Prosody in Print: Classifying Written Text by Rhythm and Sound

Nina Wang
Adviser: Christiane Fellbaum

Abstract

The notion of "style" is an important one as it relates to what makes certain types of written text appear to be different from other ones. While previous studies into the linguistic qualities that contribute to style have primarily examined lexical and text-based features, we propose that phonological features ? relating to the organization of sound in language ? are also a significant aspect of stylistic differences between texts. This paper looks specifically at prosodic rhythm, examining two measurements: the regularity of stressed beats (microrhythm) and the regularity of high/low pitch movements (macrorhythm). We hypothesize that when comparing creative and noncreative writings, the former is not only more rhythmic overall than the latter, but also has a higher degree of rhythmic variation. To evaluate how impactful these features are in differentiating texts, we use them to develop a classification model applied to three categories of text where stylistic differences seem most salient: poetry, prose, and research publications.

1. Introduction

Language is presented differently all around us. While content and meaning certainly vary greatly between writing in different contexts, the mere organization of words in a political speech, for example, appears to be quite dissimilar to that of an epic poem, which in turn seems nothing like that of a medical textbook. Texts that pay close attention to form seem somehow far more lyrical, melodic, and pleasing to the ear than texts that simply deliver function. Even changing the style in which a single sentence is presented can have drastic impact on its overall effect. In his popular writing handbook, The Elements of Style, Oliver Strunk takes for example the following famous quote by Thomas Paine:

These are the times that try men's souls.

He provides the following rewrites of it,

Times like these try men's souls.

How trying it is to live in these times!

These are trying times for men's souls.

observing that although the semantic meaning in each has remained exactly the same, something about the rewrites simply prevents them from achieving the same impact as the original. In this paper, we endeavor to clarify precisely what is meant by the notion of "style"—a variable that contributes so visibly to the differences between these four sentences of near-identical meaning, length, and complexity, not to mention between texts of entirely different genres. We propose that although lexical or semantic features are perhaps more obvious sources of differences between categories of text, phonological features are also a significant contributor to the disparity. In other words, we propose that, in different styles of written text, there are notable differences in the way sounds are organized throughout them.

Specifically, we examine the degree to which the sounds are organized in rhythmic patterns, and the degree of variety between the rhythmic patterns. We quantify this by looking at two prosodic metrics: microrhythm, the regularity of stressed and unstressed syllables, and macrorhythm, the regularity of high and low pitch tones. Because prosodic features—which consist of metrics like intonation, intensity, and duration—are phonological features that span across multiple segments and even beyond the sentence level (Wennerstrom 3), this allows us to better examine the way that they are impacted by different organizations of words and phrases.

Our choice to measure rhythm in particular is largely inspired by the fact that many creative works have a history of adhering to specific rhythmic patterns. Most forms of poetry depend on regular rhyme and meter—Shakespeare wrote his sonnets in iambic pentameter (alternating beats of short and long syllables), and Vergil's epics were in dactylic hexameter (alternating long, short, short). Haikus and limericks also follow regular patterns of line length. A cursory examination

of some of the most famous lines from literature reveal that even prose authors doubtlessly pay close attention to rhythm: Sylvia Plath in *The Bell Jar*, "I took a deep breath and listened to the old brag of my heart; I am, I am, I am;" F. Scott Fitzgerald in *The Great Gatsby*, "So we beat on, boats against the current, borne back ceaselessly into the past." Although these works do not follow any strictly prescribed pattern, they do seem to possess a sense of rhythm and even of melody. On the other hand, noncreative works like research publications, for example, are often marked by long sequences of monotonous sentences, difficult grammar, and limited structural variety.

This paper presents and examines the hypothesis that creative works, which tend to emphasize impact and lyricism, possess a higher degree of rhythm and also a higher degree of rhythmic variation. In contrast, noncreative works, which usually value information and content much more than style, have both a lower degree of rhythm and a lower degree of rhythmic variation. Furthermore, we also aim to assess the extent to which measures of prosodic rhythm can differentiate between categories of texts.

We examine three categories of text in particular, among which stylistic differences appear most significant: poetry, prose, and academic publications. We predict that these three texts will rank in the following way in terms of rhythmicity: poetry higest, then prose, and academic lowest. We predict that they will rank in this way in terms of rhythmic variation: prose highest, then poetry, and academic lowest.

2. Background and Related Work

The question of how to quantitatively measure stylistic differences between written text has been particularly relevant to the problem of authorship identification. In a 2006 study, Zheng et al. developed a framework for authorship identification of online messages by measuring four types of features: lexical, syntactic, structural, and content-specific. While their best-performing classification model achieved an accuracy rate of 97.69%, they had to take into account a total of 270 features, some of which include: total number of characters, total number of upper-case characters, average word length, frequency of punctuation, and five different measures of vocabulary richness []. Measuring

this many features is certainly more feasible and impactful when analyzing short online messages where author idiosyncrasies are more minute, but our research explores whether, in differentiating larger bodies of text, a great number of these features could be encompassed by only a few prosodic ones.

Because prosody encompasses characteristics relating to speech, studies of prosody have typically been applied to spoken rather than written language. This is particularly true when examining the way prosodic features change from one category to another. In 1991, Tench found that prosodic features tend to differ between different uses and contexts of speech (Wenner 7). For example, a speaker?s pitch and volume are likely to be more extreme in warning cries than in intimate conversation. However, the fact that in our study we are examining written rather than spoken text will not cause too much of a hindrance, since all words carry essential phonological properties that are accessible to readers fluent in the language []. Thus, we can still examine the phonological and prosodic qualities encoded in written words, using methods described in later sections.

We specifically study the rhythm of these prosodic features. Many diverse scholars suggest that rhythm is the underlying building block in phonology, and that rhythm is foundational to the stress patterns of our speech (49-50). Halliday notes that ?there is a strong tendency in English for the salient syllables to occur at regular intervals; speakers of English like their feet to all roughly the same length? (50). This is because English is a stress-timed language, which means that stressed syllables occur at regular intervals []. To illustrate this, in Figure 1, Wennerstrom provides an example utterance to illustrate how the feet are aligned in time []. Although the stresses (marked by arrows) are not perfectly regular, the intervals are indeed of roughly equal length.

This inspires our first metric for analyzing rhythm in text: the regularity of time intervals between stresses, referred to as ?microrhythm? []. Although the English language in general is expected to have approximate regularity of stresses, Halliday suggests that the degree of regularity is in different types of speech [], being particularly high in poetry and verse, but perhaps less so in other types of writing. Thus, we believe that a quantitative study of microrhythm will uncover meaningful differences between various styles of text.

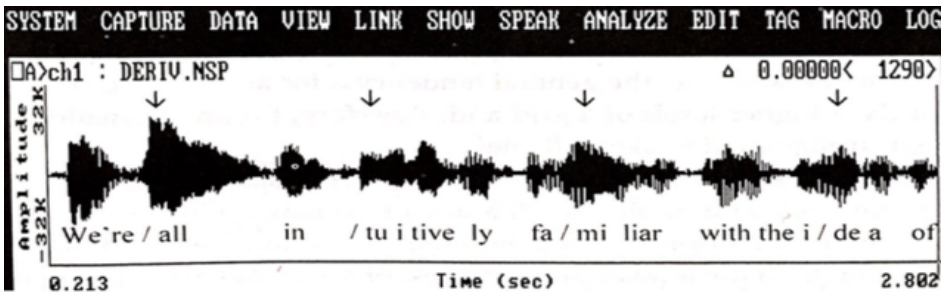


Figure 1: Waveform with stresses marked by arrows
Source:

In addition to examining stress patterns in text, we also examine patterns of intonation (pitch). Whereas the stresses in a sentence are just phonological properties of the lexical items in it, intonation provides additional information about discourse meaning?speakers assign pitch depending on meaning and context []. For example, a speaker would use very different intonation if they wanted to convey sarcasm. Just like microrhythm, the study of intonation rhythm was also proposed as method for classifying languages. Sun-Ah Jun proposed intonation rhythm—which she termed "macrorhythm"—as another approach to traditional methods of language typology (Jun 4). She proposes macrorhythm as a new way to approach prosodic typology, defining it as "a tonal rhythm characterized by the regularity of tonal pattern" (Jun 4). While microrhythm is concerned with the regular occurrence of stressed syllables, macrorhythm is concerned with regular changes of fundamental frequency (F0) in the pitch contour of the spoken utterance. Jun provides Figure 2, as an illustration of the difference between micro- and macrorhythm (524).

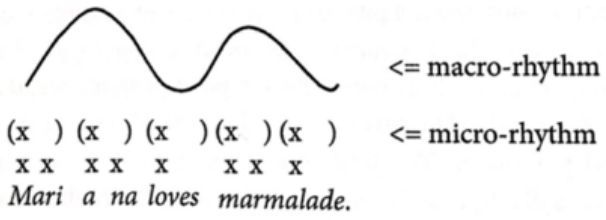


Figure 2: The difference between micro- and macrorhythm
Source:

Although macrorhythm was proposed as method to study prosodic differences between different languages rather than different styles of English text, we elect to examine it in this study because we believe that there are indeed meaningful differences in the pitch contours of different texts, and that

these differences are actually more pronounced than ones of syllable stress. Specifically, we wish to investigate whether movements in the pitch contour will provide insight into what contributes to the difference between ?melody? and ?monotony?. We predict that creative texts that tend to sound more melodic and lyrical will be full of pitch movements that are both frequent and regular. On the contrary, noncreative texts that seem more monotonous?a word which, appropriately, means to be of one tone?will perhaps contain fewer and more irregular pitch movements.

Our paper thus simultaneously presents prosody as a unique approach to the problem of textual classification, and textual classification as a previously unexplored application for prosodic and phonological study. As a broad summarization, this project is simply motivated by our curiosity about the extent to which the ?style? of different texts corresponds—consciously or unconsciously—to the rhythm produced by the sounds that compose of it.

3. Data

To examine the changes of prosodic rhythm across different types of texts, we wanted to choose a dataset composed of texts that were as stylistically disparate as possible. We decided to look at three different categories of texts—poetry, prose, and academic writing. Our final dataset included 120 texts in total: 40 in each category.

For the poetry category, which we intend to represent the most rhythmic end of the spectrum, we chose to use Shakespearean sonnets. The sonnets were taken from the Project Gutenberg dataset, a collection of 3,036 English books written by a total of 142 authors [].

For the prose category, we selected our texts again from the Gutenberg dataset. Because prose as a genre is comparatively much more fluid and diverse when it comes to structure, and we were not yet sure what kind of prosodic patterns we would find, we wanted to widen our lens of observation as much as possible by including a very diverse sampling of prose in our dataset. Thus, we included many different works across authors, genres, and time periods. Just an example of the works included are: *Pride and Prejudice* (1813), *Farewell to Arms* (1929); and President Obama's Farewell Address (2017).

Lastly, for the academic writing category, which we intend to represent the least rhythmic end of the spectrum, we selected research publications in the subject of physics. Just like for the poetry category, we did not feel the need to sample a lot of different variations of academic writing because we believe that there is little variation in the genre as a whole, and so publications in one subject can therefore reasonably represent the stylistic pattern of publications in general. We selected most of these texts from the Papers in Physics dataset [], and some from Nature magazine.

4. Methodology

4.1. Preliminary Processing

Because prosody is linked to spoken rather than written text, we must first convert our texts into audio format before we can begin looking at their prosodic features and measuring their degrees of micro- or macrorhythm. This is similar to the process of reading?as readers, we also translate written words into a voice that we ?hear? inside our heads. To do this, we use Amazon Polly, a text-to-speech software, to ?read aloud? each piece of written text and output the resulting audio recordings. Rather than sending the entirety of a text file to Polly, we actually send it one sentence at a time. We do this primarily because the character counts of our texts in their entirety simply exceed the 1500-character-per-request limit enforced by the software. However, this actually proves beneficial because it trims away of the extra silences between sentences, leading to less noise in the data for our analyses.

Obtaining these auditory readings of texts allows us to directly examine their prosodic qualities, since we now have accessibility to information like pitch, intensity, and duration through the waveform. We examine the actual waveform using Praat, a software that provides rich functionality for speech analysis and easy visualization of relevant prosodic features [].

4.2. Microrhythm

To restate, microrhythm is the degree to which stressed syllables are regularly aligned in time. In order to calculate the degree of microrhythm of a text, therefore, we first need to identify which

syllables are stressed. In the next two sections, we describe two attempts at doing this.

4.2.1. Initial Attempt Our initial attempt to determine stressed syllables actually did not require any text-to-speech conversion at all. We simply used the CMU Pronouncing Dictionary [], which provides information on pronunciation and lexical stress for a total of over 134,000 individual words. In the dictionary, each word is associated with a list of its phones. The phones that are identified as syllables are marked with a number: 0, 1, or 2. For example, this is the dictionary output for the word "intuitively":

IH2 N T UW1 IH0 T IH0 V L IY0

A marking of 0 means that the phone is not stressed; it is merely a syllable. A marking of 1 means that the phone has primary stress. A marking of 2 means that the phone has secondary stress, which means that most of weight in the word is assigned to it. Because secondary stresses seem less frequent than primary, for our purposes, we treat stress as a binary property?either a phone is stressed or not. A phone with no marking or a marking of zero would be considered not stressed; a phone with a marking of either 1 or 2 would be. Using this method, to determine the stressed syllable of a particular piece of text, we would simply use the dictionary to look up each individual word in the text one-by-one, and consider its phones stressed or not stressed depending on the output. One source of error that we encounter at this step is that the dictionary does not contain all possible words. So, when we encounter a word that it cannot not provide stress information for, we just forgo analysis on that word and assume that it contains no stresses. This is definitely a source for error, so we try to manually adjust as many unfindable words as possible?for example, splitting unrecognizable compound words into two recognizable ones.

While this approach is very simple to understand and easily computable, the results that it yields are not entirely satisfactory. This has to do with the fact that because each word is looked up individually in the dictionary, they are not being considered in context of the whole sentence. As a result, we end up with a lot of false positives?many words, such as "I" or "am" or "be," indeed contain a stressed syllable if uttered on their own, but when considering them as part of a larger

unit such as a phrase or a sentence, they are no longer considered stressed on that level, as the areas where we place intonation inevitably changes with context. For example, Figure 3 illustrates the difference between expected stress and individual word stress for the first two lines of Sonnet 18. The bolded areas are considered by the CMU Dictionary to contain stress. The underlined areas of words are where stresses should be, according to iambic pentameter. Although iambic pentameter certainly doesn?t align perfectly with natural speech, one can still see that there are quite a few bolded areas that don?t seem like they would be stressed if read aloud.

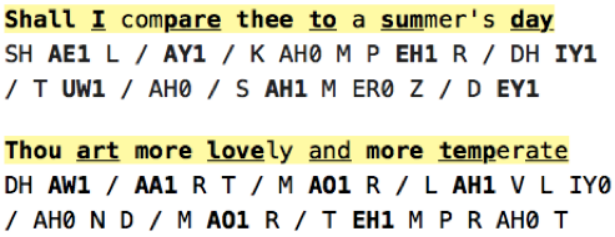
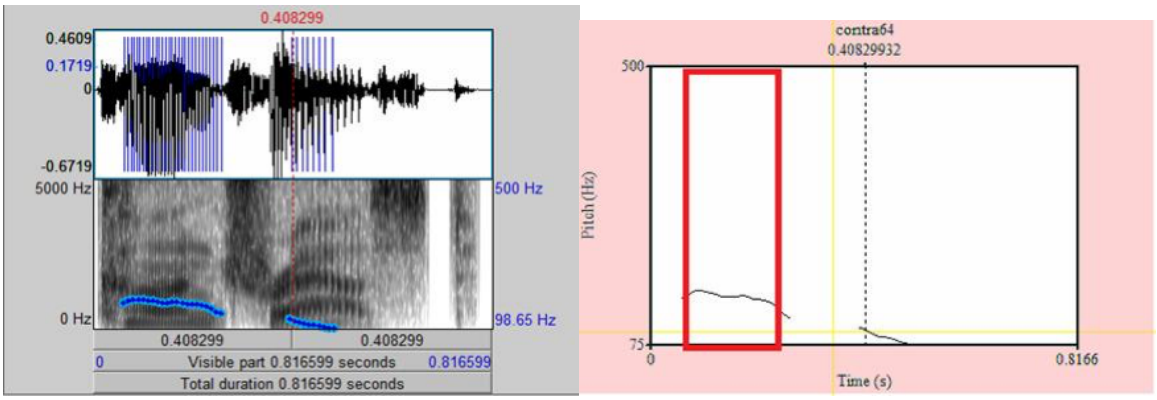


Figure 3: Dictionary stress versus expected iambic pentameter stress

4.2.2. Improved Attempt The main problem with the previous attempt was that there were too many false-positive marks of stress. We determined that better accuracy would come with taking words in context with the entire sentence. Then, we would be able to look the way each syllable compares with all others of the same sentence, and then determine stress from there. As [] notes, there are two types of stress ? word stress and sentence stress. Word stress is concerned with the stressing of individual words when they are pronounced in isolation, but we are interested in the stress on the sentence level. To distinguish stressed and unstressed syllables on a sentence level, we must examine their prosodic features when turned into speech. Roach specifies that stressed syllables all something called prominence, which is produced by four main factors: duration, pitch, loudness (intensity), and quality. These features generally work together, but the strongest effects are produced by pitch and duration; loudness and quality have much less effect []. Stressed syllables are marked by a longer duration, higher pitch, greater intensity, and higher quality. This can be seen in Figure 4, which shows the waveform (a) and the pitch contour (b) of the word “contrast,” stressed on the first syllable. The separation of the two syllables is marked by the dotted vertical line in the



(a) (b)

Figure 4: Waveform (a) and pitch contour (b) of "CON-trast"
Source:

pitch contour. As we can see, the first syllable has a higher pitch and longer duration, showing that it is the stressed one.

We thus perform prosodic stress analysis on a sentence level to examine the way phones interact with each other. This is done by applying text-to-speech to the entire sentence and then studying the resulting waveform. In this attempt, our purpose in adding another layer of analysis is to improve upon and correct the previous method, which only considered stresses on a word level. Thus, for this implementation, we endeavor to examine prosody to ?correct? the false-positives from the CMU Dictionary on a sentence level. We do not aim to correct the false-negatives, because there are much fewer of them—in English, all content words with two or more syllables have at least one stressed syllable [uni-bamberg] and the CMU Dictionary marks all of them. Occurrences of false-negatives would occur if there is special context, such as emphasis or contrast, that causes a non-content word to be stressed. However, the placing of this kind of stress relies on detecting and understanding aspects of semantic context such as emphasis, contrast, incredulity, sarcasm, etc. which our text-to-speech software is not advanced enough yet to do satisfactorily.

Our method for correcting false-positives is as follows: examine the prosodic data for every syllable (stressed or unstressed) found in all words of the sentence, assign each one prominence scores relative to each other, and then correct any syllables that are marked by the CMU Dictionary as stressed but whose prominence actually ranks amongst the ones marked as unstressed. The

intuition is that if a syllable is listed as stressed, but not actually spoken with the acoustic markers of stress, then it is a false-positive error and should actually be marked unstressed.

First, we gather all the phones of all words in the sentence, and determine which of them are marked by the CMU Dictionary as syllables. Then, we convert the sentence to audio and analyze the waveform in Praat. We are only interested in examining the phones considered to be syllables (so for the word ?intuitively,? only the [IH2, UW1, IH0, IH0, IY0] phones), so we need to locate where they are in the waveform. We accomplish this by using a forced aligner called Gentle []. Given an audio recording and a text transcript, Gentle outputs a JSON file of start and end times for each word and each phone in the words found in the audio file. We then parse this JSON file of timestamps to produce a dictionary of our own, whose keys are all the phones, and whose values are their start/end times, whether they are syllables, and if so, whether or not the CMU Dictionary lists them as stressed. There is some error correction required at this step. Sometimes Gentle does not recognize a certain word in the audio or is not able to find a certain word from the transcript in the audio, and thus is unable to provide the timestamps for it. In this case, we simply forgo analysis for this word and accept the stress information provided by the CMU Dictionary.

We then transform our final dictionary of phones into a Praat TextGrid object so that we can draw boundaries in the waveform for the start/end times of all the phones found in it. With this, we then collect pitch, duration, and intensity data for each syllable. We calculate a prominence score for each syllable, representing how prominent it is in relation to all the other ones. Because prominence score combines different units, in order to prevent features that are quantified with larger numbers from overpowering those with smaller ones, we first scale the data for each feature on the interval [0, 1]. Moreover, since these three features do not contribute to stress equally, we use the following equation for weighing and combining them:

$$prominence\ score = 0.75 \times duration\ +\ 0.15 \times pitch\ +\ 0.10 \times intensity$$

where duration is measured in seconds, pitch is measured in Hertz, and intensity is measured in

decibels. We determined that this particular scaling yielded most accurate results. An example of the way prosodic data for each syllable is converted to a prominence score is found in Figure 5.

Syllable	Hz	dB	s	score
"Thou_EH1"	[165.82,	72.58,	.15]	.96
"Art_AA1"	[171.86,	71.55,	.01]	.15
"More_AO1"	[153.77,	73.47,	.08]	.54
"Lovely_AH1"	[171.78,	70.07,	.07]	.54
"Lovely_IY0"	[135.86,	71.84,	.08]	.46
"And_AE0"	[132.09,	70.20,	.04]	.18
"More_AO1"	[153.87,	72.63,	.07]	.46
"Temperate_EH1"	[125.92,	71.92,	.09]	.54
"Temperate_AH0"	[123.13,	68.93,	.10]	.50

Figure 5: Pitch (Hz), intensity (dB), and duration (s) measurements and resulting prominence score for each syllable

We also calculated a prominence score threshold, below which any syllable is to be considered unstressed,

$$threshold = med\_score - (0.15 \times sd\_score)$$

where *med_score* represents the median prominence score of all syllables in the sentence, and *sd_score* represents their standard deviation. Intuitively, this equation says that any syllables that scored below the median by a measure of .15 times the standard deviation is deemed to have too low a prominence to be considered stressed. We change all syllables that scored below the threshold to unstressed. In the example of Figure 5, the phones "Art_AA1" and "More_AO1" would be changed to unstressed.

4.2.3. Rhythm With information on where the stresses are, we can finally calculate microrhythm—the regularity of time intervals between each stressed syllable. We look at each text file on a sentence-by-sentence basis for previously described reasons, taking measurements of both the average microrhythmicity within each sentence, and the variation of microrhythmicity across the sentences. We first define *micro_var* as the standard deviation of time intervals between stressed syllables within a single sentence, scaled by its mean time interval. Then, the average

microrhythmicity within sentences, *micro_within*, is defined as:

$$micro\_within = mean\ of\ (micro\_var\_scores)$$

where *micro_var_scores* is an array containing each sentence's *micro_var* value. Next, we examine the variation of microrhythmicity across sentences (*micro_across*):

$$micro\_across = \frac{standard\ deviation\ of (average\_intervals)}{mean\ of (average\_intervals)}$$

where *average_intervals* is an array containing each sentence's average interval length between stresses. We scale the standard deviation by the mean to provide understanding of the number in context. Examining both of these measurements allows us to see to what extent the text 1) is generally microrhythmic, and 2) has a healthy amount of variation as well. Since we are looking for regularity of intervals, it is important to note that the higher the value of *micro_within* and *micro_without*, the lower the degree of *micro_rhythm*.

To standardize things as much as possible, we measured the same number of sentences for each text. However, there was the concern of sentence length?whether disparities of macrorhythm scores would largely be caused by the fact that shorter sentences would have fewer stressed syllables and thus a lower standard deviation of the intervals between them. We plotted sentence length versus microrhythmicity for 5 texts in each genre (totaling about 275 sentences) in Figure 6, and find that the R^2 value of 0.181 represents only a very weak positive correlation. Thus, we determine that changes in microrhythmicity score are not explained by sentence length to a meaningful degree.

4.3. Macrorhythm

4.3.1. Identifying Rising/Falling Trends The calculation of macrorhythm depends on the regularity of High/Low pitch movements. This requires access to prosodic data, so we again apply text-to-speech on the texts one sentence at a time, and then examine the waveforms in Praat. Specifically, we are examining the pitch contour of each sentence, so we use a Praat script that produces the

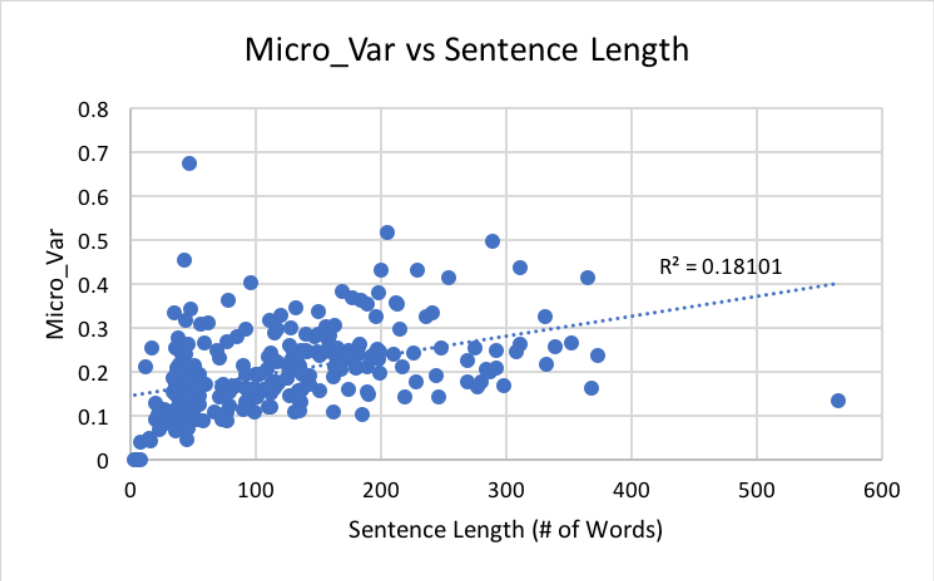


Figure 6: Scatterplot of the length of a sentence versus its microrhythmicity score

pitch listing by sampling the f0 of each frame present in the waveform. Once we have an accurate pitch listing, we then locate and examine the places where the pitch is rising and where it is falling. There is an established system for transcribing the intonation patterns and other aspects of prosody for English utterances, called ToBI (for Tones and Break Indices) []. This is a labelling convention that marks the various different types of tones within the English languages: phrasal tones (L-, H-, L%, H%), assigned at every intermediate or intonation phrase, and pitch accents (H*, L*), assigned at every accented syllable []. We choose not to use ToBI annotation for our purposes; the first reason is that it requires manual annotation, as current automatic ToBI annotation software did not appear to be very accurate. Secondly, ToBI is more complex than necessary; Jun?s description of macrorhythm only takes into account High or Low tones, not what type of High or Low tones they are.

Thus, the primary challenge is to produce an algorithm that accurately identifies the relevant trends in the pitch listing. This is difficult because the pitch contours are not perfectly smooth; oftentimes local maxima/minima are present that do not represent the most meaningful local maxima/minima for the entirety of the pitch contour. We therefore develop an algorithm to determine the meaningful pitch trends: first, we capture every single interval of rising and falling slope, regardless of how short or flat it is, and then we “smooth out” these trends according to the following general criteria:

1. *If the magnitude of the slope of a trend is too low, then merge it with the previous one.* This is so that if a period of slightly negative slope follows a period of very positive slope, for example, then it is more accurate for our analysis to consider the entire thing as one trend in the positive direction. We determine our slope cutoff to be $\frac{1}{6}$ of the average slope magnitude between all points in the pitch listing. We take this into account so that the cutoff for each sentence is dynamic, changing depending on the local environment; the cutoff is higher for pitch contours that change very drastically, and lower for flatter ones.
2. *If the duration of a trend is too low, then merge it with the previous one.* This is so that we do not count very short trends as individual trends; they are more accurately considered as part of a longer trend. The time cutoff is determined to be $\frac{1}{3}$ of the average duration of silences in the sentence, again to allow the measurement to be dynamic.
3. *If two adjacent trends are of the same direction, and that there is only a small change in pitch between the end of the first trend and the start of the next trend, then merge them together.* This is taken into account due to the fact that there are breakages in the pitch contour when the speech is not continuous. If there is a falling slope that lands at 120 Hz, a pause, and then another falling slope that starts at 125 Hz, it is more accurate for us to consider the two of them as one falling slope.

Once we have cleaned up and smoothed out the trends so that we only have the most significant pitch changes, we can label the High and Low points of the contour. We do this by going through each trend?if it a rising slope, then we mark the highest point as a High tone; if it is falling, then we mark the lowest point as Low tone. Figure 7 shows an example of our final labeling of a pitch contour. As we can see, although there are many interruptions in the contour, our algorithm picks out the most significant peaks and valleys.

4.3.2. Rhythm The process for determining macrorhythm is slightly more complicated than determining microrhythm, because there are more aspects that factor into it. Jun specified three in particular []:

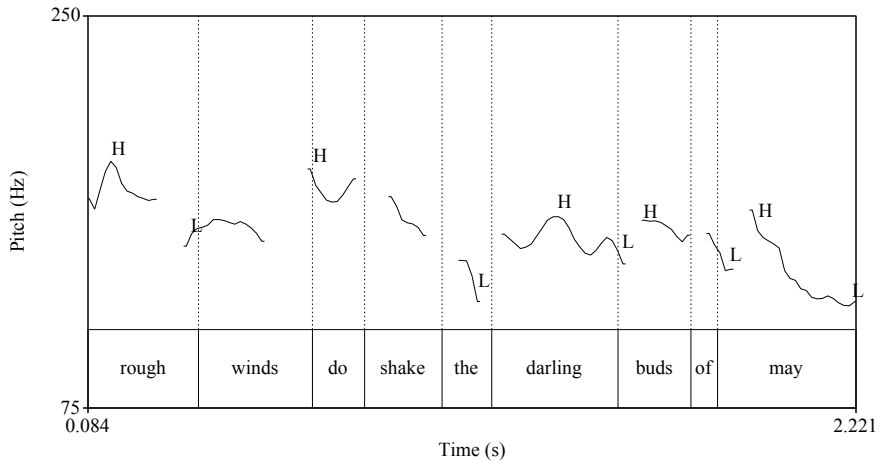


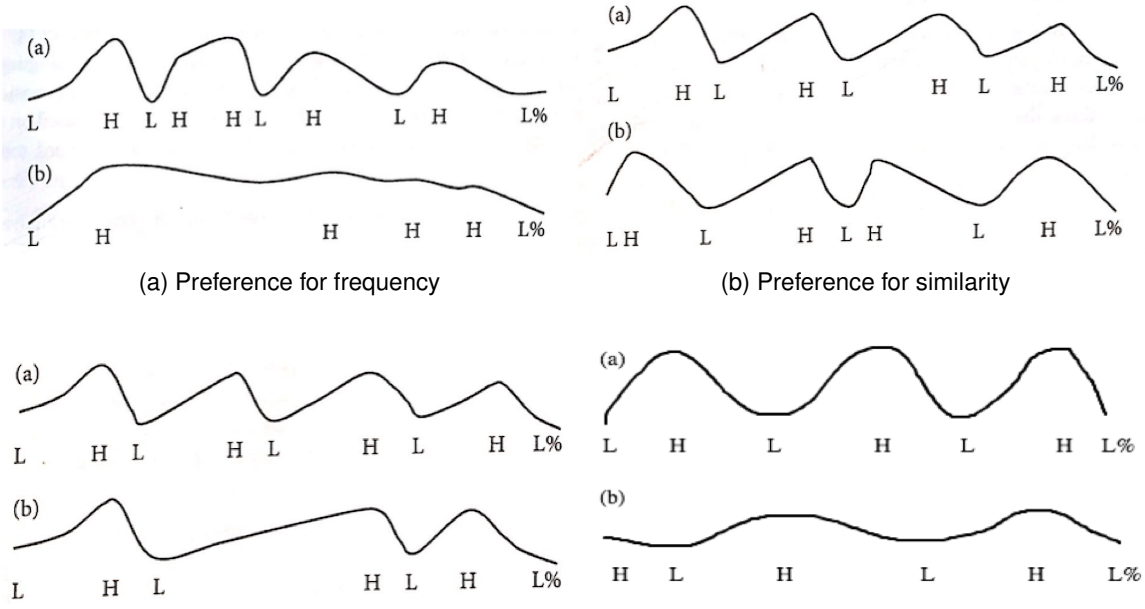
Figure 7: Our algorithmic labeling of High (H) and Low (L) tones in a pitch contour

1. *Frequency:* A pitch contour with more H/L alternations has stronger macrorhythm than one with fewer alternations. For example, in Figure 8a, pitch contour (a) is more macrorhythmic than (b).
2. *Similarity:* A pitch contour whose H/L alternations are more similarly-shaped has stronger macrorhythm than one whose alternations are more irregularly-shaped. Figure 8b, pitch contour (a) is more macrorhythmic than (b).
3. *Regularity:* A pitch contour whose H/L alternations are more regularly spaced out has stronger macrorhythm than one whose alternations are more irregularly spaced out. Again, in Figure 8c, pitch contour (a) is more macrorhythmic than (b).

On top of Jun's specifications, we also add one more:

4. *Disparity:* A pitch contour whose H/L alternations have larger magnitudes of difference has stronger macrorhythm than one whose alternations have smaller magnitudes of difference. In Figure 8d, we determine that pitch contour (a) is more macrorhythmic than (b).

While our preference for greater disparity in pitch doesn?t add to the ?rhythm? aspect of macrorhythm, we include it because of the information that it provides about melody versus monotony. A pitch contour that has very high H tones and very low L tones is quite melodic, while a pitch contour whose H and L tones are more similar in pitch is considered more monotonic. Now, we combine the above four metrics to determine a score of macrorhythmicity for a pitch contour.



(a) Preference for frequency

(b) Preference for similarity

(c) Preference for regularity

(d) Our preference for disparity

Figure 8: Rules for macrorhythm

We first considered Jun's formula [] for combining her three metrics:

$$MacR\_Var = SDp + SDv + SDr + SDf$$

Where SDp = standard deviation (SD) of length of peak-to-peak intervals; SDv = SD of length of valley-to-valley intervals; SDr = SD of rising slopes; and SDf = SD of falling slopes. The lower the $MacR\_Var$ score (i.e. the more regular everything is), the stronger the macrorhythmicity.

However, we consider this formula flawed for a few reasons. First is that metric 1, frequency, is not included in this formula. Secondly, the fact that none of the variables are scaled means that the ones are measured in Hertz (SDr and SDf) are much greater in magnitude and thus greatly overpower the variables measured in seconds (SDr and SDf). Lastly, measuring standard deviation runs into some issue when there is only one peak-to-peak interval, for example. Although the standard deviation is indeed zero when there is only one data point, this does have the unwanted effect of greatly lowering the scores of some pitch contours. In addressing these problems and

incorporating our fourth metric, we adapt Jun's formula into the following:

$$macro\_var = \frac{1}{freq\_score} + sim\_score + reg\_score + \frac{1}{disp\_score}$$

where

$$freq\_score = \frac{num\ of\ H/L\ alternations}{time},$$

$$disp\_score = \frac{avg.\ slope\ magnitude\ between\ H/L\ points}{num\ of\ H/Lpoints},$$

$$sim\_score = \frac{SDr + SDf}{2}$$

if neither SDr or SDf is zero, otherwise just $SDr + SDf$, and

$$reg\_score = \frac{SDp + SDv}{2}$$

if neither SDp or SDv is zero, otherwise just $SDp + SDv$. We scale all the SD measurements by dividing them by their mean values.

We take the inverses of $freq\_score$ and $disp\_score$ because in this formula, a lower *macro_var* indicates greater macrorhythmicity. Since both higher frequency and higher pitch disparity suggest stronger macrorhythmicity, we had to invert them in order to properly add them to the other metrics where a lower number indicates stronger macrorhythmicity. Moreover, the optional division for *sim_score* and *reg_score* is intended to take a sort of average of the two SD measurements that make it up, as a way to offset the effect of the presence of a zero in one of them. If both SD measurements are zero, then we simply accept the score for that metric as zero.

Because we are measuring standard deviations, there was again the concern that *macro_var* would be simply predicted by the length of the sentence. We then plotted sentence length versus macrorhythmicity, again for the same 5 texts in each genre (totaling about 275 sentences) in Figure 9, and find that the R^2 value of 0.094 represents another very weak correlation.

We actually do expect that *macro_score* would have a positive correlation with sentence length.

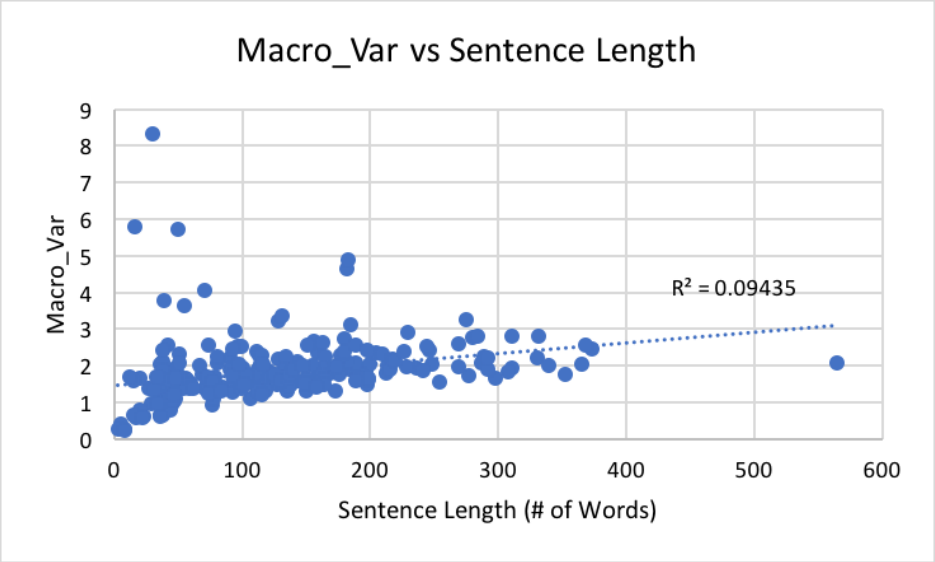


Figure 9: Scatterplot of the length of a sentence versus its macrorhythmicity score

This is because a lot of prosodic features are impacted by their position in the sentence. For example, the stressed syllable of the first content word is usually longer, louder, and higher pitched. From there, the air pressure diminishes so that later syllables tend to be shorter and of lower volume and pitch [wenner p. 50]. Thus, if the peaks and valleys of a pitch contour gradually become shorter in length and lower in magnitude, they therefore become overall less regular, so it makes sense that longer sentences have a higher *macro_score* (and are less macrorhythmic). However, we determine that the correlation is again too weak for the changes in *macro_score* to be explained by sentence length.

Now that we have a way of determining the *macro_score* for individual sentences, we can analyze the macrorhythmicity of the text as a whole. For each text, we again measure both the average macrorhythmicity within each sentence, and the variation of macrorhythmicity across the sentences. The average macrorhythmicity within sentences, *macro_within*, is defined as:

$$macro\_within = mean\ of(macro\_scores)$$

where *macro_scores* is an array containing the *macro_score* of each sentence. Next, we examine

the variation of macrorhythmicity across sentences, *macro_across*:

$$macro\_across = \frac{standard\ deviation\ of\ (macro\_scores)}{mean\ of\ (macro\_scores)}$$

where we again scale the standard deviation to give it some context. Just as we did for microrhythm, we keep things uniform as much as possible by measuring the same number of sentences for each text.

5. Results

5.1. Microrhythm Results

To reiterate, we took two microrhythm measurements for each text: the average microrhythm within sentences, and the degree of variation of microrhythm across sentences. We plotted these two metrics (*micro_within* versus *micro_across*) for all the files in our dataset in Figure 10.

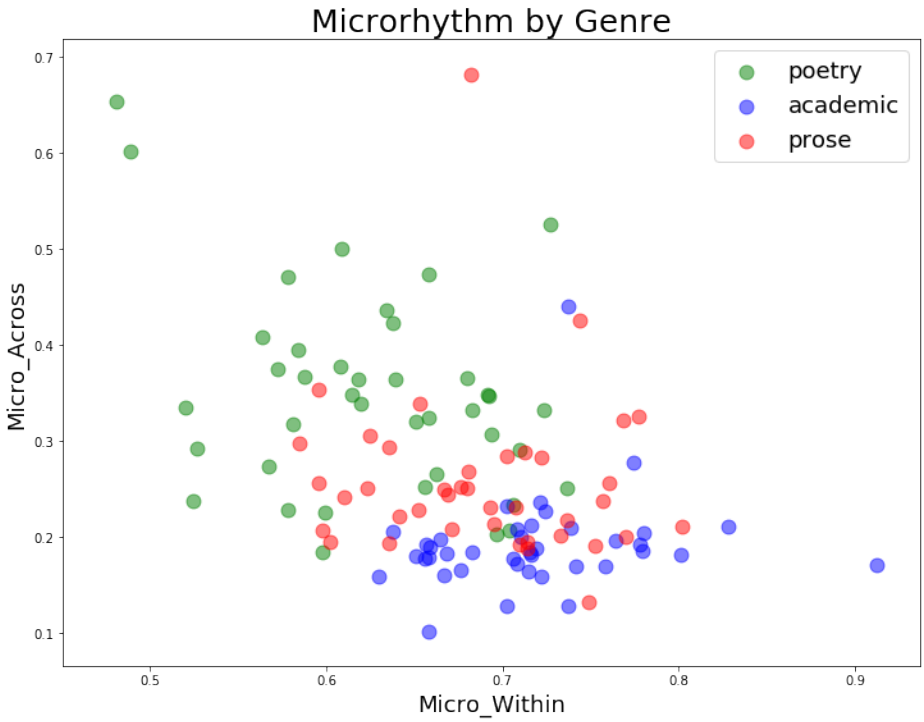


Figure 10: Scatterplot of texts according to microrhythm measurements

As we can see just by glancing at the plot, there is a fairly evident split between poetry, academic,

and prose. The separation between poetry and academic seems to be much more distinct than the separation both between prose and poetry and between prose and academic. Table 1, which gives the average distances between each category, supports this observation. Additionally, this observation is in line with our expectation that poetry would lie on one extreme of the microrhythmic spectrum, while academic would lie on the other end. Moreover, the prose category is sandwiched between poetry and academic, suggesting that when it comes to microrhythm, prose shares certain similarities with both. This is much more the case for *micro_across* than for *micro_within*, which visibly does not provide as powerful of a delineation between the three categories. *Micro_across* separates the categories much better, suggesting that, in terms of microrhythm at least, the three categories of text differ more in the way their sentences vary between one another rather than in the quality of the sentences themselves. Again, prose falls in the middle for *micro_across*.

	Poetry	**Academic**	**Prose**
Poetry	0	0.17985	0.10741
Academic	0.17985	0	0.07319
Prose	0.10741	0.07319	0

Table 1: Average microrhythm distance from each genre to the others

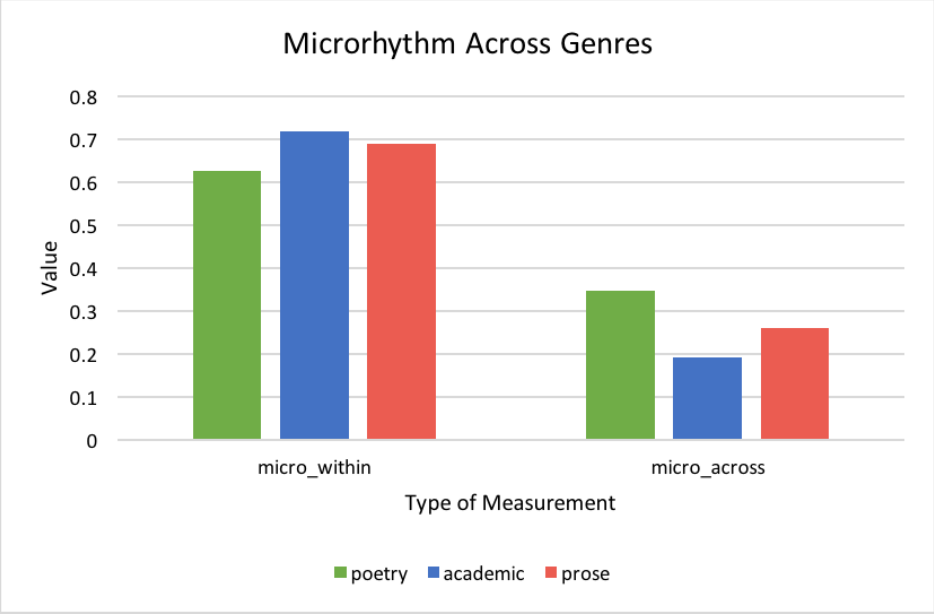


Figure 11: Bar graph of microrhythm measures for each genre

We also present a summation of these findings in Figure 11. For *micro_within*, poetry scores lowest on average, which denotes that its sentences generally have the lowest variability in its time intervals between stresses, and therefore the greatest degree of microrhythmicity. Academic scores the highest, meaning that its sentences are the least microrhythmic; its stresses are the least evenly spaced out. Prose scores in the middle. These results do support our original hypothesis; we expected that since sonnets follow a regular stress pattern, that we would find that the stresses are evenly spaced out. On the other end of the spectrum, we also expected that academic would be the least regimented, and so would have a more irregular pattern of stresses. It is also unsurprising that prose ranks in between these two, since prose is neither as strictly regimented as poetry, nor is it as scattered as academic writing.

For *micro_without*, we see that poetry actually contains the greatest amount of variation between sentences, academic contains the least, and prose again somewhere in the middle. The ranking of academic and prose both support our hypothesis, since we expected academic writing to have little structural variety, and for prose to definitely have more variety in comparison. The position of poetry, however, is somewhat surprising. We expected poetry to be similar to academic in that because it is so strictly regimented, there would be little variation across sentences.

5.2. Macrorhythm Results

For macrorhythm, we again took two measurements for each text: average macrorhythm within sentences, and the variation of macrorhythm across sentences. Figure 12 represents a scatterplot of *macro_within* versus *macro_across* values for all the files in our dataset.

Right away, we notice that plotting by measures of macrorhythm seems to do even better in separating the categories than does plotting by microrhythm. This which will be further explored when looking at the classification models that each one produces. Again, poetry and academic are much more disparate than prose/poetry or prose/academic. Table 2 shows the average distances between each category. This provides more supports our expectation that poetry would lie on opposite ends of the macrorhythmic spectrum. Prose is sandwiched between poetry and academic

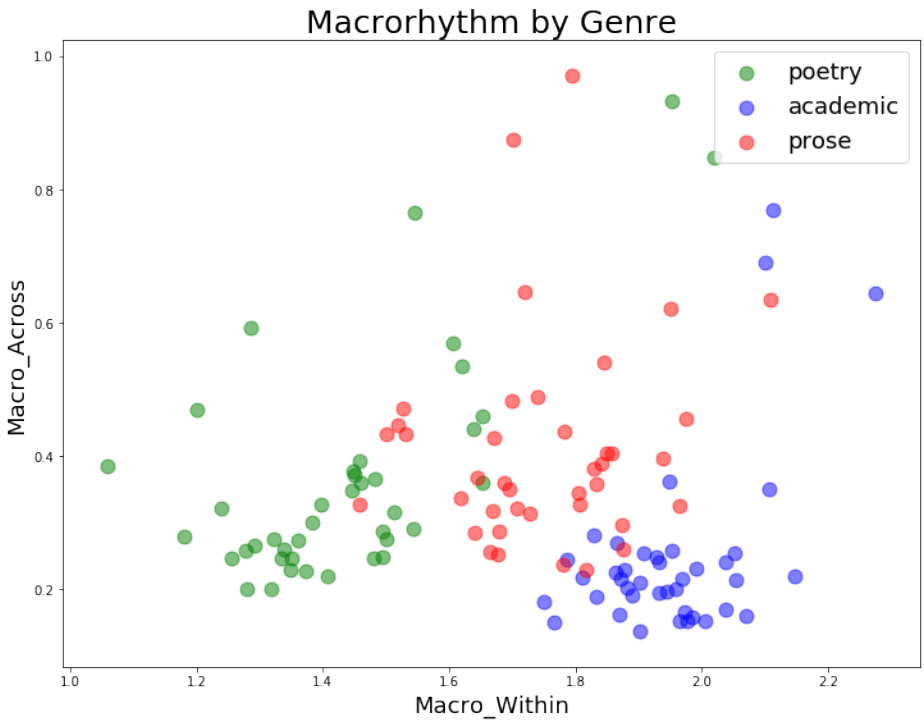


Figure 12: Scatterplot of texts according to macrorhythm measurements

here as well.

The across-sentence measure, *macro_across*, definitely plays a role in separating the categories here just like it does for microrhythm. Importantly, it appears to be quite outshone by *macro_within*, the measure of average sentence macrorhythm. One potential explanation for this difference is that macrorhythm is simply a better metric for capturing stylistic differences, and so it is able to capture the differences on the level of individual sentences, which is perhaps more nuanced than simply looking at the degree of across-sentence variation.

	Poetry	**Academic**	**Prose**
Poetry	0	0.52807	0.31726
Academic	0.52807	0	0.25878
Prose	0.31726	0.25878	0

Table 2: Average macrorhythm distance from each genre to the others

A summarization of the findings in the scatterplot is presented in Figure 13. For *macro_within*, poetry again scores lowest on average, showing that its sentences are the most rhythmic when examining both stresses and pitch movements. Academic also again scores the highest, showing

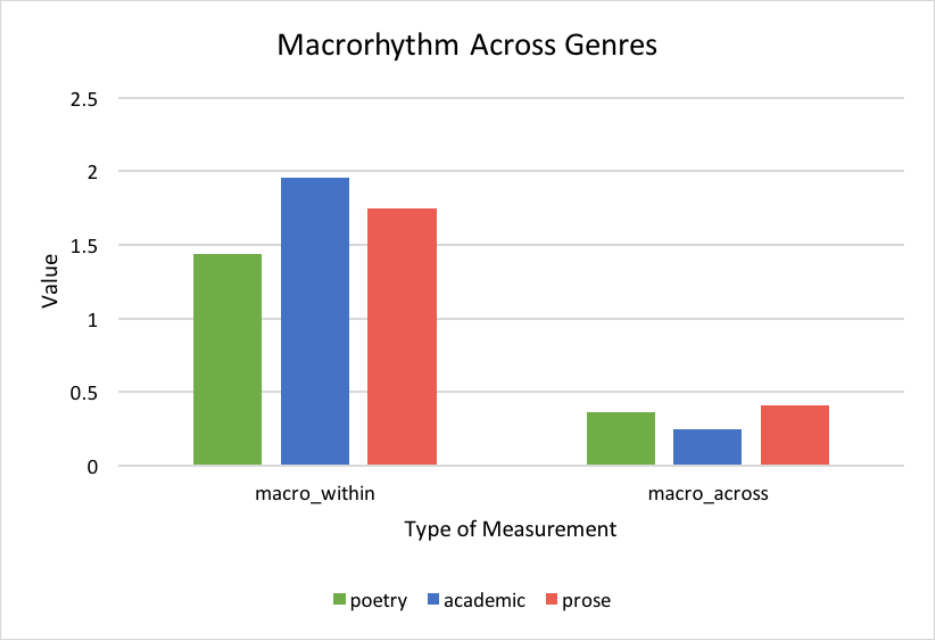


Figure 13: Bar graph of macrorhythm measures for each genre

that its sentences are the least rhythmic on both of these fronts. These results support our original hypothesis; we expected the strictly regimented sonnets would have regular pitch movements on top of regular stress placement, while academic would be the most disorganized and so would have fewer, weaker, more irregular pitch movements. It is also unsurprising that prose ranks in between these two, as prose text can be very rhythmic and melodic in some sections, such as flowery descriptions, but fairly ordinary in others, such as plain dialogue.

The results for *macro_without* also support our predictions. We see that prose contains the greatest amount of variation between sentences this time, academic the least, and poetry somewhere in the middle. This is the ranking that we expected; we expected prose to be a sort of amalgamation of some highly rhythmic sentences and some not so highly rhythmic ones, causing there to be a lot of variation. Meanwhile, we predicted that both poetry and academic texts would only have in general one "type" of sentence—either very rhythmic or not very rhythmic, thus yielding lower variation. With this, however, we still expected poetry to score just above academic because it is still a creative text, where we expected there to be enough attention paid to style that the texts avoid sounding monotonous.

5.3. Classification Models

Now, we evaluate the significance of our findings by using them to set up classification models and observing how well the models are able to use our features to distinguish between types of texts. In our case, we choose to use support vector machines (SVM), which are a supervised learning methods for classification that are versatile and effective in spaces with many dimensions []. Analysis of the our SVM accuracy is completed by looking at the resulting confusion matrices. We built our SVM using the LIBSVM library [] and their out-of-the-box software that automatically performed many preprocessing steps such as scaling data and selecting kernel functions. The most important preliminary step that we complete is dividing up our data into a training and testing set. After comparing the performances of splitting the data in 50-50, 60-40, and 70-30 ratios, we settle on reserving 70% of our data for training and 30% for testing.

We build 3 classification models: one for features of microrhythm, one for features of macrorhythm, and one for the two combined. For each one, we first evaluate their performances in distinguishing between all three categories of text, and then in distinguishing just between creative and noncreative text. A summary of the accuracy rates for these are given in Table 3.

	Poetry/Academic/Prose	**Creative/Noncreative**
Micro	69.05%	83.33%
Macro	80.95%	88.10%
Micro + Macro	83.33%	85.71%

Table 3: Accuracy rates for three models on two different classification tasks

When distinguishing between poetry, academic, and prose, we achieve quite high accuracy rates for the individual micro- and macrorhythm models—69% and 81% respectively—and an even better rate (83%) when combining these two together. Both of these models score substantially above the null accuracy, 33%, which would be the accuracy rate of a model that simply predicts the most prevalent category for every data point. This indicates that rhythmicity of stress and rhythmicity of pitch movement are indeed meaningful measurements in distinguishing between different styles of texts. We also predicted that macrorhythm would be a better classification feature than microrhythm,

which we can see in the greater than 10% increase in accuracy rate when moving from microrhythm to macrorhythm. To look closer at the different performances of the micro and macro models in distinguishing between the three categories of text, we examine their confusion matrices, shown in Table 4 and Table 5. The testing dataset consists of a total of 42 texts, with 14 per category.

Total = 42	**Poetry (predicted)**	**Academic (predicted)**	**Prose (predicted)**
Poetry (actual)	64.29%	7.14%	28.57%
Academic (actual)	0%	92.86%	7.14%
Prose (actual)	0%	50%	50%

Table 4: Confusion matrix for Micro model

Total = 42	**Poetry (predicted)**	**Academic (predicted)**	**Prose (predicted)**
Poetry (actual)	71.43%	0%	28.57%
Academic (actual)	0%	92.86%	7.14%
Prose (actual)	0%	21.43%	78.57%

Table 5: Confusion matrix for Macro model

For microrhythm, we notice that classifying academic works has the highest standalone accuracy rate (92.86%). However, when examining the errors, we see that poetry has the fewest errors across the board—missed predictions of poetry only make up 38% of the false negatives, and there are no mistaken categorizations of other texts as poetry. In contrast, the false negative errors for academic make up 8% of the total, and its false positive errors make up 61% of the total. In essence, the model is classifying many more works as academic. The model struggles the most with dealing with prose; not only is its success rate only 50%, its false negative errors make up 54% of the total, and false positives make up 38% of the total.

The macrorhythm model achieves greater or equal accuracy rates on all fronts. Classification of academic works again has the highest accuracy rate of 92.86%, though this time it has the same error rate as poetry. The model again performs the worst when dealing with prose, since even though its standalone success rate is 75.57%, slightly greater than poetry?s 71.43%, there are a lot more errors relating to prose. Just like in microrhythm, the greatest sources of error here are where poetry is categorized as prose, and prose as academic.

This greater confusion of prose is not especially surprising, since even by looking at the earlier scatterplots, we see that in both cases the boundary for prose lies in the middle and is least well defined, often bleeding into the other regions. This again lends support to our hypothesis that structural features in prose seem to share commonalities with the two other categories.

The performance of our SVM models is even better when distinguishing between only two categories—creative and noncreative text, where creative consists of poetry and prose, and noncreative is academic. Interestingly, the macro model performs even better here than the two models combined; adding the micro model slightly decreases performance, again lending support to our hypothesis that microrhythm is a slightly worse metric for categorization.

6. Conclusion

In this study, we endeavored to address the following questions: do different styles of texts "sound" different? to what extent are stylistic differences between them actually accounted for by prosodic qualities? and precisely how do we quantify this? While prosodic analyses are traditionally conducted on spoken language, we wanted to analyze written language in order to look for ways in which prosodic features are "embedded" in text, translating lexical and syntactical qualities such as clause length, word length, or use of punctuation into purely prosodic ones—pitch, loudness, and duration. After drawing heavily upon existing research, we decided to measure the microrhythm—regularity of stressed syllables—and the macrorhythm—regularity and degree of pitch movements—of each text.

In the end, we find that our results support our hypothesis that creative works possess a higher degree of rhythm and also a higher degree of rhythmic variation, and that in contrast, noncreative works have both a lower degree of rhythm and a lower degree of rhythmic variation. Specifically, poetic works have high rhythm/medium variation; academic texts have low rhythm/low variation; and prose texts have medium rhythm/high variation. The only exception was that the ranking of poetry and prose was swapped for microrhythmic variation—poetry ended up displaying more microrhythmic variation than prose did.

We also found that measures of microrhythm and macrorhythm are powerful enough to perform quite well in the task of distinguishing categories of text. The best performing models achieve an accuracy rate of 83% when distinguishing between the three types of text, and an accuracy rate of 88% when only distinguishing between creative and noncreative text. In particular, macrorhythm appears to be more powerful in distinguishing texts, suggesting that when it comes to style, greater disparity can be found in pitch contours rather than the syllables. Still, however, it is evident that both of them play a role, and if these two prosodic features can reveal difference in texts, perhaps other prosodic qualities not examined in this paper may as well.

6.1. Limitations

The main limitation in our work is the use of Amazon Polly for converting text-to-speech. While Polly ranks among the current state-of-the-art, it is certainly not perfect, and the speech that it outputs can be clearly identified as software-synthesized. There are a few issues involved, mainly that Polly is unable to adequately recognize semantic elements like sarcasm, contrast, surprise, etc. in the text, all of which have drastic effects on pitch contour. Polly was still enough for us to reveal differences among texts, but we imagine that much better results could be obtained by using better text-to-speech software, such as Tacotron 2, which Google unveiled last month [].

6.2. Future Work

Though we were concerned with written text in this study, we think it would be interesting to turn it around and actually study prosodic differences of humans reading different styles of text aloud, examining whether there are any psychological biases that come out through reading. Perhaps people would read textbooks or academic publications much more quickly and monotonously in anticipation of these subjects being stereotypically dense or boring, while reading books and novels with more tonal variation as they mimic the action of the story.

7. Acknowledgements

I would especially like to thank my advisor, Christiane Fellbaum, for her patience and support throughout the semester, which I very greatly appreciated as a first-timer to independent work. The guidance that I received will be invaluable in future endeavors. I would also like to thank Professor Byron Ahn for briefing me on the vast body of relevant research and for pointing me in the direction that eventually inspired the bulk of my work in this project.

8. Preparation Instructions

8.1. Paper Formatting

There are no minimum or maximum length limits on IW reports. We are including this template because we think it will be helpful for citing things properly and for including figures into formatted text. If you are using LaTeX [**?**] to typeset your paper, then we strongly suggest that you start from the template available at http://iw.cs.princeton.edu – this document was prepared with that template. If you are using a different software package to typeset your paper, then you can still use this document as a reasonable sample of how your report might look. Table 6 is a suggestion of some formatting guidelines, as well as being an example of how to include a table in a Latex document.

Field	**Value**
Paper size	US Letter 8.5in × 11in
Top margin	1in
Bottom margin	1in
Left margin	1in
Right margin	1in
Body font	12pt
Abstract font	12pt, italicized
Section heading font	14pt, bold
Subsection heading font	12pt, bold

Table 6: Formatting guidelines.

Please ensure that you include page numbers with your submission. This makes it easier for readers to refer to different parts of your paper when they provide comments.

We highly recommend you use bibtex for managing your references and citations. You can add bib entries to a references.bib file throughout the semester (e.g., as you read papers) and then they will be ready for you to cite when you start writing the report. If you use bibtex, please note that the references.bib file provided in the template example includes some format-specific incantations at the top of the file. If you substitute your own bib file, you will probably want to include these incantations at the top of it.

8.2. Citations and Footnotes

There are various reasons to cite prior work and include it as references in your bibliography. For example, If you are improving upon prior work, you should include a full citation for the work in the bibliography [**?**, **?**]. You can also cite information that is used as background or explanation[**?**]. In addition to citing scholarly papers or books, you can also create bibtex entries for webpages or other sources. Many online databases allow you to download a premade bibtex entry for each paper you access. You can simply copy-paste these into your references.bib file.

Sometimes you want to footnote something, such as a web site.[1] Note that the footnote number comes after the punctuation.

8.3. Figures and Tables.

Figure 14 shows an example of how to include a figure in your report. Ensure that the figures and tables are legible. Please also ensure that you refer to your figures in the main text. Make sure that your figures will be legible in the expected forms that the report will be read. If you expect someone to print it out in gray-scale, then make sure the figures are legible when printed that way.

In Section 8.1, an example of a table was given. (Note that the "S" in Section is capitalized. Here's one more example - see Table 7.

Here's an example that shows how you can have side-by-side figures - see Figure 15 and Figure 16. (Note that the the "F" in Figure is capitalized.

[1] http://www.cs.princeton.edu

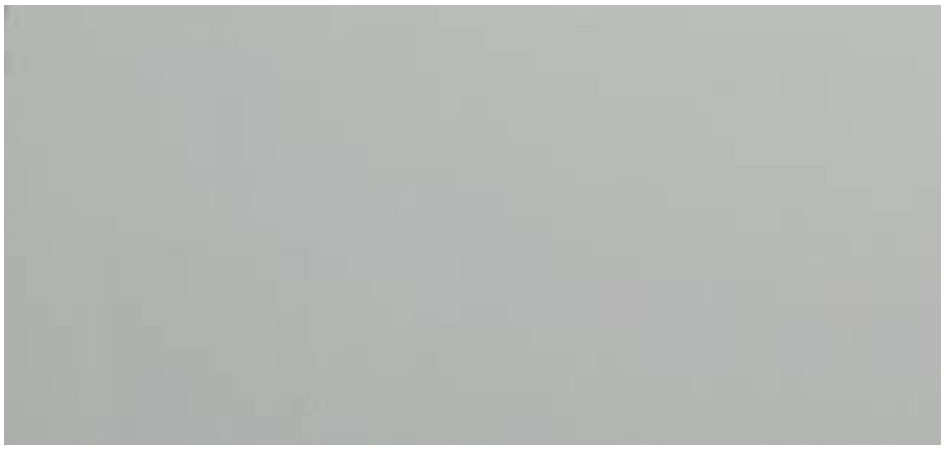


Figure 14: This is a gray image.

Some field	Another field
200	10000
400	20000
800	40000
1600	80000
3200	160000
6400	320000

Table 7: Some data in a table.

8.4. Double Quotes.

Latex double quotes are not the same as the double quote key on your keyboard. The standard way of writing quotes and double quotes in LaTeX is with “ and ” not with " and ".

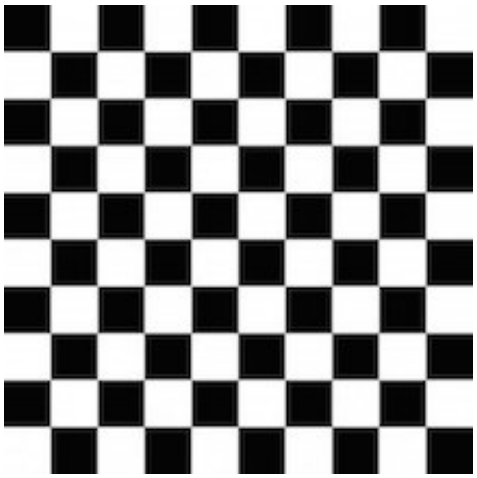


Figure 15: Plain checkerboard.

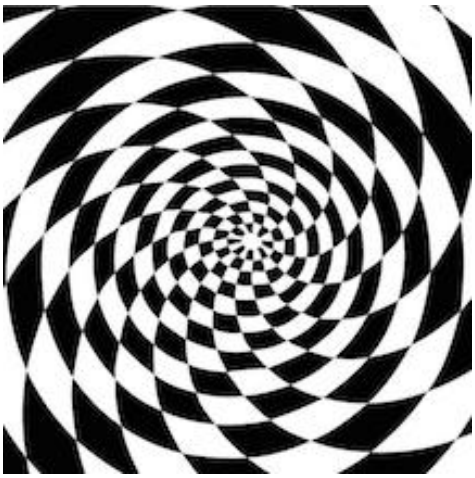


Figure 16: Cool checkerboard.

Now that may be confusing, so you may want to use the \{quotes} command. For example “The quick brown fox.”

8.5. Main Body.

Avoid bad page or column breaks in your main text, i.e., last line of a paragraph at the top of a column or first line of a paragraph at the end of a column. If you begin a new section or sub-section near the end of a column, ensure that you have at least two (2) lines of body text on the same column.

9. Outline

The following is a possible outline for your paper.

9.1. Introduction

- Motivation and Goal (The goal of this project is...)
- Overview of challenge and previous work
- Approach
- Summary of implementation
- Summary of results
- (optional) Roadmap: The remainder of this paper is organized as follows....

9.2. Problem Background and Related Work

- Survey of prior work with similar goals
- For each previous approach, explain what has been done and why it does not meet your goal

9.3. Approach

- Key novel idea
- Why it is a good idea

9.4. Implementation

- System overview (flow chart of key steps?)
- Subsection for each step or issue you addressed
  - Problem statement
  - Possible approaches
  - Chosen approach and why
  - Implementaton details

9.5. Evaluation

- Experiment design...
- Data...
- Metrics...
- Comparisons...
- Qualitative results...
- Quantitative results...

9.6. Summary

- Conclusions...
- Limitations...
- Future work...

10. Ethics

Your independent work report should abide by the basic standards of scholarly ethics and by the Princeton Honor Code. If you have any doubts about how to cite other work, how to quote or include text or images from other works, or other issues, please discuss them with your project adviser or with the IW coordinators.