# Entropy and Information Content of Graphical Symbols in Written Language

Daniel T. Citron*
*Department of Physics, Cornell University, Ithaca, NY 14850, U.S.A.*

Michael Flashman† and Isabel Kloumann‡
*Center for Applied Mathematics, Cornell University, Ithaca, NY 14850, U.S.A.*
(Dated: May 11, 2012)

Entropic methods provide for useful quantification of the complex structure of language. Here we review recent work on the application of relative entropy to the structure of language at the level of word ordering. Following this review, we examine the visual structure of alphabets across languages by measuring the entropy of different five-pixel motif across the alphabet. Preliminary results suggest a common pattern in the entropic properties of many alphabets in spite of visual differences between alphabets. In particular, the entropy associated with visual motifs is roughly constant across all alphabets under study.

## I. INTRODUCTION

With his seminal 1948 paper, "Mathematical Theory of Communication," Claude Shannon developed a set of tools for formally studying and quantifying how the content of messages is transferred from source to receiver. He argued that a quantity called entropy best quantified the important aspects of the message. Previously used in the context of thermodynamics to quantify disorder, Shannon's entropy measured the information contained within a message. Until Shannon's theory, interpreting the semantic importance of words in a sentence, or of letters in a word, would have been considered a problem appropriate for scholars of literature or linguistics. By using entropy to connect the physical concept of disorder to the problem of information, Shannon's theories invited a mathematical analysis of the organizing principles at work in everyday human language [1].

We pursue an entropic analysis of the visual components of alphabets from a variety of real languages. After all, in a basic sense one may think of the written word as visual communication from a document to the reader. Previous studies of the graphical forms of alphabets used worldwide have argued that letters are chosen so as to be most immediately recognizable and distinguishable from other letters [2]. Knowing this, prompts the question, what visual features of alphabets make them well-suited for communication? Following the analytical methods used by previous studies to quantify the importance of organization found in text, we make entropic measurements of the visual organization seen in characters of many different alphabets .

*Electronic address: dtc65@cornell.edu
†Electronic address: mtf53@cornell.edu
‡Electronic address: imk36@cornell.edu

### A. Entropy and Communication

The entropy of a closed system quantifies one's ignorance of the configuration of that closed system. The entropy of a microcanonical ensemble is proportional to the logarithm of the number of states available to the system ($\Omega$) [3][4]. This definition of entropy assumes that each configuration occurs with equal probability. One can extend this definition of entropy to include out-of-equilibrium systems by generalizing to probabilities that are unequal [4][5].

$$H = -k \ln \frac{1}{\Omega} = -\sum_{i=1}^{\Omega} p(x_i) \ln p(x_i) \qquad (1)$$

This latter definition of entropy, known as the information entropy, is particularly useful in the study of communication. It provides a convenient way to quantify the amount of information that a message communicates, be it a sentence, a word, a bit string, a sound, or a picture. To give a simple example, suppose Alice sends Bob single bit message, "0" or "1". If half the bits typically arrive with value "1" and half typically arrive with value "0", then Bob has a 50% chance of correctly guessing the content of the next message; his ignorance of the next message is maximized. If instead Alice mostly sends "1"s, then Bob will be much more likely to correctly guess the content of the next message. Not coincidentally, the information entropy in this example is at a maximum for the complete ignorance case and at a lower value for the other case [5]. It is straightforward to show [4] generally that the information entropy of any probability distribution of messages (not just single bits) is maximized at equal probabilities (maximum ignorance) and decreases as the probability distribution deviates from uniform (less ignorance).

### B. Entropy and Disorder

One can also argue that entropy (Eq. 1) is an appropriate measure of disorder. The Ising model of a fer-

romagnet serves as a good example of how disordered systems have more entropy than ordered systems. In the context of statistical mechanics, one can think of an ordered system as possessing patterns that persist across long distances. For the Ising model, we define an order parameter that quantifies how many of the spins are aligned parallel to the same direction: in an ordered Ising system, the majority of spins are aligned the same way, while in a disordered Ising system there is no such majority alignment of such spins. As the Ising system orders and the magnetization increases, one can show that the number of states available to the system decreases [6]. We therefore see that the system entropy decreases as the organization in the system increases.

These two definitions of entropy – a measure of ignorance and a measure of disorder – there are really one and the same, and this equivalence becomes clear in the context of studying language as a form of communication. One can expect a lesser degree of disorder in a text composed in a language with many rules of spelling, syntax, and other organizing principles. The patterns that one observes in language, in the word content, word ordering, spelling, character content, or pictorial representations of language, all serve to reduce ones ignorance of what is being communicated. We therefore turn to entropy as a way of quantifying the amount of information contained in several structural aspects of human language.

## II. ENTROPY OF WORD CONTENT AND ORGANIZATION

### A. Word Frequency and Zipf's Law

We consider the information content contained in the word content of languages. The relative frequencies of different words used in text or speech show evidence of nontrivial organization. Consider a simple model of language as a series of words transmitted from one person to another. One can think of each word available for use in communication as being part of an ensemble of $N$ words. We define the probability distribution $P(X)$ by measuring the frequency of each word. Using the definition of information entropy, one can obtain a measurement for the amount of information entropy associated with a single word taken at random from the ensemble of available words.

Measurements of $P(X)$ for real human languages yield a probability distribution for words known as Zipf's law [7] [8]). If one examines words in text (or spoken words) and ranks all words according to their relative frequencies of use $F(X)$, one obtains a power law relationship.

$$P(X) \propto F(X)^{-1} \tag{2}$$

For example, if "the", "and", and "of" are ranked as the first, second, and third most frequently used words in English, one finds that "the" is two times more frequently used as "and" and three times more frequently used than "of". What is most remarkable about this result is how this probability distribution appears in many human languages, suggesting a universal property of the structure of languages independent of context, geography, or the specific history of a language [8][9].

Comparing Eq. 2 to a uniform probability distribution, entropy can be used to demonstrate that the word content of language is organized, not random. If one were to take a naive view of language as a completely random sequence of words, one would expect the probabilities of all available words to be uniform. Instead, the probability distribution is far from uniform. Comparing the entropy of Zipf's law to the naive consideration of language, one concludes that word content of real language is by no means close to a maximum entropy state.

However, the entropic interpretation of Zipf's law hardly gives the full picture of the information communicated through words because Zipf's law only considers frequency ranking and gives no consideration to word ordering [8][7].

### B. Importance of Word Ordering

The patterns observed in how words are organized together in texts prompt an entropic analysis of word ordering. How much information is contained in the specific order in which words are communicated, and how does a meaningful text differ from a text with the same word content but arranged in a random sequence? One recent study by Montemurro and Zanette used measurements of entropy to quantify the importance of word ordering by comparing real texts to randomized texts composed of the same words [8] . Both the random and real texts had identical statistical properties with regard to word content, but only the real text contained any deliberate organization or meaning.

In their study, Montemurro and Zanette compared randomized text to real text by measuring the relative entropy. Relative entropy for two probability distributions $P(X)$ and $Q(X)$ is defined as follows [5]:

$$H_{rel}(P\|Q) = \sum_{\{x\}} P(x) \log \frac{P(x)}{Q(x)} \tag{3}$$

Relative entropy can be thought of as the amount of information lost by assuming that a particular variable $x$ follows the distribution $Q(x)$ when it actually follows $P(x)$. Relative entropy is particularly useful for measuring the importance of correlations between variables. Replacing the first probability distribution with the joint probability $P(x, y)$ of two variables $x$ and $y$, and the second probability distribution as the product of the marginal probability distributions $P(x)P(y)$, we obtain [5][8]:

$$D = \sum_{\{x\}} \sum_{\{y\}} P(x, y) \log \frac{P(x, y)}{P(x)P(y)} \tag{4}$$

Using the definition of conditional probability, it is straightforward to show that this form of the relative entropy, can be rewritten as [5]:

$$D = H(x) - H(x \mid y). \tag{5}$$

We therefore interpret the relative entropy as the entropy difference between a random variable $x$ and the entropy of that variable given information about some other variable $y$. Any correlations between $x$ and $y$ reduce the entropy in $x$, since knowledge of $y$ reduces our ignorance about the variable $x$.

Relative entropy can be used to study the ordering of words by examining the statistics of long sequences of words in text (n-grams). Let each variable $x_i$ represent a word, and let the probability of any particular word be $P(x_i)$. Let the probability of a particular sequence of n words be $P(x_1, x_2, ..., x_n)$. The entropy contained in these word sequences can be computed using [8]:

$$H = -\frac{1}{N} \sum_{\{x_1 \ldots x_n\}} P(\{x_1 \ldots x_n\}) \log P(\{x_1 \ldots x_n\}) \tag{6}$$

where the sum is over all n-grams of a particular length. By considering the probabilities associated with specific n-grams, Eq. 6 accounts for word correlations. In the absence of word correlations, the joint probability $P(\{x_1 \ldots x_n\})$ becomes the product of the independent probabilities of the n words that compose the n-gram: $P(\{x_1\}) \ldots P(\{x_n\})$. In the absence of any word correlations, the average entropy per word combination is:

$$H_r = -\frac{1}{N} \sum_{\{x_1 \ldots x_n\}} P(x_1) \cdots P(x_n) \log \left[P(x_1) \cdots P(x_n)\right]. \tag{7}$$

Equation 6 is a measure of entropy that accounts for word correlations, while equation 7 is the entropy a text would have in the absence of word correlations. The relative entropy (Eq. 3) is found by subtracting the two other entropy measures $D = H_r - H$ [8]. In this sense, relative entropy captures the decrease in the per-word information content resulting from word correlations.

This definition of relative entropy allows one to measure the importance of correlations between words in text. To lend some intuition to the notion of word correlations in the context of language, suppose one wanted to guess the next word in a text. Outside the context of all other words in the text, it is most likely that the next word is "the" or some other highly common word. But, if there are preceding word sequences, such as "by the", one can tell from the presence of other words that the next word is most likely a noun such as "house". The preceding words provide context that changes the probability distribution of the next word in the sequence and reduce the entropy of that next word.

The most surprising result obtained from the study by Montemurro and Zanette is how the relativity entropy measurements of texts shows a common pattern across many languages. Fig.1 shows their results [8]. The blue columns represent the entropy of the randomized texts, calculated using Eq. 6, and the green columns represent the entropy of the real texts, calculated using Eq. 7. Subtracting the two different entropies (similar to Eq. 3) yields the entropy due to word ordering only, shown in red. While the heights of the blue and green columns vary across all languages studied, the heights of the red columns are roughly constant across all languages. That all languages mentioned in the study have this entropic property in common suggests a universal importance of the deliberate ordering of words in text.

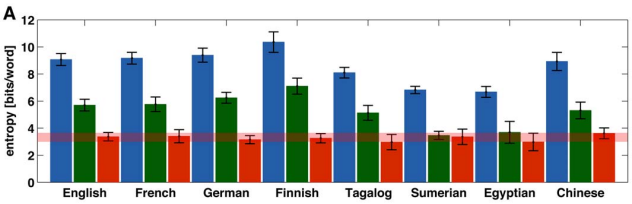


FIG. 1: Blue represents the entropy of randomized text, green the entropy of real texts, and red the relative entropy (the difference between blue and green). Taken from [8]

.

## III. ENTROPY OF GRAPHICAL COMMUNICATION

We propose to extend these entropic methods for studying linguistic structures to examine the information content of visual representations of texts. Our discussions of previous work have shown that the concept of entropy is useful for studying the structure of language and quantifying the importance of word content and ordering. These studies considered a single word to be the smallest building block of language. This choice of scale was arbitrary, as linguistic structures exist at all levels of organization in text [7][8]. In our study we choose the graphical representation of a single letter as the smallest building block of language.

Casually defined, an alphabet is a collection of graphical symbols (made of lines, curves, dots, etc.) which are arranged on a page to visually encode language. These symbol sets provide the building blocks for non-verbal communication. There is great variety in the alphabets that have been invented through the course of human history, but in each case the symbols serve the same principal function: to communicate complex ideas by encoding words. It is this common function which prompts us to explore whether information theoretic measures of the entropy associated with alphabets will bring to light any common features.

Let's now draw an analogy between the organization of words and the graphical organization of symbols. Languages possess significant syntactic and grammatical structure. This may not be the most efficient way to com-

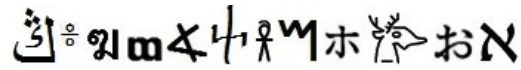

FIG. 2: A sampling of ancient and modern alphabets from around the globe used in our study.

.

municate in the sense that many words merely structure a sentence, rather than providing meaningful information. However, this added structure provides for unambiguous and robust communication. Similarly, we expect that alphabetic symbols will be visually distinguishable and robust to perturbations. An alphabet needs to be read easily and unambiguously, so that individual words can be quickly parsed and assembled into meaningful sentences.

### A. Data and Methods

In the present study we consider a collection of 30+ alphabets. The selected alphabets were chosen to represent the diversity of symbol forms across geographical regions and time, see Fig. 2.

Graphical images are built from pixels, and in a given character a pixel can be in one of two states: black or white. In exploring the amount of order in a graphical image, we are quantifying the extent to which pixels within an alphabet are correlated. In our study we consider the most basic shape that can capture pixel correlations, which consists of a central pixel and its four nearest neighbors (see Fig. 3; hereafter we refer to the shapes associated with these combinations of pixels as 'cross-motifs'.

FIG. 3: Examples of the cross-motifs used to compute the pixel-correlated entropies. The ones shown here are eight of the 32 possible combinations.

We measure the entropy of cross motifs using Eq. 6, where each $x_i$ encodes whether a pixel is black or white, and the sum is over all binary sequences of length 5.

### B. Results

Fig. 4 displays our measurements of $H$ and $H_r$ for the 30+ alphabets in our dataset. A clear pattern emerges in Fig. 4: every alphabet in our data set exhibits an average cross-motif entropy $H$ (in green) distinctly less than $H_r$ (in blue), indicating that all alphabets have substantial pixel correlations. In other words, the cross-motifs that compose alphabet characters are not random, some cross-motifs are distinctly more likely to be observed than if characters were composed of random pixels.

Fig. 4 indicates that the average shape entropy of cross-motifs (green bars) is roughly constant across all alphabets. Further, the relative entropy due to pixel correlations (red bars) is roughly constant across all alphabets.

Our latter result recalls what was found in the study by Montemurro and Zanette, in that the value accounting for the entropy of word correlations was constant across all of the languages in their data set (see Fig. 1). Their study, however, was able to make estimations of the errors in their entropy measurements due to the fact that there are a multitude of texts in each language and so multiple trials of their measurements could be made. To make similar statistically significant conclusions we would need to provide meaningful error estimations of the quantities $H$ and $H_r$ (Fig. 4). However, the extent to which this is possible is inherently limited due to the fact that each alphabet has only a few symbols and only occurs once per language.

## IV. CONCLUSION

We explored the use of relative entropy as a tool for studying the structural organization of languages. This method was first demonstrated in a previous study that used it to quantify the importance of deliberate organization of words in text. We extended this method to examine graphical representations of characters by measuring the probability distribution of various 5-pixel motifs that appeared in each image. We found that the relative entropy of alphabets was roughly constant across all languages studied. Although the study of word organization and the study of the organization of a visual image are by no means identical, this method is effective at quantifying the importance of structural organization in a linguistic context.

Future work should develop a statistically rigorous measure of the graphical information content in a graphical character, though it is possible that the extent to which this is possible may be inherently limited due to the fact that alphabets are composed of a small number of symbols and that each alphabet only 'occurs once'. One proposed solution to this problem would be to compare the relative entropy of various fonts in a given language as a way of estimating the variations that appear in different realizations of a single alphabet.

[1] J. Gleick, *The Information: A History, A Theory, A Flood* (Vintage, 2011).

[2] M. Changizi, Q. Zhang, H. Ye, and S. Shimojo, The Amer-

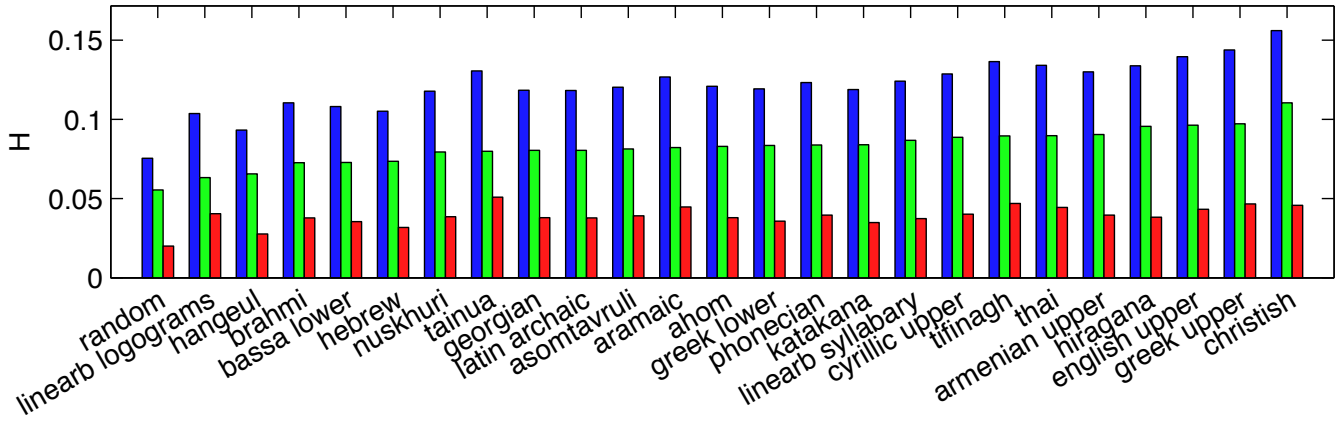


FIG. 4: Blue represents the entropy of randomized pixels, green the entropy of real characters, and red the relative entropy (the difference between blue and green).

ican Naturalist **167**, E117 (2006).

[3] M. Kardar, *Statistical Physics of Particles* (Cambridge University Press, 2007).

[4] J. P. Sethna, *Statistical Mechanics: Entropy, Order Parameters, and Complexity* (Oxford University Press, 2010).

[5] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (John Wiley & Sons, 1991).

[6] P. M. Chaikin and T. C. Lubensky, *Principles of Condensed Matter Physics* (Cambridge University Press, 2000).

[7] M. Montemurro and D. H. Zanette, Advances in Complex Systems **5**, 7 (2002).

[8] M. Montemurro and D. Zanette, PLoS One **6**, e19875 (2011).

[9] M. E. J. Newman, Contemporary Physics **46**, 323 (2007).