A CRITICAL COMPARISON OF RHYTHM IN MUSIC AND NATURAL LANGUAGE

Mihaela BALINT[1], Ştefan TRAUŞAN-MATU[2]

Abstract. *This paper presents an analysis of the resemblances and differences between music and natural language starting from the rhythmic dimension. It then applies the results for the natural language processing in order to discriminate types of texts. The first part of the paper contains an extended state of the art on the rhythm in both music and natural language. In the second part, starting from the ideas discussed previously, an experiment is presented, which was done for comparing, from the rhythmic point of view, two corpora: one of famous speeches and a second containing Wall Street Journal articles.*

Keywords: *Rhythm, natural language processing, musical rhythm, discourse analysis, poetry analysis, rhetorics*

1. Introduction

Rhythm may be viewed in several ways. First of all, it is a permanent reality for us, at least if we take into account heart beats, breath, the sequence of days, nights, seasons and, not the last, music and dance. Meanwhile, it refers also to the quest for harmonious proportions in all creative acts.

Rhythm can be temporal, as in the regularity of musical beats, the pattern of stressed and unstressed syllables in speech, or spatial, as in the alternance of colors and distances in paintings, sculptures, or architecture. Through rhythm, the thoughts and feelings of the artists, performers, and audiences are brought to resonance, which facilitates the understanding and recognition of the creative piece.

1.1. Rhythm production across arts

Hebert [11] identifies three stages of rhythm production: the segmentation into units, the arrangement, and the seriation of units. Examples of units include pulses and beats (in music), phonemes, graphemes, and semes (in linguistics), colors and shapes (in paintings). Arrangement refers to the temporal and/or spatial distribution of units over the creative piece. Seriation is also of temporal and/or spatial nature.

[1]Eng., PhD student, Faculty of Automatic Control and Computers, University ”Politehnica” of Bucharest, Bucharest, Romania, (mihaela.balint@cs.pub.ro).
[2]Full member of AOSR. Prof., PhD, Faculty of Automatic Control and Computers, “Politehnica” University of Bucharest and Senior Researcher, Research Institute for Artificial Intelligence of the Romanian Academy, Romania. stefan.trausan@cs.pub.ro

Some disciplines constrain both time and consecution (a musical performance is typically not slowed down or accelerated, and one does not interrupt the listening to go back to previous fragments), others constrain consecution but not time (texts are read from left to right, but one can take one's time for certain passages, or go back, or skip ahead), others constrain neither time, not consecution (there is no proper duration or direction to watch a painting). We can observe that this taxonomy correlates with performing arts (constrained time and consecution), literary arts (unconstrained time, constrained consecution), and visual arts (unconstrained time, unconstrained consecution). Rhythm in performing and visual arts is sensory, typically pertaining to sound and vision. In literary rhythm there is no direct stimulation of the senses, therefore, this kind of rhythm is more subtle and more subjective. Based on literary technique and one's unconscious tendency towards proportion, the recipient's imagination creates a pattern of mental images and sounds.

1.2. The structure of rhythmic patterns

Rhythm is built as a succession of identical or different units. To find order in this type of construction is essential to human cognition and for the tasks of understanding, remembering, or learning. Rhythmic patterns can be classified according to the number of positions in a sequence (e.g. a quatrain has four lines), the number of units per position (e.g. a syllabic position has multiple phonemes, but only one syllable – so it depends on what we consider a unit), the total number of units to choose from (e.g. the number of primary colours in a painting which uses only primary colours), the organization of the pattern (e.g. rhyme units A and B can be organized as AABB – in couplet rhymes, ABAB – in cross rhymes, or ABBA – in envelope rhymes), the duration of units (rhythms can be *isometric* when units share the same (range of) duration, *allometric* when they use different ranges, or *parametric* when various ranges distinguish various groups of units) etc.

1.3. The complexity of rhythmic patterns

The rhythmic complexity can be judged from several perspectives. Hierarchical complexity refers to the existence of structure at various levels (such as time, space, and meaning). Mathematical complexity refers to mathematical regularity in the distribution of units. Cognitive complexity deals with how easy it is for recipients to perceive and understand the aspects of the rhythm. Performance complexity is similarly focused on the recipient, namely on the ability to perform a rhythm after witnessing it.

Pressing [19] distinguishes three main reasons for the importance of hierarchical complexity. First, there is an increase in the range of elements that may have an

impact on the recipient. Second, different levels of complexity will correspond to different levels of sophistication in the recipient. Third, multiple levels of order will allow for multiple experiences with the piece, each with a different focus.

The next sections use these criteria in a critical view of the study of rhythm across music and language.

2. Rhythm in music

Music can be studied in either its intended form (musical scores) or in its realized form (actual musical performance).

2.1. Intended form (musical scores); Structure

Lerdahl and Jackendoff's influential Generative Theory of Tonal Music [14] emerged from the desire to find a grammar for music, similar to Chomsky's generative grammar for language. Unlike related efforts, Lerdahl and Jackendoff did not aim to "translate" Chomsky's grammar, but to describe the cognitive processes used by an educated listener to make sense of a musical idiom. According to their theory, rhythmic and pitch patterns are recognizable at four different levels: the *grouping structure*, the *metrical structure*, *time-span reduction*, and *prolongational reduction*. The grouping structure divides music into motives, which are further grouped into themes, phrases, periods, theme-groups etc. The metrical structure is the hierarchy of strong and weak beats throughout the score.

Time-span reduction captures the relative prominence of pitch events in a piece. Sequences of events are grouped into units with exactly one dominant event upon which other (subordinate) events elaborate. The term "reduction" refers to the recursive grouping of events into single sequences, which yields a tree structure. Left branches in the tree correspond to events which elaborate on the following event, right branches to events which elaborate on the previous event. Dominant events bear the name of *heads*. Prolongational reduction also refers to events which elaborate on other events, but in this case, the elaboration is of harmonic and melodic nature, rather than metrical. The structure takes again the form of a tree. There are two kinds of elaboration: in *prolongation*, a pitch event (represented as a circular node in the tree) is elaborated into several copies of itself (represented as branches of equal priority); in *contrast*, left branches signify *delay* in relation to the bass, while right branches mean *progression*. The structure at each of the four levels is established through rules divided into three categories: *well-formedness* rules, *preference* rules, and *transformational* rules. Well-formedness rules specify which structures are valid, preference rules are mechanisms of selection from the set of all possible structures, and transformational rules provide a way to change structures into other structures

(typically into distorted structures not covered by the well-formedness rules). Rules reflect the tendency to reduce complexity. Preference for symmetry, parallelism and evenly spaced beats is a preference for mathematical regularity. The distance between beats at higher levels is chosen as a multiple of the distance at the immediately lower level. Cognitive complexity is reduced by inter-relating the 4 levels. For example, dominant events in the time-span reduction coincide with the strongest beats in the metrical structure. Beats at higher levels are also beats at lower levels. Groups in the grouping structure become time-spans in the time-span structure. Groups do not overlap, tree branches do not cross, and an event cannot belong to several branches.

2.2. Pattern matching

Musical rhythm is parametric. Units are notes of shorter or longer duration, which creates the potential for various prototypical patterns. Christodoulakis et al. [7] address the problem of classifying songs according to their rhythm. In their paradigm, a rhythm is a sequence of two types of units: S (slow) and Q (quick), which correspond to the relative duration of notes, with the convention that S = 2Q. They propose an efficient algorithm which, given a musical text as a string of durations of events, identifies the longest subsequence that matches a particular rhythm. The actual duration of units is not a priori known, which makes the problem more difficult than traditional string matching problems. If the size of Q is assumed to be q, then Q *matches* a substring of durations if and only if q equals the sum of durations in the substring. When the substring has a single element, the match is said to be *solid*. A rhythm *covers* a substring of durations if this substring can be split into consecutive parts such that each n-th unit in the rhythm matches each n-th part. The algorithm works in four stages. First, for every possible value, it identifies all occurrences of solid matches for S = s. Second, the neighboring areas are transformed into sequences of Q's and S's. Third, the algorithm looks for matches of the given rhythm. Fourth, the maximum cover is found for every possible duration of a Q unit and the algorithm returns the best overall cover. The running time of the algorithm is *O(n log H)*, where H is the maximum value in the string of durations.

2.3. Complexity

Thul and Toussaint [22] studied the ability of six rhythm complexity measures to predict the human difficulty of performance. They define rhythm as a cyclic binary sequence of pulses evenly spaced on a circle. These pulses can be sounded or silent. Sounded pulses correspond to beginnings of notes and are called *onsets*. For the purpose of this study, *offsets* (ends of notes) are considered simultaneous to onsets. The *meter* designates a sequence of special pulses called *beats*.

Beats are evenly spaced, occurring at multiples of n/p, where n equals the total number of pulses and p spans the set of proper divisors of n.

The studied measures could be assigned to three categories. Two measures quantify the regularity of the rhythmic pattern (*weighted note-to-beat distance* uses a formula based on where the onsets lay between beats, *off-beatness* counts the onsets that do not align with beats). Two measures are cognitive, based on the possibility to detect specific musical or rhythmic events in the score (*Pressing's cognitive complexity* considers the occurrence of five events: *fill*, *run*, *upbeat*, *subbeat*, and *syncopation* – each weighted from 1 to 5, respectively, while *Keith* assigns weights from 1 to 3 for the events of *hesitation*, *anticipation*, and *syncopation*). The focus of the study is on the third category of measures, which are based on a metrical hierarchy proposed by Lerdahl and Jackendoff [15] and Yeston [24]. For all proper divisors of the number of pulses n, for all positions that are multiples of such divisors, the weight of the corresponding pulse position is increased.

For example, if n equals 16, the pulse position 8 has its weight increased four times because 8 is multiple of 1, 2, 4, and 8. The resulting weights of the positions of the onsets enter the formula for the complexity measure. This is the framework for both measures in the category, namely *metrical (onset-normalized) complexity*, and *Longuet-Higgins and Lee*. The study confirmed that measures based on a weighted metrical hierarchy are best at predicting human rhythmic performance complexity.

2.4. Realized form (musical performance)

Honing [12] moves the focus from the structural to the perceptual aspects of musical time. Performed rhythm adds to the rhythmic structure the dimensions of *tempo* (the speed of performance) and *timing* (units may be lengthened, shortened, delayed, or anticipated). A listener is able to separate judgment about rhythmic classes from judgment about particular variations which belong to expressive performance. In fact, such variations are better perceived in reference to the traditional rhythmic classes. The tempo is obtained by measuring the distance between onsets and referring it to the corresponding distance in the score. The timing variations have been shown to adapt to the global tempo, but information from both tempo and rhythm class needs to contribute to a good prediction of timing. Honing uses the familiar representation of a performance as a string of durations, and defines the mathematical concept of *rhythm space* (the set of all possibly performed rhythms), an n-dimensional space where n is the number of durations in the string. He advises to study the placement of prototypical rhythmic structures and preferred tempo and timing variations inside this infinite space. A proposed experiment asks participants to listen to a number of performed rhythms,

with each rhythm corresponding to a particular location in the rhythm space. A *centroid* is defined as the location of a rhythm frequently identified by listeners as a particular rhythmic class. Honing expects centroids to be slightly shifted from the placement of purely mechanical renditions of a rhythm, suggesting that successful interpretations add to the communicative value of a piece.

3. Rhythm in language

As with music, we can divide language into written language (intended form) and spoken language (realized form).

3.1. Intended form (written language); Structure

Phonology is the branch of linguistics which investigates the systematic organization of sounds in languages. For rhythm studies, we refer to *metrical phonology*. Based on the phenomena of *syllabification* and *stress* (the relative emphasis placed on a syllable), metrical phonology uses non-overlapping binary trees to create a hierarchy of stresses inside utterances. This hierarchy is strongly related to the constituency parsing of the text.

In "The Sound Pattern of English" (SPE) [6], Chomsky and Halle devise rules to transform the (English) text as a sequence of phonemes into the phonetic form that is uttered by a speaker. Word-level rules are the *Main Stress Rule*, the *Alternating Stress Rule*, and the *Stress Adjustment Rule*.

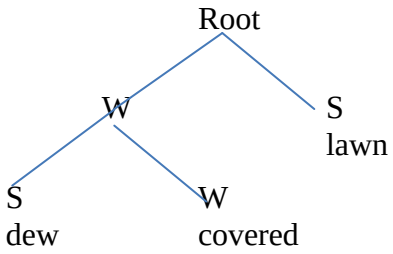
The Main Stress Rule is a formula based on the word's part of speech and on its affixes. The Alternating Stress Rule captures the pattern that words of three or more syllables tend to have primary stress on the antepenultimate vowel and tertiary stress on the final vowel (e.g. húrricàne, ánecdòte).

The Stress Adjustment Rule states that all non-primary stresses within a word should be weakened by one. Chomsky and Halle use a cyclic reassignment of primary stresses, so this rule applies whenever a certain primary stress is assigned. Two rules apply at and beyond word level: the *Nuclear Stress Rule* and the *Compound Rule*. The Nuclear Stress Rule states that the last heavy stress in a phrase receives the nuclear stress (e.g. the main stress of "equality" is heavier than the main stress of "absolute" in the phrase "absolute equality"). The opposite happens for lexical compounds such as "toy factory" or "sugar cane", as stated by the Compound Rule. Some of the rules in SPE are *ordered disjunctively*. That is a type of interaction that prevents a rule B to apply to a certain unit after a rule A has already applied to it. SPE served as the point of departure for Liberman and Prince [16]. They argue that stress subordination is best represented by a tree structure. The resulting hierarchy is a function of two basic ideas: *relative prominence* and the *metrical grid*.

Relative prominence is based on the constituent structure of the phrase. Liberman and Prince use the phrase “dew-covered lawn” (and its bracketed representation [[[déw][cóvered]][láwn]]) to compare their approach to that of Chomsky and Halle. Chomsky and Halle tell us to assume primary stress for the main stresses of each of the three words, and start applying rules from the innermost set of brackets, i.e. from “dew-covered”.

This is a compound adjective subject to the Compound Rule, so the primary stress stays on “dew”, and the stress of “covered” is demoted by one. Then, the Nuclear Stress Rule applies to the phrase “dew-covered lawn” and leads to a primary stress on “lawn”, a secondary stress on “dew”, and a tertiary one on “covered”.

Liberman and Prince suggest to represent relative prominence by annotating the nodes in the syntactic tree with the symbols “S” (for “strong”) and “W” (for weak). For “dew-covered lawn”, this yields the following tree:

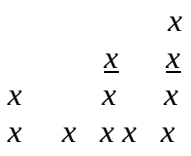


In practice, stress is an *n*-ary feature (although it is often treated as binary). In this representation, no special numbering is necessary to account for primary, secondary, tertiary etc. stresses. Trees also make the structure easier to navigate. Relative prominence is preserved under embedding, so cyclic reassignment of stress is no longer necessary. The authors rephrase the Nuclear Stress Rule and the Compound Rule as: “In a configuration $[_{C}AB_{C}]$, if C is a phrasal category, B is strong; if C is a lexical category, B is strong if and only if it branches”.

In Liberman and Prince’s theory, the hierarchical representation of stress is equally characteristic of words, lexical categories, and phrasal categories. But words do not benefit from a ready-made tree, as compounds and phrases do. Word trees are erected using the SPE model of assigning stress features to vowels. Strong constituents (syllables) then correspond to stressed vowels.

The metrical grid functions separately from the rules of rhythm assignment. It is based on the observation that in some cases, to avoid clashing stresses, stress is not preserved under embedding. The pressure to move the stress in such situations proves the tendency to maintain an alternating rhythmic pattern in language. The authors note that the same phenomenon happens in English, German, and Hebrew, but its manifestations differ across languages. The metrical grid attaches to every syllable a column whose height reflects the relative prominence of that syllable.

All columns start with a height of one, and columns are progressively erected until they satisfy the *Relative Prominence Projection Rule*: "In any constituent on which the strong-weak relation is defined, the designated terminal element of its strong subconstituent is metrically stronger than the designated terminal element of its weak subconstituent" [16]. Ill-sounding cases such as "achromatic lens" will be represented as follows:



achromatic lens

Regarding the *x*'s in the columns as elements, we say that two elements are metrically *adjacent* if they are on the same level (at the same height) and no other elements of that level intervene between them.

The elements just below adjacent elements, if any, can also be adjacent - and, in this case, the original elements are metrically *clashing*, or not - and in this case the original elements are metrically *alternating*. It can be observed that the underlined *x*'s are clashing in the example above.

In English, clashing stresses are solved through *Iambic Reversal*: a [weak strong] configuration is replaced by its [strong weak] counterpart. Iambic reversal is optional, and it is usually applied when the alternative is particularly unfluent.

3.2. Pattern matching

Prototypical rhythmic patterns occur in poetry, in the form of *metrical feet*. The basic metrical feet in English poetry have exactly one stressed syllable and one or two unstressed syllables. *Rising meters* use *iambs* (marked as *x* /, meaning 2 syllables: first unstressed, second stressed) or *anapests* (*x x* /). *Falling meters* use *trochees* (/ *x*) or *dactyls* (/ *x x*). According to the number of metrical feet, a line can be *monometric*, *dimetric*, *trimetric*, etc. Other rhythmic features of poems refer to the number of lines, the number of syllables, the organization of rhymes etc. To check whether a line matches a metrical foot, word stress patterns have to be obtained.

The straight-forward solution is to use a dictionary such as the CMU Pronouncing Dictionary (http://www.speech.cs.cmu.edu/cgi-bin/cmudict), but there are several drawbacks. The dictionary is not exhaustive and it provides no good indication for how to treat monosyllabic words (in practice they are sometimes stressed, sometimes unstressed). Some poems include stress reversals, and sometimes syllables are added, subtracted, or merged, for rhythmic convenience. To overcome these difficulties, Greene et al. [9] use unsupervised learning to extract

word stress patterns directly from poetry, a method that easily ports to other languages. They acquire data from sonnets because of their fixed form (iambic pentameter), and then use the word patterns to generate new poems and to translate Italian poetry into English.

3.3. Complexity

Many forms of poetry display simple rhythmic patterns as the ones mentioned above. The case is more complicated for prose texts, where no pattern is obvious apart from the general tendency to alternate stresses with non-stresses. Beeferman [3] takes the purely lexical stream and its associated word stresses extracted from the CMU Pronouncing Dictionary (discarding secondary stresses), and trains an n-gram model to predict probabilities of the form $p(s_k|s_0, s_1,\ldots,s_{k-1})$, where s_i is the stress class of syllable *i* in the text. *N*-grams that do not fit between sentences boundaries are disregarded. Beeferman then computes the stress entropy rate on the same corpus of over 60 million syllables of Wall Street Journal text, obtaining a rate of 0.795 bits per syllable for $n = 6$. To make sure that the result doesn't solely reflect intra-word stress regularity but also correlates with word arrangement, he randomizes word order inside sentences and computes a new entropy rate, which is slightly higher than in the original setup. This confirms that sentences with a higher probability of occurrence, i.e. sentences that are actually preferred by writers, are more rhythmical.

3.4. Realized form (speech)

Prominent units in speech may refer to stress, duration, or fundamental frequency, and differ from one language to another. Pike [18] and later Abercrombie [1] proposed the existence of two language classes, according to the preferred method to achieve isochrony: stress-timed languages and syllable-timed languages. The first class is characterized by equal duration inter-stress intervals and is represented by languages such as English, German, Russian, Arabic etc.

The second class displays equal duration syllables, encountered in French, Italian, Turkish, Cantonese Chinese etc. Later, Ladefoged [13] postulated a third class - that of mora-timed languages, such as Japanese. A mora is a unit that defines syllable weight. Roughly, we can say that short syllables have one mora, while long syllables can have two or three.

While appealing, this taxonomy has failed to find support in empirical studies, and isochrony should be regarded mainly as a perceptual phenomenon. Isochrony aside, research has found that there are quantitative differences between the hypothesized classes, which Ramus, Nespor and Mehler [20] captured in the following sentence indicators: *%V* (the proportion of sentence time devoted to vocalic intervals), ΔV (the standard deviation of vocalic intervals), and ΔC (the

standard deviation of consonantal intervals). Ramus et al. found that a combination of %V and ΔC correlates best to rhythm classes. English, which presents the phenomenon of *vowel reduction*, has a smaller %V and a higher ΔC compared to French. The possibility of an increased number of consonants per syllable leads to more syllable types, which correctly individualizes English, Dutch, or Polish (with more than 15 syllable types) on the (%V, ΔC) scale. In a related study, Grabe and Low [10] captured the difference in duration between successive vocalic intervals, in the form of the *normalized Pairwise Variability Index (nPVI)*, where the denominator is used to normalize for speech rate.

$$n\,PVI = 100 \times \left\lfloor \sum_{k=1}^{m-1} \left| \frac{d_k - d_{k+1}}{(d_k + d_{k+1})/2} \right| / (m-1) \right\rceil$$

In the above formula, m is the number of intervals, and d_k is the duration of the k-th interval. Using nPVI, stressed-timed languages such as English, German, or Dutch were demonstrated to display greater durational variability between successive measurements.

French and Spanish were classified as syllable-timed. Mora-timed languages like Japanese behaved similarly to the syllable-timed class, while previously unclassified languages (Greek, Romanian, Mandarin, Welsh) filled the space in-between traditional classes.

Barbosa and Bailly [2] argue that listeners have an internal clock that needs to synchronize with the external clock of the stimulus (the beat of the perceived rhythm). They depart from Campbell's model, according to which any phoneme *i* inside a syllable has its duration determined by the same formula: $Dur_i = \exp(\mu_i + k*\sigma_i)$, where μ_i and σ_i are the mean and standard deviation of the log-transformed durations of the realisations of the phoneme *i*, and k is a factor calculated from the total syllable duration. *k*-factors are used to create gradual lengthening towards stress realization.

An experimental setup using French sentences generated either with or without a gradual pattern shows that the gradual pattern is preferred by listeners. The authors use the result to generate duration automatically, including the automatic generation of pauses when it is needed for the two presupposed clocks to synchronize.

3.5. Cultural aspects

Languages display particular rhythmic patterns both in relation to other languages and in relation to other standards of the same language. This observation has been investigated by Galves et al. [8] in a case study involving European versus Brazilian Portuguese.

Using corpora of newspaper articles written in the two dialects, they encode the texts according to rhythmic features and feed the encodings of the two corpora to Variable Length Markov Chains (VLMC) [5]. Each syllable is encoded based on two binary features: whether the syllable is stressed or not, and whether it is the beginning of a *phonological word* or not. A phonological word is a word together with the functional nonstressed words which precede it (e.g. "the boy").

Based on the property that each new symbol in a chain of rhythmic encodings is a function of its predecessors, and using the term *context* for the end string of its predecessors, together with Rissanen's [21] observation that the set of all contexts can be represented as the set of leaves of a rooted tree, the problem becomes a problem of context tree modeling and context tree selection.

The authors use the *context tree weighting* algorithm [23] to produce a set of *champion trees* for the two datasets. They devise their own method (the *smallest maximizer criterion*) to select a single tree from the set of candidates. The hypothesized rhythmic differences between European and Brazilian Portuguese is supported by the different champion trees obtained for the two corpora.

4. Experiments

4.1. Experimental setup

The second part of this paper is devoted to experiments on rhythm in written texts. First, we propose a set of rhythmic features, then we calculate these features on texts belonging to two different corpora: a corpus of famous speeches extracted from http://www.famous-speeches-and-speech-topics.info/famous-speeches/, and the raw texts from the RST-DT corpus of Wall Street Journal articles.

The values are calculated using the Python programming language, particularly the NLTK package for natural language processing, and the sqlite3 package for interfacing with SQL databases. Both the training set and the test set of the RST-DT corpus contributed to this analysis.

The relevant properties of the two corpora are listed in table 1.

Corpus	# of documents	# of sentences
Speeches	110	13762
RST-DT	380	7901

Table 1. Properties of the corpora used in the experiment

The workflow of this feature extractor is as follows:

- *Load and preprocess data*: the two sets of texts (corresponding to the two corpora) are loaded into two separate databases. Each sentence is parsed using the Stanford Parser (http://nlp.stanford.edu) and the resulting trees of constituents are added to the databases
- *Extract features*: for each document, the proposed features are computed and added to the respective database
- *Calculate overall statistics*: features of the two corpora are obtained from the features of individual documents.

4.2. Units

Prior to feature extraction, the text is segmented into units. The most natural unit is the sentence and it is used by a majority of features. The second type of unit used in these experiments is the fragment in-between punctuation markers.

For example, there are four such units in the sentence "Shall we expand(1), be inclusive(2), find unity and power(3); or suffer division and impotence(4)." This follows the intuition that punctuation functions as an offset (in musical terminology). In the remainder of this paper, we will call these units "punctuation units".

4.3. Features

We used five kinds of features: *organizational, lexical, grammatical, phonetical, and rhythmical*.

Each feature set is presented in Table 2.

Organizational. Organizational features capture the average word length, the length of units in either words or syllables, and patterns of length variation. The number of syllables is calculated using the CMU Pronouncing Dictionary (http://www.speech.cs.cmu.edu/cgi-bin/cmudict).

It is considered a good writing style to alternate long and short units. Rising, falling, or repetitive patterns, and the maximum length of such patterns are also of rhetorical interest. For frequent words in a document, we check whether they tend to appear in the beginning (first third), in the middle (second third), or at the end (last third) of units.

Lexical. Lexical features capture types of lexical repetition. Words or n-grams in a document are deemed frequent if their number of occurrences surpasses the value given by the formula *(text_length * threshold / n_gram_length)*. The threshold can be varied.

Filtering stop words is performed only in the case of frequent words because some of the stop words contribute to relevant n-grams. Duplicated units are identical units found within at most *delta* units of each other, again with a variable parameter *delta*.

The same parameter is used to count *anaphoras* (units starting with one or several identical words), *epistrophes* (units ending the same), *symploces* (units presenting a combination of the anaphora and epistrophe phenomena), *anadiploses* (a second unit starting the way a first unit ends), *epanalepses* (units starting and ending with the same word(s)).

We only count the maximal and non-redundant occurrences of these phenomena. That means that if *n_units* neighboring units start the same, that is considered to be a single anaphora.

If they share *n_words*, the anaphora is counted only once, not once for every initial substring of the maximal one. These values are calculated in turn for both sentence units and punctuation units.

Feature set	Feature
Organizational	Average number of syllables per word
	Average number of words per unit
	Average number of syllables per unit
	Percentage of rising/alternating/falling/repeating length structures
	Length of the maximum uninterrupted rising/alternating/falling/repeating series
	Percentage of frequent words located in the first/second/third part of a unit
Lexical	Average number of words deemed frequent per document
	Average number of n-grams deemed frequent per document
	Number of duplicated units, anaphoras, epistrophes, symploces, anadiploses, epanalepses – normalized by text length
Grammatical	Part-of-speech frequencies
	Number of commas normalized by number of words
	Percentage of sentence boundaries that are full stops, question marks, exclamation marks
	Number of neighboring syntactically parallel sentences – normalized by total number of sentences
Phonetical	Number of assonances, alliterations, rhymes – normalized by text length
Rhythmical	Percentage of units with an odd number of syllables
	Percentage of units with a stress on the final syllable

Table 2. Features of rhythm in written texts

Grammatical. We report the part-of-speech frequencies in order to check for affinities with different types of discourse. For the same reason, we report the frequency of commas and types of sentence boundaries (full-stops, question marks, and exclamation marks). Syntactic parallelism is detected only between neighboring sentences (located within a distance of given parameter *delta*). Parallelism can be checked only up to a given depth in the tree, but other than that it has to be pure or almost pure.

That means that, after deleting the terminal nodes (corresponding to actual English words), the remaining sentence trees should be almost the same. "Almost" means that corresponding nodes should be labeled with the same main part-of-speech category; another kind of noun, verb, adjective etc is allowed in place of a kind of noun, verb, adjective, but a noun cannot be in place of a verb, for example.

The following example from Jesse Jackson's speech "Common ground and common sense" illustrates this point. The nodes which differ but still fulfill the standard for syntactic parallelism are showed in boldface.

(ROOT
(S
(NP (PRP We))
(VP (VBP have) (NP (JJ public) (**NNS** accommodations)))
(. .)))
(ROOT
(S (NP (PRP We)) (VP (VBP have) (NP (JJ open) (**NN** housing))) (. .)))

Example 1. Two syntax trees marked for syntactic parallelism.

Phonetical. Similar to lexical features, phonetical features refer to stylistic devices based on repetition. An *assonance* is the repetition of a vocalic phoneme in a small amount of text. *Alliteration* is the analogous phenomenon for consonants.

A *rhyme* is defined as a repetition of the same sequence of several phonemes, not necessarily at the end of words. The distances and the required number of identical adjacent phonemes in a rhyme can be varied in the program.

Rhythmical. The complete stream of stresses (primary, secondary, or no-stress) is extracted from each document.

Following the theory that units having an odd number of syllables („odd units") and units ending in a stressed syllable make a text more dynamic, we use two features to calculate the frequency of occurrence of such units.

4.4. Results

Table 3 displays a comparison between the features' values for the two corpora. Significantly different values are showed in red color.

Feature	Speeches corpus	RST-DT corpus
Average number of syllables per word	1.475	1.587
Average number of words per sentence	22.474	21.543
Average number of syllables per sentence	33.186	34.861
% of rising/alternating/falling/repeating word length structures	0.18,0.62,0.17, 0.03	0.18,0.6,0.19,0.03
Maximum uninterrupted rising/alternating/falling/repeating series	5, 22, 5, 3	5, 14, 5, 3
% of rising/alternating/falling/repeating syllable length structures	0.18,0.62,0.18,0.02	0.17,0.61,0.2,0.02
Maximum uninterrupted rising/alternating/falling/repeating series	5, 22, 7, 3	5, 13, 5, 3
% of frequent words located in the 1st/2nd/3rd part of a sentece	0.155, 0.653, 0.192	0.279, 0.423, 0.298
% of frequent words located in the 1st/2nd/3rd part of a punctuation unit	0.158, 0.518, 0.323	0.271, 0.362, 0.367
Average number of words deemed frequent per document	39.909	53.037
Average number of n-grams deemed frequent per document	657.755	438.666
Sentence duplications/anaphoras/epistrophes/ symploces/anadiploses/epanalepses	0.0, 0.006, 0.002, 0.0, 0.001, 0.0	0.0, 0.005, 0.001, 0.0, 0.0, 0.0
Punctuation unit duplications/anaphoras/epistrophes/ symploces/anadiploses/epanalepses	0.0, 0.013, 0.004, 0.001, 0.001, 0.0	0.0, 0.005, 0.003, 0.0, 0.001, 0.0
Number of commas normalized by number of words	0.058	0.06
% of sentence boundaries that are '.', '?', '!'	0.919, 0.066, 0.015	0.992, 0.006, 0.002
Average number of syntactically parallel sentences per document	0.01	0.01
Assonances/alliterations/rhymes	0.047, 0.016, 0.003	0.013, 0.006, 0.003
% of units with an odd number of syllables	0.497	0.508
% of units with a stress on the final syllable	0.968	0.938

Table 3. Comparison between the two corpora

5. Discussion of results and future work

The main purpose of this feature extractor experiment was to study rhythmic differences between texts belonging to two very different categories: speeches and newspaper articles. We found several interesting results:

- *The small number of syllables per word*: This is a characteristic of the English language, but speeches (1.475 syllables/word) and newspaper articles (1.587 syllables/word) seem to be, on average, more compact than other textual categories. It should be noted that we counted in this statistics also stop words (which are mainly mono-syllabical)
- *Similar length of units and a strong preference for alternating short and long units*: This is true both for sentences and for punctuation units.
- *There are more words deemed frequent in newspaper articles*: This is probably due to business terminology, while in speeches repetitions have a rhetoric purpose (more interesting for rhythmic studies). The inverse property holds for frequent n-grams.
- *Frequent words and n-grams tend to be located in the middle of units*: Through repetition, these words become the main themes (or *voices*, in polyphonic terminology) of the text. This phenomenon is similar to the gradual pattern observed by Barbosa and Bailly [2] in stress realization. It seems that the receiver's attention needs to be repeatedly gained and released and the middle of units act as peak moments. This aspect is particularly exploited in speeches, where frequent words appear in the middle of sentences in 65% of cases (as opposed to 42% in newspaper articles).
- *The units considered are important*: For speeches, the number of anaphoras doubles when we consider punctuation units instead of sentence units. In future implementations, we intend to experiment with other kinds of units, such as EDUs (elementary discourse units obtained by own segmenter).
- *Interrogative and exclamatory sentences are a lot more common in speeches*: This was to be expected, and it is interesting to compare with values for other textual categories.
- *Lexical repetition is a much weaker separator than phonetic repetition*: There are a lot more assonances and alliterations in speeches, compared to newspaper articles. Anaphoras, epistrophes and similar features have only slightly higher values in speeches.
- *Similar values for syntactic parallelism correlates with similar values for lexical repetition*: In the future we intend to extend the notion of syntactic parallelism, to hold between smaller units of text.

6. Conclusions

This paper is firstly meant as an overview of the tools and methods of rhythm analysis in music and language. We observe similarities and differences between the two disciplines and we propose a set of rhythmic features for the evaluation of written texts. These features could be used in text type recognition, authorship recognition, measures of complexity, or evaluations of what is a well-written text. They also provide a valuable starting point for the identification of rhythmic patterns. Significant differences can be observed between the rhythm of speeches and the rhythm of newspaper articles.

R E F E R E N C E S

[1] Abercrombie, D. (1967) *Elements of general phonetics* (p. 97). Chicago, IL: Aldine.

[2] Barbosa, P., Bailly, G. (1994) Characterisation of rhythmic patterns for text-to-speech synthesis. *Speech Comm* 15:127-137.

[3] Beeferman, D. (1996) The rhythm of lexical stress in prose. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, UK.

[4] Boychuk, E., Paramonov, I., Kozhemyakin, N., Kasatkina, N. (2014) Automated Approach for Rhythm Analysis of French Literary Texts.

[5] Buhlmann, P., Wyner, A. J. (1999) Variable length Markov chains. *Ann. Statist.* 27 480–513.

[6] Chomsky, N., Halle, M. (1968) *The Sound Pattern of English*, Harper and Row, New York.

[7] Identifying Rhythms in Musical Texts. Christodoulakis, M.; Iliopoulos, C.S.; Rahman, M.S.; Smyth, William. (2008) *International Journal of Foundations of Computer Science* 1937-51.

[8] Galves, A., Galves, C., Garcia, J., Garcia, N., Leonardi, F. (2012): *Context tree selection and linguistic rhythm retrieval from written texts*. Ann. Appl. Stat. 6(1), 186–209.

[9] Greene, E., Bodrumlu, T., Knight, K. (2010) *Automatic Analysis of RhythmicPoetry with Applications to Generation and Translation*. EMNLP.

[10] Grabe, E., Low, E. L. (2002) Durational variability in speech and the rhythm class hypothesis. In C. Gussenhoven & N. Warner, *Laboratory phonology* (pp. 515–546). 7. Berlin: Mouton de Gruyter.

[11] Hebert, L. (2011) A Little Semiotics of Rhythm. Elements of Rhythmology. *Signo*. http://www.signosemio.com/semiotics-of-rhythm.asp

[12] Honing, H. (2002) Structure and interpretation of rhythm and timing. *Tijdschrift voor Muziektheorie [Dutch Journal of Music Theory]* 7(3): 227–232.

[13] Ladefoged, P. (1975) *A Course in Phonetics*. New York: Harcourt Brace Jovanovich.

[14] Lerdahl, F.,Jackendoff, R. (1982) A grammatical parallel between music and language. In: Clynes, M.: Music, Mind, and Brain. New York: Plenum. 1982 83-117.

[15] Lerdahl, F., Jackendoff, R. (1983) *A Generative Theory of Tonal Music*. MIT Press, Cambridge, MA.

[16] Liberman, M., Prince, A. (1977) On Stress and Linguistic Rhythm. *Linguistic Inquiry* 8, 249-336.

[17] A.D., Daniele, J.R. (2003) An Empirical Comparison of Rhythm in Language and Music. *Cognition*, 87: B35-B45

[18] Pike, K. L. (1945) *The Intonation of American English*. University of Michigan Press, Ann Arbor.

[19] Pressing, J. (1997) Cognitive Complexity and the Structure of Musical Patterns. *Proceedings* of the Fourth Conference of the Australasian Cognitive Science Society. Newcastle, Australia.

[20] Ramus, F., Nespor, M., Mehler, J. (1999) Correlates of linguistic rhythm in the speech signal. In *Cognition*, volume 73(3), pages 265–292.

[21] Rissanen, J. (1983). A universal data compression system. *IEEE Trans. Inform. Theory* 29 656–664.

[22] Thul, E., Toussaint, G.T. (2008) On the Relation Between Rhythm Complexity Measures and Human Rhythmic Performance. Proceedings of the 2008 C3S2E Conference, pp. 199-204

[23] Willems, F. M. J., Shtarkov, Y. M., Tjalkens, T. J. (1995) The context-tree weighting method: Basic properties. *IEEE Trans. Inform. Theory* 41 653–664.

[24] Yeston, M. (1976) *The Stratification of Musical Rhythm*. Yale University Press, New Haven, Connecticut.