12-2018

Vowel transitions in the sonnets of Shakespeare: an information theoretic analysis

Josh Freeman

University of Tennessee at Chattanooga, dkl445@mocs.utc.edu

Follow this and additional works at: https://scholar.utc.edu/honors-theses

Part of the English Language and Literature Commons

Recommended Citation

Freeman, Josh, "Vowel transitions in the sonnets of Shakespeare: an information theoretic analysis" (2018). Honors Theses.
Vowel Transitions in the Sonnets of Shakespeare:
An Information Theoretic Analysis

Joshua Allen Freeman

Departmental Honors Thesis
The University of Tennessee at Chattanooga
English Department

Examination Date: 4/9/2018

Joseph P. Jordan
Assistant Professor of English
Thesis Director

Katherine Rehyansky
Professor of English
Department Examiner

Earl Braggs
Professor of English
Department Examiner
Abstract

This paper argues for the importance of articulatory phonology in the study of poetic form and style, showing how a shift of focus away from symbol-probability centric analyses and towards vowel-transition probabilities improves the researcher's ability to understand poetic sound structure. I use information theory and its principal measure, entropy, as a mathematical basis for this argument. I apply information entropy, a mathematical measure of the predictability of a signal source, to the probabilities of vowel sound occurrences in the sonnets of William Shakespeare. I first show how entropy can be applied to the probabilities of vowel sounds as identified by the International Phonemic Alphabet. Then, I present an alternative approach that eschews symbol-based vowel probabilities and instead calculates the probability of the next sound in a series based only on the physical location in the mouth where the previous sound was produced. The principal argument of this paper is that, as the receiver of the poetry signal, one can more accurately predict the next symbol in a series by taking into account the state of the articulator (the tongue) than one can by observing the symbol-based identity of the last sound produced. Said in reverse, I mathematically show that Shakespeare as a source of poetry adheres more strictly to a pattern of physical articulation than to patterns of specific sounds.
Intro

Most poetry scholars would not consider themselves phonologists, yet to the extent that they concern themselves with the structure and function of sound in language, these scholars will find themselves facing phonological problems. In the practice of close reading, critics often use words like “assonance,” “alliteration,” and “rhyme” to describe sound structure. These descriptors are representative of the structure of a poem, but they begin to fail if the scope of the observed relations is too wide. Through history, meter has been the most prominent tool for poets and scholars seeking recognized sonic frameworks. The recognition of the stressed and unstressed syllables in meter gave rise to a host of poetic forms from the villanelle to the sestina, to the sonnet (and more). Modern attempts to go beyond meter in describing these forms have met with varying degrees of success. Many attempts have been made at visually representing sound structure, but none of the range of poetry visualization schemes currently available have gained significant popularity among scholars (Rahman et. al.). The goal of this paper is to use information-theoretic models to explore the relationship between the articulation of vowels and the patterned occurrence of vowels in poetry. Using measurements of entropy, I show how this patterned occurrence, known as assonance or “vowel harmony,” can be more readily understood by taking into account the physical constraints of human speech production. In doing so, I provide what may be a more convenient framework for the detailed study of phonological structure in poetry at large.
1. Goals and Assumptions

The main goal of this paper is to draw attention to the predictive power of the physical aspect of speech. This work is neither mathematical nor functional phonological research, but I draw heavily from each, making generalizations from functional phonology and using the information-theoretic definition of entropy as a measure of symbol predictability. In order to demonstrate the effectiveness of a functional phonological approach, I first calculate the conditional entropy of the poetry data using APA symbols as the most basic unit. Then, the results are compared for the same sets, but with the transitions between symbols used as the most basic unit.

My analysis holds the following assumptions: (1) the body of Shakespeare’s sonnets represents an approximately ergodic source and is therefore a good candidate for an information-theoretic analysis; (2) the view of the relationships between vowels--specifically my proposal that relative backness and height in vowels can be taken as a primitive unit--are accurate and phonologically distinctive; and (3) there is some relation between written poetry and the articulation of spoken words.

This work is organized as follows: sections 2, 3 and 4 pose the need, in the phonological analysis of poetry, for a new system of representation. Section 5 presents a brief introduction to information theory and probes the possibility of entropy as a tool for examining poetry. Section 6 describes the relationship between entropy and encoding algorithms and introduces the “tree of choice.” Sections 7 and 8 present the analysis of Shakespearean sonnets and further discusses the findings obtained from the present study.
2. The Alphabet of Sounds

Most English speakers and writers understand that the symbols we use to write are not inextricably tied to the sounds we produce when speaking. A smaller subset of these people may be familiar with the International Phonetic Association (IPA) and its members’ ongoing work to normalize and represent symbolically all of the possible sounds that a speaker in any language might produce. To this end, the IPA has organized many of the symbols used in this study. While these symbols have appreciated near-ubiquitous usage by linguists, the actual chart used in this paper comes from the American phonetic tradition, primarily because it’s layout and structure better suit my purposes. While some of the symbols and standards of the American phonetic alphabet differ from the International Phonetic Alphabet, both have representations of many of the same sounds, and only minor conversions were needed in my limited usage.

As is standard in any phonetic chart, the vowel chart in figure 1 identifies symbols based on the locations of their underlying articulators in the human mouth:

![Vowel Chart](image_url)

Fig. 1: The American Phonetic Vowel Chart (Milward i)
In the APA chart above, “High/low” and “front/back” refer to the actual position of each vowel in the mouth relative to the others. These positions are supported by empirical research into speech articulation (Milward ii). In my analysis, I use this chart--converted to an (x, y) grid--to track vowel occurrences in poetry. Vowels in the texts examined have been analyzed with software that recognizes and assigns probabilities to repeated structures. This type of analysis is becoming increasingly common in the study of poetry and literature in general. Statistics-based analyses of sound structure have recently been used in software designed to aid scholars in close-readings and, viewed line-by-line, produce structures like those seen along the left side of the following phonetic analysis of “A Drinking Song” by William Butler Yeats:

Fig. 2: “macro Glyphs” (Rahman et. al. 105)

Where this paper differs from these visualizations is in my identification of significant structures. Significant structures, or “primitive units,” are the smallest and identifiable recurring vowel structures in a given analysis. In a typical symbol-based analysis, the symbols of the
International Phonetic Association are used as the significant structures. It is my hypothesis that these symbols alone obscure some of the relationships between vowels. Similarly, in their research into improving speech recognition software, Zhuang, Nam, et al. identify a particular limit of the symbol representational system: “A sequence of segmental phonological units, e.g., phones is the most widely used representation of speech. Since these abstract units are not allowed to temporally overlap with each other, they have difficulty accounting for [all] phonetic variations” (Zhuang 1498). Zhuang et al. are referring here to the specific case of coarticulation, but in the same way that these symbols fail to relay this element, they also fail to express much about the physical degree to which the most recent vowel symbol differs from the last one produced.

3. From Symbols to Gestures

As I previously mentioned, most schemes for interpreting and presenting the phonological data of poetry focus on the phoneme as the most primitive unit. So, one might calculate the probability of the /i/ sound following /o/. For the current analysis, the basic unit is not the phoneme symbol but instead the relative difference between the current symbol and the last symbol that was produced. My following interpretation of the APA vowel chart takes into account the relative nature of backness and height in each of the 11 English vowels identified, with /ə/ being (0, 0):
Using the above relations, I identify the “transition” between two sequential points as the difference between the two points \((x,y)\) value. Taking this information, I look for the repeated usage of a single movement, say \((-1,-3)\), or even series of movements, for instance, logging every time \((+1,-3)\) is followed by \((+1,-1)\). The \((x,y)\) view serves our purposes only as it most conveniently expresses the relative nature of vowel position. These values are strictly categorical and are not meant to reflect real measurable distances.

4. The Advantage of Transitions

A few key factors to note about this viewpoint are central to the hypothesis that vowel position is superior as a predictive and descriptive measure. First, it must be shown that certain transitions may occur at many different locations. One of the simplest is the transition \((+1,-1)\) or its inverse \((-1,+1)\). Referring to figure 3, one can see how this simple diagonal move occurs in any sequence from /i/ to /I/, from /I/ to /e/, and many others. To the contrary, another transition (-10, 0) can occur only in a sequence from /u/ to /i/. These constraints inform the model and lead
to more accurate predictions. The second advantage of this scheme over a typical symbol-based approach is that it encapsulates a notion of complementing and contrasting position. Where an analysis using symbol statistics, observing a sequence /iIe/, would gauge probability based only on the likelihood for each symbol $S$ that could follow /iIe/, a relative-location analysis might determine that the poet in question has an extremely high probability of choosing a distant, contrasting symbol given the locations of those that just occurred. I must stress here that “complement” and “contrast” refer only to vowel articulation, and that the actual auditory properties of the sound signals produced share no known linear relation to their location of articulation (Lee 2). Still, it is my initial hypothesis that this inherent location information will hold more predictive power than the strictly symbol-based approach; the following section goes into detail about how to mathematically verify this hypothesis.

5. Information Theory

Information theory has its roots in late nineteenth-century research concerning the encoding of messages sent across telegraph wires. Operators at the time were in need of increasingly efficient encoding algorithms for the growing length, number, and complexity of messages that they were sending. Many advances were made within the field during this time, but a true mathematical measure of information did not arrive until 1948, when Claude Shannon published the paper “A Mathematical Theory of Communication.” Shannon termed this measure “information entropy” (1).

Information entropy as a measure is given in binary digits--or “bits,” the exact measure used in computer science today. For a given source, the entropy in bits is the smallest number of
binary digits that it should take to encode the ensemble of messages taken from the
source--“ensemble” being a mathematical term that imagines a source producing an infinite
number of messages. Alternatively, entropy could also be viewed as the average minimum
number of yes/no questions one could “ask” to decode a message from the source. In An
Introduction to Information Theory: Symbols, Signals & Noise, John R. Pierce describes entropy
as a measure of uncertainty:

Entropy increases as the number of messages [from] which the source may choose
increases. It also increases as the [source’s] freedom of choice (or the uncertainty [felt
by] the recipient) increases. [Entropy] decreases as the freedom of choice and the
uncertainty are restricted. (81)

Shannon provides several methods of calculating entropy, the simplest case of which we will
now examine. For two possible options X and Y, we say X will have a probability \( p_0 \) and Y a
probability \( p_1 \). The entropy, \( H \), is calculated as follows:

\[
H = - (p_0 \log_2 p_0 + p_1 \log_2 p_1) \text{ bits per symbol}
\]

As Peirce explains, “entropy is the negative of the sum of the probability \( p_0 \) that X will be chosen
times by the logarithm of \( p_0 \) and the probability that Y will be chosen times the logarithm of this
probability” (Pierce 81). In the case of the flip of a coin:

\[
P_0 = \frac{1}{2} \\
P_1 = \frac{1}{2} \\
H = - \left( \frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{2} \log_2 \frac{1}{2} \right) \\
H = - \left[ \left(\frac{1}{2}\right)(-1) + \left(\frac{1}{2}\right)(-1) \right] \\
H = 1 \text{ bit per toss}
\]

Thinking back to the idea of entropy as a measure of the number of questions asked by the
receiver of a message, we see that 1 bit makes sense; after a coin flip, one would ask, “Is it
heads?" When we receive the answer to this one question, the “message” is relayed, and so, the entropy equals 1. If there were two coins, two questions must be asked to relay the message, totaling 2 bits. In a different case, taking a single dishonest coin that shows heads 75% of the time and tails 25%, entropy is equal to 0.811. Notice that the entropy is lower. Entropy can be interpreted as a measure of the amount of information received by a message; Because we are more informed about the possible results of the flip, we actually receive less information. In fact, as the probability $p$ moves either above or below 50%, the entropy $H$ lowers. The relation between known probability $p$ and entropy $H$ is shown in the graph of a binary entropy relation in figure 4. As $p$ diverges from .5, the less surprised we are and the less new information we actually receive.

![Fig. 4 (Pierce 84)](image)

The relation between probability and entropy is more obvious if you consider flipping two coins, where one coin is a dishonest coin that shows heads every time. The formula is as follows:
Heads: \[ P_0 = \frac{1}{2} \quad P_2 = 1 \]
Tails: \[ P_1 = \frac{1}{2} \quad p_3 = 0 \]

\[
H = - \left( \frac{1}{2} \log \frac{1}{2} + \frac{1}{2} \log \frac{1}{2} + 0 \log 0 + 1 \log 1 \right) \\
H = - \left[ \left(\frac{1}{2}\right)(-1) + \left(\frac{1}{2}\right)(-1) + (0)(-\infty) + (1)(0) \right] \\
H = 1 \text{ bit per toss}
\]

Though you would normally expect two coins to yield two total bits of information, the entropy for this case actually equals 1 bit. Because we know about the dishonest coin beforehand, we always know the result, and so, receive no information from looking at the result of that flip. If the dishonest coin operated only 75% of the time, it would have an entropy of 0.811 and would be added to the honest coin for a total entropy of $H=1.811$. Simple entropy can be calculated in this way for any arbitrary number of input (number of coins) $n$ with probabilities $P_n$.

6. Tree of Choice

Another representation of entropy, and one that is critical to my later analyses, is the “tree of choice” (Pierce 74). This diagram can be used to illustrate two components necessary for understanding my thesis: conditional entropy and binary-digit encoding. The tree below shows relations between the probabilities of the appearance of given words. At each fork, each branch is labeled with either a 0 or a 1. Each represents a decision to take the left fork or the right in the tree.
This type of tree is always organized so that the word or group of words with the highest probability occur at the first fork in the tree. If we were to send this message across telegraph wires according to the given probabilities, 50% of the time we would send “the.” The other 50% of the time, we would intend to send one of the seven other words. The receiver (aware beforehand of our encoding scheme) would know that the first digit will always answer the first question: “is ‘the’ the word that was sent?” 50% of the time, the message we send will be “1,” meaning “yes” or “take a the first right turn.” So, 50% of the time we need only 1 binary digit to relay our word. In order to relay the word “man,” we would send “001,” totaling three bits, or “left, left, right.” In the unlikely event of any of the worst case scenario words, we require at most 5 bits, or five turns, to specify the “path” to take to receive a given word in the message. Thus far, the example is in line with the simple entropy described earlier, and it may shed light on how the ensemble of all the possible outcomes can result in entropy being a decimal value.
This decision-tree example is currently static and operates with no reference to what symbol has come before. This could easily be changed to conditional entropy by assigning probabilities for each word $S$, given that some other word $B$ has just occurred. So, one would have a collection of trees, one for each possible symbol. The operator, having just received the word “man,” would consult the “man” tree, under which, the likelihood of the very next word being “horse” is incredibly low. The math for this and its effect on entropy is not too far removed from simple entropy and is handled by a simple summation. This version of entropy is the most crucial to my analysis.

7. Analysis Using Entropy

The first subject of this analysis is a string of one thousand randomly generated vowels. Vowel occurrences in the set are equiprobable but with a random distribution. This is meant to show something about what we might expect from calculating conditional entropy from a source that imposes no structure other than equiprobability. The pie charts and bar charts in figures 6 and 7 visualize this distribution. Figure 6 visualizes the probabilities of individual symbols in the random set.
Of all the inputs tested, simple entropy is highest in the random set, coming in at 3.45 for these individual symbols. The higher entropy here results because equiprobability leads to a balanced “tree of choice” with many branches and symbols evenly spread across them. In contrast, an unbalanced tree is indicative of lower entropy. As one traverses the branches of the tree, entropy is reflected in the number of options that are eliminated at each turn. In an unbalanced tree, all options but one might be eliminated at the first turn. This is exactly the case in the earlier figure 5 example, where “the” occurs 50% of the time and is an option at the first fork. In a perfectly balanced tree, the most one could eliminate at each turn would be half of the current options, because all symbols are evenly distributed across all branches. Thus, equiprobability leads to a wider, taller tree, increasing the maximum number of branches one would need to traverse in the
worst case scenario--and directly increasing entropy. Figure 7 shows the same random set, but with the probabilities of pairs of symbols. Again, equiprobability results in evenly distribution.

Fig 7: probabilities of symbol pairs

Compare this to figures 8 and 9, which show the entropy of a single sonnet, Shakespeare’s Sonnet 15; Sonnet 15 is distinctly unbalanced, not evenly distributed, and therefore, has a lower entropy. Something has obviously changed. The lower entropy in this case is due primarily to the fact that we have 20% probability of finding /l/ as the next symbol. The second most probable symbol /a/ occurs around 17% of the time, and the next two, /e/ and /a/ occur at lower but similar rates. Any of the remaining seven symbols combined (that’s over half of all the symbols possible) occur only 37% of the time. The factor that has changed--the factor that produces the lower entropy measured--is selectivity. Shakespeare as the message source is
being much more selective with his use of sound, and it shows in this analysis. Figure 9 reveals
the same for pairs of symbols in Sonnet 15. One key thing to note in figure 9 is the total number
of pairs (easily seen by the total number of bars in the bar chart) is drastically lower. This
reduced variability in Sonnet 15 has dual causes. First, the repeated selection of some pairs by
the artist excludes the selection of the variations of the random set as shown in figure 7. Second,
and most importantly, the random generation of figure 7 shows no regard to the natural sound
structure of words and phrases in the English language. Figure 9 is less varied both because the
Shakespeare follows both an individual artistic patterning of sound and an implicit sound
structure enforced by the English language.

![Pie chart and bar graph](image)

**Fig. 8**

Until now, I’ve shown only measurements of simple entropy, which assume the symbols
occurring have no effect on the symbols to come. We know this is not the case, but simple
entropy lends itself well to graphical representation and as an introduction to the concept. The
focus of this paper, though, is on conditional entropy. Conditional entropy is difficult to graph, because the probabilities as shown in figures seven through ten--

--change both with the identity of the current symbol and with number of symbols into the past we choose to look. Figure 10 shows a comparison of digram, trigram, quadrigram entropies between the random set and Sonnet 15. These “n-grams” used here are terms for how far back in time we are looking, or more specifically, the size of the “block” for which we are calculating probabilities. “Digram” implies looking at only the last symbol that has occurred in order to predict the next. “Trigram” takes into account the last two to predict the third, and so on. N-gram is a term used in general and in this paper, because there is no hard limit on the number of symbols one can choose to consider.

In the following graph, entropy for the random set does not decrease significantly until the final series. Equi-probability is the likely cause of this trend: while it is quite difficult to determine what the next or second removed symbol will be in the random series, equi-probability makes it very unlikely that the same symbols will occur repeatedly in quick succession. It seems
that quadigram probabilities (i.e., the probabilities gleaned from looking at the last three symbols to determine the fourth) give us enough information about the past to make a fairly educated guess. In contrast, conditional entropy for Sonnet 15 drops much more consistently and dramatically with each step. This result suggests that the occurrence of phonemes in Sonnet 15 is, at each level, more structured than the random assortment. This steep decline in entropy can be seen in all five sonnets observed in the study.

Fig. 10

I have up to this point only shown results for symbol-based probabilities. Now, I will show side-by-side comparisons of the same data using my scheme for the generalization to transitions. Figure 11 below shows that encoding based on transitions results in conditional entropy that is not only lower, as hypothesized, but is in almost every case less than half that of the symbol-based approach.
The results in figure 11 seem to support my hypothesis that patterns of articulation are more predictable than patterns of phonemes—at least, in the selected sonnets by Shakespeare.

8. On the Significance of the Model

I must acknowledge that, even in the random set, the entropy of transitions was lower than symbol entropy. This leads one to question whether or not transitions are truly a useful way to view the structure of poems or whether the drop in entropy shown in my analysis is only the result of oversimplification. One might ask what comments can be made about the structure of poetry using transitions. To that end, I attempted a final comparison taking advantage of the pigeonhole effect. “The pigeonhole effect” is a term in information theory that describes a peculiar trait of the “tree of choice.” The tree of choice is a basic way to imagine a message
encoding algorithm, as I described earlier. The pigeonhole effect is an attribute of any such algorithm. In short, a given algorithm is only efficient at encoding the messages it was designed to encode. If one were to try to use the original tree of choice from figure 5 to encode some other message (ignoring the fact that the sender would probably fail for lack of sufficient vocabulary), that person would soon find out that their message-passing would become incredibly inefficient. The encoding algorithm of figure 5 is designed to send the word “the” 50% of the time. Any message sender that does not follow that probability and the others that coincide with it will not succeed in sending their message efficiently. In essence, this principle says that, for an encoding scheme to send more varied messages, the encoding scheme must be adjusted and made more varied. This adjustment means adding more branches to the tree of choice, and more branches mean an equivalent rise in entropy. This relates to my analysis of the sonnets, because I view the patterned occurrences of phonemes as the encoded message. Up until now, I have calculated entropy for each sonnet individually. The pigeonhole principle holds that a calculation of entropy for a set of two poems (or any set of two messages) will always be higher than either set calculated individually, unless the two are equivalent. In addition, to the degree that the patterns in the pairs of sonnets differ, a direct increase in entropy should be visible. In the following figure, I show the results of these calculations performed on pairs of sonnets. I chose to compare Sonnet 15 to the other four sonnets covered in this study. Before performing the calculation, I thought that one or more of the sonnets chosen would have a comparatively low or high entropy in all n-gram categories. That is to say, I thought one of the sonnets would be clearly revealed to be most similar to Sonnet 15. To my surprise, not only was there no clear “winner,” but there was also no clear connection between each individual pair’s own n-gram entropies.
While the pairing of Sonnet 15 and Sonnet 18 had the lowest digram entropy, the trigram entropy of that pairing was the highest, and, while the pairing with Sonnet 116 had the highest digram entropy, its 4-gram entropy was the lowest. Instead of identifying one sonnet as “most similar” to Sonnet 15, I can only say that it seems that Sonnet 18 is most similar at the digram level, Sonnet 129 at the trigram, Sonnet 116 at the 4-gram. Overall, though, there is very little difference in the entropies of any of the pairs, which I attribute to the fact that all five Sonnets were written by the same author and are therefore likely to share a similar sound structure. I believe the results in figure 12 show that a transition-based analysis is not a blanket reduction of complexity. Instead, transitions seem to encode more complexity than the current analysis can dissect.
Conclusion

There is more to the aesthetics of poetry than the sounds that are heard when it is spoken. Branches of literary criticism from structuralism to moral criticism and beyond operate with little regard to sound structure. Even this study considers only the physical--not the audible--aspect of speech in poetry. This study does not seek to place the analysis of articulation above other methods, but focuses instead on how sound and articulation complement the many other overlapping characteristics of poetry and the written word at large. It is undeniable that repetition, specific distribution, and correlation of sounds do exist in great variety and complexity in poetry. The difficulty in reaching greater conclusions about these characteristics comes from the breadth and variability of language. This study seeks to generalize the discussion of sound structure; it hopes to move toward a convenient framework to argue, for example, that the eerie tone of a poem is intensified by the use of frontal vowels that mimic whispering, or that a poet’s trademark style is characterized in part by over- or under-use of certain vowel patterns.

I do not believe that any particular style of sound structure will ever prove intrinsically “better” than any other, but I do believe that certain artists operate within self-similar patterns, and I believe this research shows that our comprehension of poetic sound structure can be improved by a renewed focus on articulation and how articulation shapes our notions of style and form. While this analysis seems to support my hypothesis that transitions are a superior predictive measure, the use of transitions to make greater structural statements about poetry is still in question and may only be revealed by further research performed on larger quantities of written work.
Works Cited


