemmatization, and constituency parsing (using the Stanford Parser; http://nlp.stanford.edu).

Lexico-syntactic features include the lemma and part-of-speech (POS) of each word in the pair, and information about the largest syntactic constituent starting/ending in each of the two words: top syntactic tag, top production rule, and depth in the constituency tree.

Contextual features capture the same kind of information about the previous word and the next word in the text, relative to the examined pair.

Global features are only used in the second pass, and they take their name from the fact that they consider the EDU structure of the entire document (as it was assessed during the first pass). They include the lemmas and POSs of the neighboring predicted EDU boundaries, the distance to these boundaries, and information (for now, only the top syntactic tag) about the two syntactic subtrees that span from the current position to each of the left and right neighboring predicted boundaries.

For the RST-DT test set, the discourse segmenter achieved accuracies of 86.38% for the class of true boundaries and 99.24% for the class of non-boundaries, with a micro-average of 98.36%.

IV. Feature Extraction

Both the discourse segmenter and the rhythmic feature extractor were implemented in the Python programming language, making use of the NLTK (for Natural Language Processing; http://www.nltk.org), sklearn (for Linear Support Vector Machines; http://scikit-learn.org/stable), and SQLite3 (for interfacing with SQL databases; https://www.sqlite.org) packages. Other useful tools were the Stanford Parser (used to detect syntactic parallelism), and the CMU Pronouncing Dictionary (used for syllabification and the identification of stresses; http://www.speech.cs.cmu.edu/cgi-bin/cmudict).

We tested our features by performing text categorization on the same datasets reported in previous work, for purposes of comparison: three corpora containing famous speeches (http://www.famous-speeches-and-speech-topics.info/famous-speeches), student essays written by non-native English speakers (the Uppsala Student English corpus; http://ota.ox.ac.uk/desc/2457), and newspaper articles taken from the RST Discourse Treebank. The data were balanced by keeping only the longest (in number of sentences) 110 documents in each corpus. Table I presents the properties of the three balanced datasets. The studied documents are of significantly different lengths, but they provide a balanced input to the subsequent statistical analyses, in the form of 330 observations for each feature (110 corresponding to each genre), with the feature values already normalized by text length.

TABLE I. Properties of the three balanced datasets.

Dataset	Number of documents	Number of sentences
Speeches	110	14,111
Essays	110	7,553
Articles	110	5,269

In previous work [1], we grouped our features into five classes, based on the type of linguistic item that was used as a building block of rhythm: organizational, lexical, grammatical, phonetical, and metrical. For the purposes of this study we introduce a new classification, into (1) *generic features*, whose values do not depend on a particular segmentation into rhythmic units, and (2) *unit-based features*, which describe individual properties and interactions of rhythmic units. These features are calculated for every type of rhythmic unit that we consider: sentence (S), punctuation unit (PU), and elementary discourse unit (EDU). Sentences are delimited by sentence boundaries (full stops, question and exclamation marks) and define the syntactic structure of the text. Punctuation units are delimited by any kind of punctuation (mostly sentence boundaries and commas) and define the grammatical structure of the text. Finally, EDUs are delimited on (subjective and ambiguous) rhetorical grounds, and define the rhetorical structure of the text. We used gold standard EDUs for the corpus of newspaper articles (taking advantage of the rhetorical annotations of RST-DT), and used the segmenter described in section three to produce rhetorical structures for speeches and essays.

Generic features include the frequency of commas, the number of content words and n-grams that are assessed as frequent in the document, the average length of words in number of syllables, or the frequencies of assonances (the repetition of a single vocalic phoneme over a small piece of text), alliterations (same for consonants), and rhymes (the repetition of a sequence of phonemes).

Unit-based features refer to the length of rhythmic units in either words or syllables, and patterns of length variation over longer sequences of units (e.g., increasing length, or alternation between longer and shorter units). They also capture the preferred placing (in the beginning, middle, or the end) of frequent content words (considered the main themes of the document) inside units, and patterns of lexical repetition at the beginning or end of units (e.g., anaphora, epistrophes). From a metrical point of view, we assess whether a unit contains an odd or even number of syllables, and whether it ends in a stressed syllable (applying pressure for a continuation).

Four features not included in previous work are introduced in order to take into account the entire metrical schema of a unit: the real-valued rhythmic index (*k_real*), the most frequent integer-valued rhythmic index in a document (*k_int*), the percentage of units with the integer-valued index 2 (*k_2*), and the percentage of units with the integer-valued index 3 (*k_3*). They are based on Solomon Marcus's definition for the rhythmic index of a textual span, as the smallest natural number *k* for which the following inequality holds [10]:

$$number\_of\_words/k \leq rhythmic\_length \leq number\_of\_words*k \quad (1)$$

The rhythmic structure is defined as a string of distances between primary stresses, and there is at most one primary stress in a word (maybe zero, because in determining metrical schemas we remove stress from stop words), hence the rhythmic index is always at least 1. Our real-valued rhythmic index is an approximation of the smallest real number for which the above inequality holds, and the *k_real* of an entire document is obtained as an average of the values for individual units. To also consider the integral value proposed by Marcus, for each document, we originally reported only the integer-valued index that resulted for a majority of units. We noticed that, for all studied documents, *k_int* is either 2 or 3, but the distribution of these two values differs significantly from one text genre to another, which motivated the inclusion of two additional metrical features: *k_2* and *k_3*.

Most rhythmic features are based on *repetition* of organizational, lexical, grammatical, or phonetical nature. Such repetition may be spatially *constrained*, as in the case of lengths of units (organizational), anaphora (lexical), syntactic parallelism (grammatical), or assonances (phonetical), or *unconstrained*, meaning that the repeated items may occur however far from each other, as in the case of frequent words and n-grams (lexical), or the frequency of commas (grammatical).

Other features capture patterns of *alternation* or *progressive/regressive sequencing,* for example the frequency of sequences that alternate longer and shorter units, or the longest uninterrupted sequence of units in increasing word-length order.

V. Feature Selection and Classification

Two standardized statistical methods were used to perform feature selection and ultimately text categorization: Discriminant Function Analysis (DFA) [19] and Multivariate Analysis of Variance (MANOVA) [20, 21].

DFA is a statistical classification method that works with normally distributed features which are not linear combinations of each other. Several steps were undertaken to meet these criteria while retaining as many predictive features as possible in the model. First, we duplicated each feature in order to consider both its original value and its log value, with the intention to select later the most predictive representation (if any) that has met all the statistical requirements. Then, we removed all features which demonstrated non-normality (log values were introduced so that more features would survive this step). Second, we removed the features that did not pass Levene's test for the equality of variances, a step that was not undertaken in our previous work [1] and that would have reduced the accuracy of that model from 81.51% to 68.20%. Third, we assessed multicollinearity based on pair-wise correlations with a correlation coefficient $r > .70$, and, from each pair of multicollinear features, removed the feature with a weaker effect. After a MANOVA, we were left with 35 features whose values discriminate significantly between the three datasets (Wilks' $\lambda = .108$, $F(96, 560) = 11.88$, $p < .001$, partial $\eta 2 = .671$).

Eight of these features were deemed significant by a stepwise DFA. They can be observed in Table II, in decreasing order of effect size. The tags [S], [PU], or [EDU] in front of a feature name mean that, in computing that feature, the reference units were sentences, punctuation units, or EDUs, respectively. The [Log] tag means that the log value of

the feature was used instead of its original value. The DFA retained two canonical discriminant functions ($\chi^2(df = 7) = 187.657$, $p < .001$, leading to the separation of genres depicted in Figure 1), which were used in classification, with the results reported in Table III. It can be seen that the genre was correctly predicted for 285 out of 330 documents, resulting in an accuracy of 86.36%, an improvement of 18.36% over our previous model, if we replicate the standards of feature selection. Using a leave-one-out cross-validation, the accuracy mildly dropped to 84.24%. The resulting weighted Cohen's Kappa of 0.795 demonstrates substantial agreement between the actual genre and the predicted genre.

TABLE II. TESTS OF BETWEEN-GENRE EFFECTS FOR THE MANOVA AND DFA SELECTED FEATURES.

Feature	df	*F*	*p*	η^2 partial
[Log][PU] % of units with integer-valued index = 3	2	272.769	<.001	.625
[EDU] Real-valued rhythmic index (k_real)	2	188.551	<.001	.536
Number of syllables per word	2	68.483	<.001	.295
Normalized number of commas	2	55.410	<.001	.253
[PU] % of units with a stress on the final syllable	2	19.867	<.001	.108
[PU] % of alternating word length structures	2	16.603	<.001	.092
[PU] Longest rising word length sequence	2	14.304	<.001	.080
[S] % of units with a stress on the final syllable	2	13.048	<.001	.074

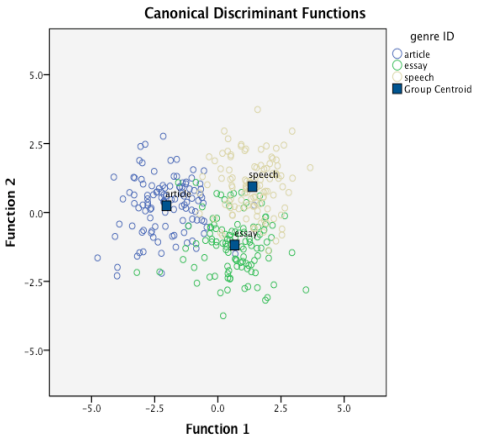

Fig. 1. Separation of genres using canonical discriminant functions.

TABLE III. CONFUSION MATRIX FOR DFA CLASSIFICATION.

	Genre	Predicted Group Membership		
		Article	Essay	Speech
Original	Article	104	5	1
	Essay	8	87	15
	Speech	3	13	94
Cross-validated	Article	101	7	2
	Essay	8	86	16
	Speech	4	15	91

VI. COMPARISON BETWEEN RHYTHMIC UNITS

The DFA presented in the previous section retained eight features, out of which two are generic, one is based on sentences, four are based on punctuation units, and one is based on EDUs. This suggests that punctuation units are the most appropriate choice of rhythmic units for discriminative purposes. To further test this hypothesis, we performed classification using each set ([S], [PU], and [EDU]) of unit-based features in isolation (supported only by generic features). The results depicted in Table IV show that both punctuation units and EDUs perform better than sentences.

TABLE IV. PERFORMANCE FOR ISOLATED TYPES OF RHYTHMIC UNITS.

Unit type	Selected predictors	Accuracy
[S]	Number of syllables per word	81.8%
	Frequency of commas	
	% of units with a stress on the final syllable	
	% of units with integer-valued index = 3	
	[Log] Longest rising word-length sequence	
	[Log] Real-valued rhythmic index	
[PU]	Number of syllables per word	85.5%
	Frequency of commas	
	% of rising word-length structures	
	% of repetitive word-length structures	
	Longest rising word-length sequence	
	% of units with a stress on the final syllable	
	[Log] Number of words per unit	
	[Log] Longest alternating syllable-length sequence	
	[Log] Longest repetitive syllable-length sequence	
	[Log] % of units with integer-valued index = 3	
[EDU]	Number of syllables per word	85.2%
	Frequency of commas	
	Longest rising word-length sequence	
	Longest rising syllable-length sequence	
	% of units with a stress on the final syllable	
	Real-valued rhythmic index	
	% of frequent words found in the 3rd part of units	

In classification, even learning algorithms that are able to transform the data into more discriminative representations work better if the original data is normalized. Therefore, we prefer, for machine learning applications, rhythmic units that produce normally distributed feature values. The number of features in each feature set that survived the test of normality and Levene's test for equality of variances are depicted in Table V. The first value between parentheses refers to original feature values, while the second one refers to log values. Removing duplicates refers to choosing either the original or the log value for a feature, in case both values met all the criteria.

TABLE V. HOW FEATURE SELECTION AFFECTS UNIT-BASED FEATURES.

Unit type	Original size	Normally distributed	Pass Levene's test and remove duplicates
[S]	66 (33*2)	38 (17+21)	13 (7+6)
[PU]	66 (33*2)	46 (23+23)	19 (13+6)
[EDU]	66 (33*2)	45 (22+23)	14 (13+1)

It can be seen that punctuation units contribute the most to the model (quantitatively speaking), and that values based on EDUs are very well (normally) distributed prior to any normalization. Also, the EDU-based original values were consistently chosen as more predictive than their transformed values. From the results of Tables IV and V, we conclude that both punctuation units and EDUs produce more useful features than sentences (although in discourse studies the sentence is often taken for granted).

While *k_int* was not deemed a significant predictor in the task of text categorization (it never passed the normality test), it proved helpful to assess the different nature of rhythmic units, and to measure the uniformity of a text, genre, or language. For all rhythmic units, the integer-valued index tends to be either 2 or 3 (sometimes higher, and seldom lower), but k_int (characterizing entire documents) is more often 2 in the case of punctuation units and EDUs, and 3 in the case of sentences, indicating that most smaller units are denser in content compared to entire sentences.

Based on Marcus's idea that the uniformity of rhythmic indices across units defines the uniformity of the language [10], we define the uniformity of a textual genre as the uniformity of its *k_int* across documents. Table VI depicts the uniformity of speeches, essays, and newspaper articles based on sentences, punctuation units, and newspaper articles. What is specified is the percentage of documents that have the *k_int* displayed between parentheses. We find that articles have the lowest and speeches the highest *k_int,* and that EDUs produce the most uniform results (more than 90% of documents have the same $k\_int = 2$). A lower *k_int* suggests more content words, meaning that the text is denser in information (as was expected from newspaper articles).

TABLE VI. GENRE UNIFORMITY AS A FUNCTION OF RHYTHMIC UNIT.

Genre	[S]	[PU]	[EDU]
Article	83.64% **(2)**	100.0% **(2)**	100.0% **(2)**
Essay	80.00% **(3)**	83.64% **(2)**	93.64% **(2)**
Speech	92.73% **(3)**	62.73% **(2)**	77.27% **(2)**

At a micro-level, if we take into account the index of each unit, rather than the predominant index of each document, we obtain the uniformity values depicted in Table VII. What is specified is both the percentage of units with index 2, and the percentage of units with index 3 (the most frequent indices).

TABLE VII. DISTRIBUTION OF RHYTHMIC INDICES ACROSS UNITS.

Genre	[S]	[PU]	[EDU]
Article	59.62% **(2)** 35.21% **(3)**	60.42% **(2)** 18.50% **(3)**	64.23% **(2)** 21.68% **(3)**
Essay	32.89% **(2)** 57.17% **(3)**	45.45% **(2)** 36.89% **(3)**	49.46% **(2)** 33.91% **(3)**
Speech	26.85% **(2)** 59.56% **(3)**	42.85% **(2)** 37.51% **(3)**	46.14% **(2)** 36.79% **(3)**

The results (which motivated the inclusion of *k_2* and *k_3* in the model) show a very good separation between articles and the other two textual genres, and moderate separation between essays and speeches.

VII. CONCLUSIONS AND FUTURE WORK

There seems to be an underlying rhythmic structure for everything we perceive or do, which helps organize the flow of information between us and the surrounding world. The main purpose of this work is to recognize the rhythmic structure of written texts, identify its building blocks, and establish relevant connections between rhythmic properties and other textual properties that are interesting for NLP. We chose to focus on text categorization because this gives us access to large amounts of pre-categorized data on which to test our hypotheses. In previous work, we classified texts pertaining to three different genres: speeches, essays, and newspaper articles; this study relies on the same datasets in order to quantify the improvements brought to the model.

There are two original additions to the previous model. First, we augmented the set of rhythmic units on which to compute rhythmic features with RST-style EDUs. Unlike sentences and punctuation units, EDUs are hard to obtain, and human agreement on the results of rhetorical segmentation is significantly lower compared to other annotation tasks. For the RST-DT corpus we had access to gold standard (human) segmentation, and for the corpora of speeches and essays we had to implement our own version of a discourse segmenter, which was tested and proven to perform well on RST-DT. We also contrasted the results of EDU-based features for the three corpora, and found the variations to follow the variation patterns of other rhythmic units, which suggests an adequate functioning of the segmentation model.

Second, we introduced the concept of rhythmic index, inspired by the work of Solomon Marcus, and used it to assess the informational density and the rhythmic uniformity of the three genres. We also found that the distribution of the rhythmic index across units in the document correlates with the genre of the document (articles produced a significantly different distribution).

The fact that the index of a document varies least when based on EDUs, the naturally good (normal) distribution of rhythmic features over EDUs, and the 85.2% accuracy obtained in the EDU-based classification experiment, all concur with our hypothesis that there is an affinity between rhythm and the rhetorical structure of a text.

In future work, we plan to test the power of rhythmic features in other NLP tasks, such as the analysis of involvement in conversations, sentiment analysis and rhetorical relation labeling. The difficulty in tackling this last task is the limited annotated data that is available, which is why we chose to refine our features by first approaching other problems. Other phenomena worth investigating are the affinity between genres and particular rhythmic units, or the affinity between individual texts and particular rhythmic indices. Maybe the rhythm of different texts would be studied best in relation to different units. Or maybe, just like musical scores have time signatures, written texts have hidden "rhythmic index signatures", and we may come to talk about a text that is written in a "$k\_int = 3$" rhythm. In fact, the musicality of chat was proven by its sonification, which became a starting point for music composition [22].

ACKNOWLEDGMENT

The work presented in this paper was partially funded by the EC H2020 project RAGE (Realising and Applied Gaming Eco-System) http://www.rageproject.eu/ Grant agreement No 644187. We would like to thank Scott Crossley and Laura Allen for their support in conducting the statistical analyses.

REFERENCES

[1] M. Balint, M. Dascalu, and S. Trausan-Matu, "Classifying Written Texts through Rhythmic Features", in Proceedings of AIMSA 2016, Varna, Bulgaria, 2016.

[2] W. C. Mann, and S.A. Thompson, "Rhetorical structure theory: Toward a functional theory of text organization". Text-Interdisciplinary Journal for the Study of Discourse, 8(3), 1988, pp. 243-281.

[3] L. A. Hebert, "Little Semiotics of Rhythm. Elements of Rhythmology". Signo. Available online http://www.signosemio.com/semiotics-of-rhythm.asp [Accessed July 10th, 2016]

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

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

[6] S. Trausan-Matu, G. Stahl, A. Zemel, "Polyphonic Inter-animation in Collaborative Problem Solving Chats", Research Report, Drexel University, Philadelphia, 2005, Available online http://mathforum.org/wikis/uploads/Stefan_Interanimation.doc [Accessed July 23th, 2016]

[7] E. Boychuk, I. Paramonov, N. Kozhemyakin, and N. Kasatkina, "Automated approach for rhythm analysis of french literary texts", in Proceedings of the 15th Conference of Open Innovations Association FRUCT, IEEE, 2014, pp. 15-23.

[8] I. D. Niculescu, S. Trausan-Matu, "Rhythm analysis of texts using Natural Language Processing", RoCHI Conference, in press, 2016

[9] 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, 9(1), 2016, pp. 43–60.

[10] S. Marcus, "Poetica matematică". Bucharest: Editura Acad. Rep. Soc. România, 1970.

[11] R. Soricut, and D. Marcu, "Sentence level discourse parsing using syntactic and lexical information", in Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, Association for Computational Linguistics, 2003, pp. 149-156.

[12] R. Subba, and B. Di Eugenio, "Automatic discourse segmentation using neural networks", in Proc. of the 11th Workshop on the Semantics and Pragmatics of Dialogue, pp. 189-190, 2007.

[13] S. Fisher, and B. Roark, "The utility of parse-derived features for automatic discourse segmentation", in Annual Meeting-Association for Computational Linguistics, vol. 45, no. 1, 2007, p. 488,.

[14] S. Joty, G. Carenini, and R. T. Ng, "A novel discriminative framework for sentence-level discourse analysis", in Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Association for Computational Linguistics, 2012, pp. 904-915.

[15] H. Hernault, D. Bollegala, and M. Ishizuka, "A sequential model for discourse segmentation", in International Conference on Intelligent Text Processing and Computational Linguistics, Springer Berlin Heidelberg, 2010, pp. 315-326.

[16] N. X. Bach, N. L. Minh, and A. Shimazu, A., "A reranking model for discourse segmentation using subtree features", in Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Association for Computational Linguistics, 2012, pp. 160-168.

[17] V. W. Feng, and G. Hirst, "Two-pass Discourse Segmentation with Pairing and Global Features", arXiv preprint arXiv:1407.8215, 2014.

[18] L. Carlson, D. Marcu, and M. E. Okurowski, "Building a discourse-tagged corpus in the framework of rhetorical structure theory", in Current and new directions in discourse and dialogue, Springer Netherlands, 2003, pp. 85-112.

[19] W. R. Klecka, "Discriminant analysis", No. 19, Sage Publications, Thousand Oaks, CA, 1980.

[20] J. P. Stevens, "Applied multivariate statistics for the social sciences". Routledge, 2012.

[21] G. D. Garson, "Multivariate GLM, Manova, and Mancova", Statnotes: Topics in multivariate analysis, 2009.

[22] A. Călinescu, S. Trausan-Matu, "A system for sonification of chat conversations", Annals of the Academy of Romanian Scientists, Series on Science and Technology of Information, vol.6, nr.2, 2013, pp.23-42.