Entropy in Written English

Alexandra Jurgens

[1]Department of Physics , University of California, Davis,
101 Physics Building, 1 Shields Avenue, Davis, CA 95616 USA

June 10, 2016

Abstract

Information theoretics as developed by Shannon and Crutchfield offer power tools to examine stochastic and deterministic processes [1], [2]. We apply these tools to written English, following earlier work by Shannon [3]. Methods are compared and proposed to minimize sampling error. We also examine using information theoretic models to study structure at different levels of language. Following earlier work, we use autocorrelation functions to examine structure in novel length texts [4].

1 Introduction

1.1 Motivation

The increased understanding of natural language through mathematics is an endeavor with a long history. Its modern incarnation owes much to C. E. Shannon's influential work *A Mathematical Theory of Communication*, in which Shannon defines what has come to be known as the Shannon entropy [1]. This work has been examined and expanded over the years, resulting in a coherent and well defined set of quantities that form the basis for Information Theory [2]. This field has also given rise to various predictive models, among them ϵ-machines, a minimal optical predictive model of dynamical systems.

Shannon's original purpose in studying information theory was to examine English messages through communication lines for the Bell Telephone Company. In pursuit of this goal, he attempted a measurement of the entropy of printed English in 1950 that has held up well, with some limitations [3]. Shannon's work, while groundbreaking, was constrained by available data and measurement techniques. Futhermore, Shannon makes entropy measurements based on letter frequencies, but we expect there to be interesting structure on the level of words and syntax that can be further studied. These things have been previously discussed, with varied results [4], [5].

1.2 Synopsis

The original goal of this project was to study the information theoretic quantities of written language as defined by Crutchfield et al. using the large corpus of written language made available by information gathering companies like Google [2]. A measurement of the block entropies up to word length nine was performed, along with a measurement of block entropy rates. This study prompted a short look at designing predictive word models, using Markov chains and word blocks of different lengths. This mimics the approach by Shannon in 1948 to replicate natural language through purely statistical means [1].

This report also includes a cursory overview of methods designed for studying natural language at the word level, which have been performed in largely separate disciplines. Words are defined here to be sequences of letters that exist between two spaces in written language. Words have been examined mostly through statistical word frequency models such as Zipf's law, as well as an examination of word auto-correlation in text [4], [5], [6].

2 Background

2.1 Definitions

The definitions of commonly used information theoretic quantities are produced below for ease of reference in this report. Any lesser used quantities will be defined in the main text as necessary.

Definition 2.1. ***Shannon entropy*** *Let X be a random variable that assumes the values $x \in \chi$, where χ is some finite set. Denote the probability that X assumes some value x by* $\Pr(x)$. *The Shannon entropy of X is defined by*

$$H[X] = -\sum_{x \in \chi} \Pr(x) \log_2 \Pr(x)$$

.

The units of $H[X]$ are bits. Note that to go from the Shannon entropy of a single variable to the Shannon entropy of a length-L sequence we simply replace x with s^L, a block of L consecutive variables. We call this the block Shannon entropy, or simply block entropy.

Definition 2.2. ***Entropy Rate*** *The source entropy rate h_μ is the rate of increase with respect to L of the block Shannon entropy in the large-L limit, defined by*

$$h_\mu = \lim_{L \to \infty} \frac{H(L)}{L}$$

.

Definition 2.3. ***Entropy Gain*** *Entropy gain is the growth in $H(L)$ with increasing L. It is calculated by taking the numerical derivative of H(L),*

$$\Delta H(L) = H(L) - H(L-1)$$

.

Note that $\Delta H(0) = \log_2 |\mathcal{A}|$.

We consider the entropy gain to be an estimate of how random the source appears is only blocks of length L are considered. With this definition we identify $H(L)$ to be the finite-L approximation to the entropy rate h_μ, $h_\mu(L)$. We also refer to this as the *block entropy rate*. Of course, we expect the block entropy rate to approach h_μ at increasing L.

Definition 2.4. ***N-Gram*** *An N-gram is a sequence of length N variables $x \in X$ where X is the alphabet belonging to a written natural language. While in principle the size of the set of N-grams at any given N is $|\mathcal{A}|^N$, in general there are forbidden words, so that the set of N-grams that occur in written lenguage at length N is much smaller than the set of all possible N-grams.*

Definition 2.5. ***Zipf's Law*** *Zipf's law is an empirical law stating that certain types of data will follow a Zipfian distribution, a kind of power law distribution. The simplest Zipfian form is*

$$f(r) \propto \frac{1}{r^\alpha}$$

where r is rank in a frequency table, $f(r)$ is frequency, and α is a parameter that varies by dataset.

2.2 System

2.2.1 Letters

In general, we model the written English language as a stationary stochastic process. On the level of letters, the alphabet $\mathcal{A}$, may seem well defined, but it is actually a matter of debate. The simplest definition would include the 26 letters of the English alphabet and no other characters. However, we may also include the space, punctuation marks, and Arabic numerals. Additionally, while uncommon in English, accented and unusual characters do occasionally appear. To control the size of possible N-grams, in this paper we take the simplest definition of $\mathcal{A}$, the 26 letters of the English alphabet.

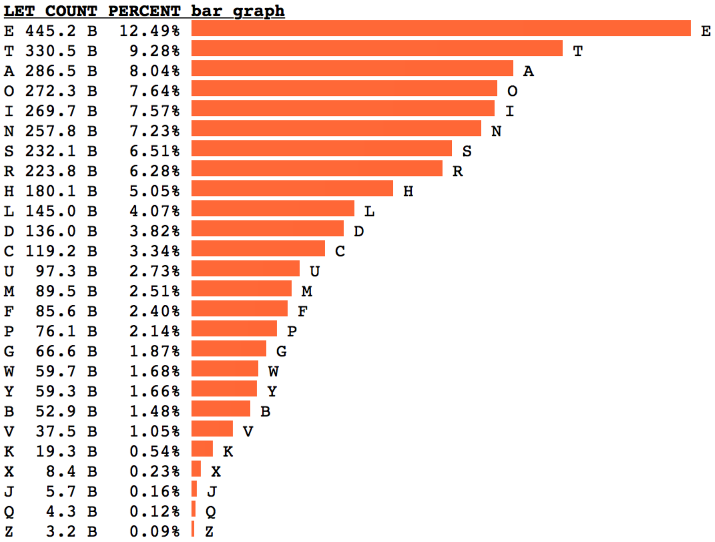
We consider the English language to be a stationary process, although it is not unreasonable to imagine the probabilities of N-grams changing in time. However, the probability spectrum of 1-grams has remained remarkably consistent for hundreds of years, and in general time differences in N-grams for $N > 2$ are averaged out in large corpora [7].

2.2.2 Words

Obviously, the space of all English words is too large to be reasonably represented by an information theoretic alphabet. Although with a large enough corpus it may be possible to get a good measurement of single word entropy, for the short term work represented by this study it is more appropriate to restrict methods to ones not requiring such a large database.

As covered in section 3.2, auto-correlation functions are another way to examine information structures in written English. In this construction, we only consider the words present in a novel length text, after eliminating words that occur with a frequency above some threshold. This alphabet size, depending on the setting of the threshold parameter, ranges between approximately twenty thousand and half a million words in the work done by Alvarez et al. [4].

The method of treating words based on their frequency can be recast as separating words out based on their use in a text. This technique stretches back to 1958 when Miller separated out “function words” and “content words”, identifying the former as words that serve mostly grammatical purposes and the latter as mostly nouns, verbs, adjective and adverbs [5]. Miller then showed that these two classes, although very roughly partitioned, produced length-frequency relationships with very different shapes.



LET	COUNT	PERCENT	bar graph
E	445.2 B	12.49%	E
T	330.5 B	9.28%	T
A	286.5 B	8.04%	A
O	272.3 B	7.64%	O
I	269.7 B	7.57%	I
N	257.8 B	7.23%	N
S	232.1 B	6.51%	S
R	223.8 B	6.28%	R
H	180.1 B	5.05%	H
L	145.0 B	4.07%	L
D	136.0 B	3.82%	D
C	119.2 B	3.34%	C
U	97.3 B	2.73%	U
M	89.5 B	2.51%	M
F	85.6 B	2.40%	F
P	76.1 B	2.14%	P
G	66.6 B	1.87%	G
W	59.7 B	1.68%	W
Y	59.3 B	1.66%	Y
B	52.9 B	1.48%	B
V	37.5 B	1.05%	V
K	19.3 B	0.54%	K
X	8.4 B	0.23%	X
J	5.7 B	0.16%	J
Q	4.3 B	0.12%	Q
Z	3.2 B	0.09%	Z

Figure 1: The frequency of single letters, ranked, in Peter Norvig's data sets [7].

3 Methods

3.1 Letters

For the calculation of the information theoretic quantities defined in section 2.1, we use the American Google Books corpus, which contains about 155 billion words [8]. Helpfully, this huge database was distilled by Google's Peter Norvig. The methodology for distillation is as follows:

- Norvig began with the Google books N-grams raw data set, which gives word counts of individual words in all books scanned by Google. Note - while we defined N-grams in section 2.1 as any possible sequence of variables from our alphabet, Google defines N-grams as sequences between two spaces.

- He then combined data from the over 200 years spanned by the Google books corpus. Any entry with a character other than the 26 letters of the alphabet was discarded, as well as case sensitivity.

- Norvig then produced tables of counts for words and letters, including specific counts for various letter positions within words.

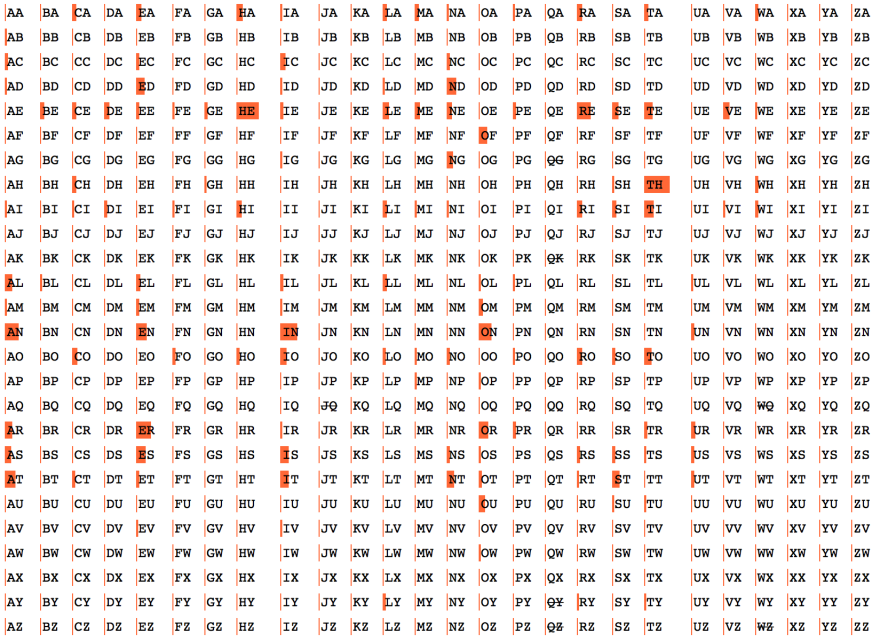
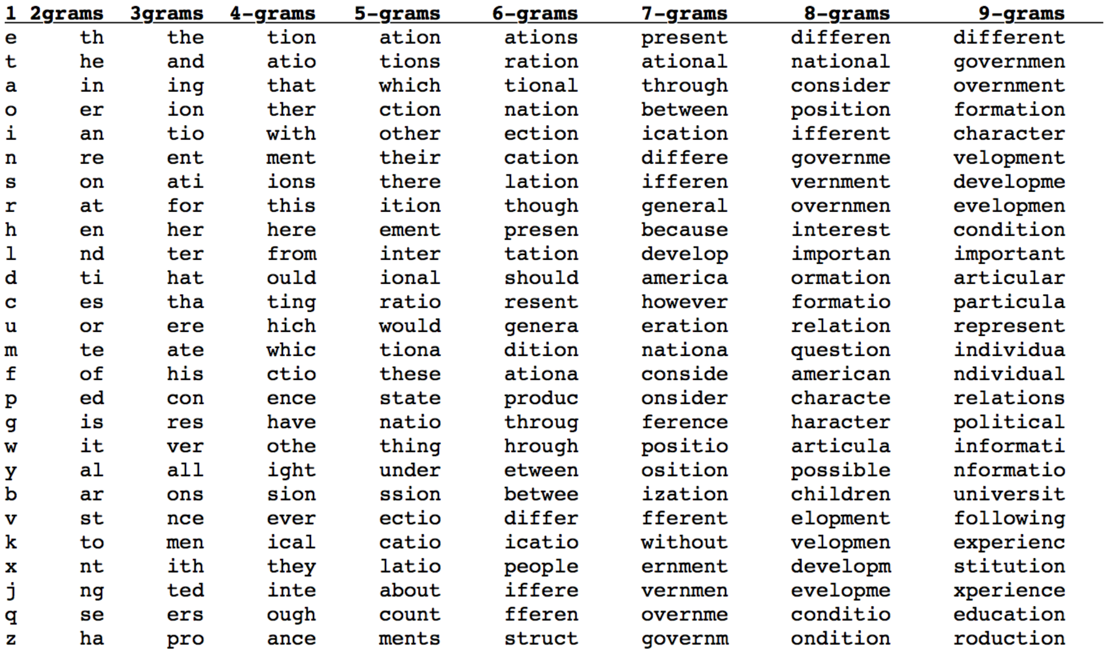


Figure 2: A plot of relative frequencies of various bigrams. Bigrams not appearing in Norvig's data are struck through [7].

Using these data sets, it was easily to pull the counts for N-grams up to $N = 9$, from a distilled database that consists of about 3.5 trillion letters. The frequencies of 1- and 2-grams can be seen in figures 1 and 2, respectively. Although the counts of N-grams for $N > 2$ are too extensive to show here, the 26 most common N-grams up to $N = 9$ are shown in figure 3.

Using these frequency counts, we make the assumption that by normalizing these counts we can immediately produce the probability of of N-grams (as defined in definition 2.4) up to $N = 9$. This assumes, of course, that the counts produced by Norvig's data sets accurately represent the frequencies of N-grams in all written English. The level of confidence we can place in this assumption is further discussed in section 4. Once the probabilities of N-grams is determined, the block entropies can be determined using definition 2.1. This was done using python, with the code below. Furthermore, the block entropy rates were found using definition 2.2, again using python.

```
def percentage(ngram):
```



1	2grams	3grams	4-grams	5-grams	6-grams	7-grams	8-grams	9-grams
e	th	the	tion	ation	ations	present	differen	different
t	he	and	atio	tions	ration	ational	national	governmen
a	in	ing	that	which	tional	through	consider	overnment
o	er	ion	ther	ction	nation	between	position	formation
i	an	tio	with	other	ection	ication	ifferent	character
n	re	ent	ment	their	cation	differe	governme	velopment
s	on	ati	ions	there	lation	ifferen	vernment	developme
r	at	for	this	ition	though	general	overnmen	evelopmen
h	en	her	here	ement	presen	because	interest	condition
l	nd	ter	from	inter	tation	develop	importan	important
d	ti	hat	ould	ional	should	america	ormation	articular
c	es	tha	ting	ratio	resent	however	formatio	particula
u	or	ere	hich	would	genera	eration	relation	represent
m	te	ate	whic	tiona	dition	nationa	question	individua
f	of	his	ctio	these	ationa	conside	american	ndividual
p	ed	con	ence	state	produc	onsider	characte	relations
g	is	res	have	natio	throug	ference	haracter	political
w	it	ver	othe	thing	hrough	positio	articula	informati
y	al	all	ight	under	etween	osition	possible	nformatio
b	ar	ons	sion	ssion	betwee	ization	children	universit
v	st	nce	ever	ectio	differ	fferent	elopment	following
k	to	men	ical	catio	icatio	without	velopmen	experienc
x	nt	ith	they	latio	people	ernment	developm	stitution
j	ng	ted	inte	about	iffere	vernmen	evelopme	xperience
q	se	ers	ough	count	fferen	overnme	conditio	education
z	ha	pro	ance	ments	struct	governm	ondition	roduction

Figure 3: A listing of the 26 most frequent N-grams, up to $N = 9$ [7].

```
    for i in range(len(ngram.iloc[:,0])):
        print(‘Percentage of’, ngram.iloc[i,0], ‘is’
                (ngram.iloc[i,1]/sum(ngram.iloc[0:,1]))*100)
```

```
def entropy(ngram):
    for i in range(len(ngram.iloc[:,0])):
        print('Entropy of', ngram.iloc[i,0], 'is',
                -(ngram.iloc[i,1]/sum(ngram.iloc[0:,1]))*
                log((ngram.iloc[i,1]/sum(ngram.iloc[0:,1])))/log(2))
```

In his original 1948 paper, Shannon attempted to make series approximations to English using a 27 symbol alphabet, with 26 letters and a space [1]. This model was designed to use the conditional probability distribution of N-grams to determine the next letter in a sequence. For example, to zeroth order the approximation assumes all symbols are independent and equally probable. To first order, the probability distribution of 1-grams is used to sample the next symbol. To second order, the next symbol is chosen from sampling from the probability distributions of bigrams starting with the previous letter. In other words, the next letter is chosen based on the conditional probability distribution of the 27 possible symbols that could follow E, or any other symbol. This method can be expanded further by sampling from the distribution of 3-grams, 4-grams, ect. This simple Markovian model, while not particularly intelligent, does a good job of producing English-esque text at around $N = 5$, for a large enough sample size. Examples of our Markovian text, for different sample spaces, are given in section 4.2.

3.2 Words

Although results from this project for studies on the level of words are still very preliminary, the methods from previous work will be discussed, with intent to reproduce and refine.

Beginning with methods developed by Zipf and Mandelbrot, connections long been made between word frequency, word rank in a frequency table, and word length [5], [6]. For natural languages, Zipf found that $\alpha \approx 1$, where α is defined in definition 2.5 [9]. Mandelbrot further generalized this result, proposing a "Zipfian" distribution of the form

$$f(r) \propto \frac{1}{(\beta + r)^{\alpha}}$$

where $\alpha \approx 1$ and $\beta \approx 2.7$ for natural languages [6]. Miller shows that the function form of this result can be reproduced with minimal assumptions, using only a Markov chain model in which there is some probability $p(*)$ of a space and probability $1 - p(*)$ of a letter. This model was found to work best with $p(*) \approx 0.2$, as shown in figure 4.

While this method reproduces the rank-frequency relation for written English fairly well, it does not produce a good approximation to the length-frequency curve. Such a curve ranks all possible N-grams from an alphabet of 26 letters by length, and plots against frequency. However, the Markovian model over estimates short words at the expense of mid-length ones. To attempt to understand this, Miller defines "function words" and "content words", showing that the two have qualitatively different length-frequency curves. These are defined intuitively, with function words including article, pronouns, numbers, conjunctions and more. A full list may be found in [5].

To further attempt to capture the importance of "function" words, Alvarez-Lacalle et al. devised the idea of a vector space defined by words within a text, and attempted

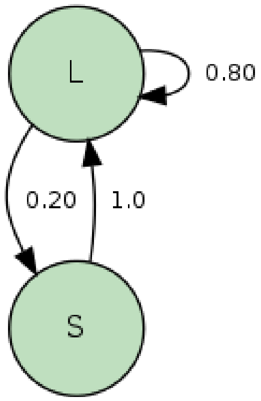


Figure 4: The simple model of words proposed by Mandelbrot.

to find long-range dynamical correlations [4]. Beginning with the set of words $w = (w_1, w_2, \ldots, w_i, \ldots, w_N)$, each word is associated with the number of times it appears in the text m_i and a basis vector $\mathbf{e}_i = (0, 0, \ldots 1, \ldots)$, where the 1 occupies the ith position in the basis vector. Alvarez et al. considered a sliding window A of text of length $a = 200$. At each point of the text this window of words comprises a "vector of attention", defined

$$\overrightarrow{V} = \left[\sum_j m_j^2(a)\right]^{-\frac{1}{2}} \sum_j m_j(a)\mathbf{e}_j$$

which may be considered the normalized sum over the vectors of the words within the window, weighted by their frequency $m_j(a)$. To capture the the most important concepts or themes in the text, Alvarez-Lacelle et al. built a connectivity matrix M based on the co-occurrence of words within the window A. The entry M_{ij} counts how often w_i occurs within a distance $\frac{a}{2}$ of the word w_j. This matrix is normalized, and to keep the matrix a manageable size a threshold value m_{thr} is defined to be the number of occurrences a word must exceed to in included in M. Finally, single value decomposition is performed, changing the basis of M into the space of "concept" vectors, which have entries consisting of the thematic elements of a text. We propose that these concept vectors $\mathbf{v}_j$ are a more mathematical representation of the intuitive idea of "function words", which aim to capture the thematic elements of a text, minus its grammatical glue. A sample of these concept vectors may be seen in section 4.

We return to the idea of the vector of attention $\overrightarrow{V}$, now defined in terms of the new

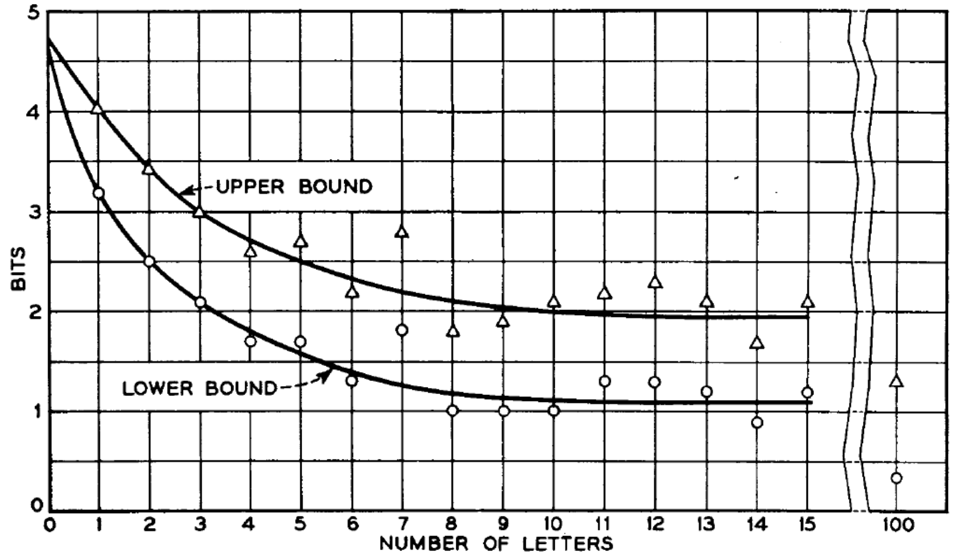


Fig. 4—Upper and lower experimental bounds for the entropy of 27-letter English.

Figure 5: Shannon's original measurements of the upper and lower bounds on entropy rate of written English, up to $N = 100$.

basis $\mathbf{v}_j$,

$$\overrightarrow{V} \approx \sum_{j=1}^{d} S_j(t)\mathbf{v}_j$$

where d is an adjustable parameter representing how many concept vectors we chose to include in our vector space. As our sliding window A moves through the text, $\overrightarrow{V}$ moves through the vector space spanned by our concept vectors. If this vector space failed to capture structure in the text, we would expect $\overrightarrow{V}$ to perform a random walk. To test this, Alvarez-Lacelle studied its autocorrelation $C(\tau) = \langle \overrightarrow{V}(t) \cdot \overrightarrow{V}(t+\tau) \rangle_t$, where $\langle \cdot \rangle_t$ is the time average. Note that t is defined as $t = l + \delta t$, where l is the distance into the text and δt is the time it takes for an average reader to read one word [4].

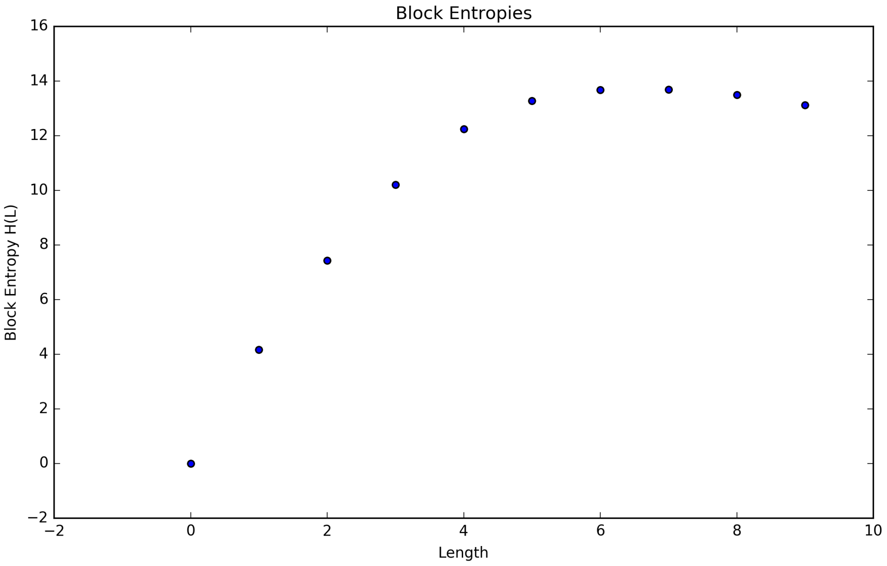


Figure 6: The block entropies up to length 9, as calculated from Norvig's data.

4 Results

4.1 Letters

Shannon's original measurements of the entropy have stood up surprisingly well, considering the limitations on data and methods when he performed his experiments in 1950. Unable to access large databases for $N > 3$, he performed measurements for N up to 100 by asking human subjects to guess the next letter in a sequence and normalizing the number of guesses required to reach the correct letter. Despite this method, his estimate of $h_\mu = 1$ bit/letter has been a robust measurement. Shannon's upper and lower bounds are reproduced in figure 5.

The measurements performed in this project, shown in figures 6 and 7, are somewhat problematic. As easily seen, around $N = 6$ the block entropy rate becomes impossibly low, and even dips below zero, which is obviously incorrect. This measurement stems from a systematic sampling error in the database.

One large problem, which can also be seen in figures 2 and 3, is the lack of spaces represented in Norvig's database. The database, due to the way the Google books corpus

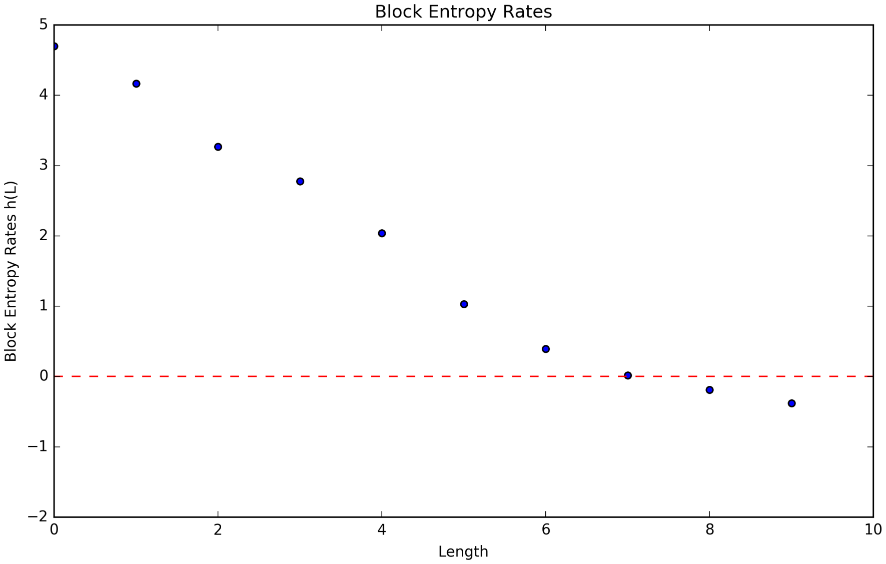


Figure 7: The entropy rates up to length 9, as calculated from Norvig's data.

defines N-grams, contains only sequences of letters between spaces, and does not retain positional data for these sequences. In other words, the phrase "red skies at night", as analyzed by Norvig, will give 11 different 1-grams, 11 2-grams, 6 3-grams, 3 4-grams, 1 5-gram and no grams of length greater than 5. This artificially lowers the number of N-grams for $N > 2$, and the problem becomes worse with increasing N. This problem is systematic to the databases as compiled, and though sampling errors may be adjusted for using statistical methods, the loss N-grams across words is a huge problem, espcially considering the average length of a single word in the English language is 4.8 letters [7]. This may be rectified using a few different methods.

Firstly, Google books does offer large datasets of N-grams greater than $N = 1$, using their definition. Again, Google's definition of an N-gram is a sequence of letters, bounded by spaces. If the average length of a word in the English language is 4.8, the average length of a Google defined 2-gram should be 9.6. This is already a better sampling environment for our version of N-grams. By starting with Google's data-base of 3-grams, and performing Norvig's algorithm on this dataset, it should eliminate the problems outlined above.

The second method is to pull a database that retains complete positional data of words and sentences, such as Project Gutenberg's large database of full text books. While this corpa is not as large as the Google book corpa, it does automatically eliminate the problems outlined above. It may also be more appropriate a database to use for some of the word-level analysis outlined in section 3.2. Moving forward, we plan to implement these ideas and compare results across corpora.

As mentioned in section 3.1, Shannon used a Markovian approximation to English to produce English-esque texts. For fun, this was reproduced in this project, with results given below.

After feeding the model the full works of H. P. Lovecraft:

> N = 2: ed ing em th had then yiners. an ithome se unt whosecto cur putch eprate of thiscie, hathe not rand but thic arrat wasen, th the stude hily tomed, arts ing asird las shose mr. he agodit.
>
> N = 4: inted to themself would ancient complex, even tumuli, which were seldom was known shelves, true storillare punge precians about bring the new-four eyes about of his could his possibility
>
> N = 6: the books the grey membrane rolls back on the most made, and here some charles's noticed that terrible by little had been my brain that the backs others lurking in the end to the alienists

The works of Jane Austen:

> N = 2: ho!–ither for he mords quaid le on elf man voling exprouloverescomence onny pure ince, the beinnot and not lifes of mot.empactel!
>
> N = 4: i assurance thoughts as you must like hide found liable. but i to impatiencies about their confide in these easy time been, or batest indeed, and elings her.
>
> N = 7: prejudiced again; but so think it influence on the sound apeice to take him never was greatly, very bright at hartfield; acknowledge that spirit, every sentence

And the complete works of Shakespeare:

> N = 4: thy sworn dare my bucking, as than speak our coast hath base you lord.- wherefore, undone. me? base you now- who ham. o my lord the gross. 'tis trumpet!

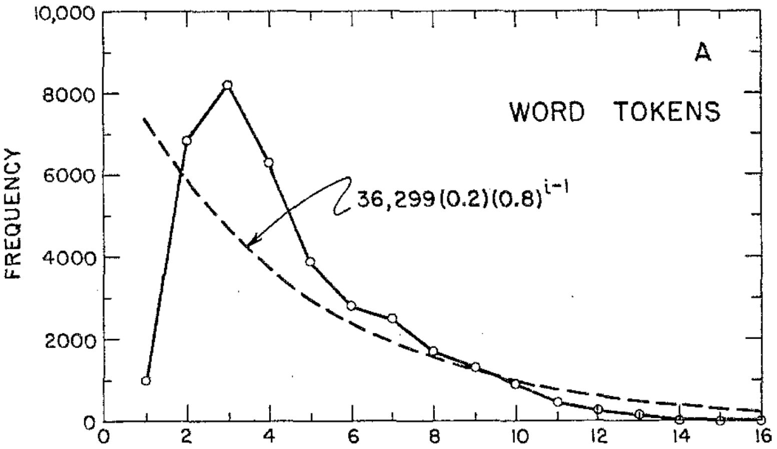


Figure 8: The frequency of words produced by Madelbrot's model versus length, as plotted against a statistical sampling by Miller [5].

> N = 7: suffolk's cloud, and cull'd the lord i have seen the king; hear them lie till i rouse yesterday suspire, are thine until he be she? julia. what title to alter not at all?
>
> N = 9: act v. scene i. petruchio. nay, he must not know why i should a villain. believe' said she's gone, thou art hermione.

We see, as expected, that the approximation becomes more like realistic English as N is increased. However, the downfall of increasing N is also clear in the $N = 9$ sample of Shakespeare - *act v. scene i. pertruchio* has been taken wholesale from *The Taming of the Shrew*, because of the large N and relatively short sample size.

4.2 Words

Returning to Mandelbrot's construction of words as a simple Markovian process, shown in figure 4, we examine the length-frequency statistics implied by such a model. This model produces

$$N_i = N * p(S|L)p(L|L)^{i-1}$$

words of length i, where N is the number of words generated in the sample. This

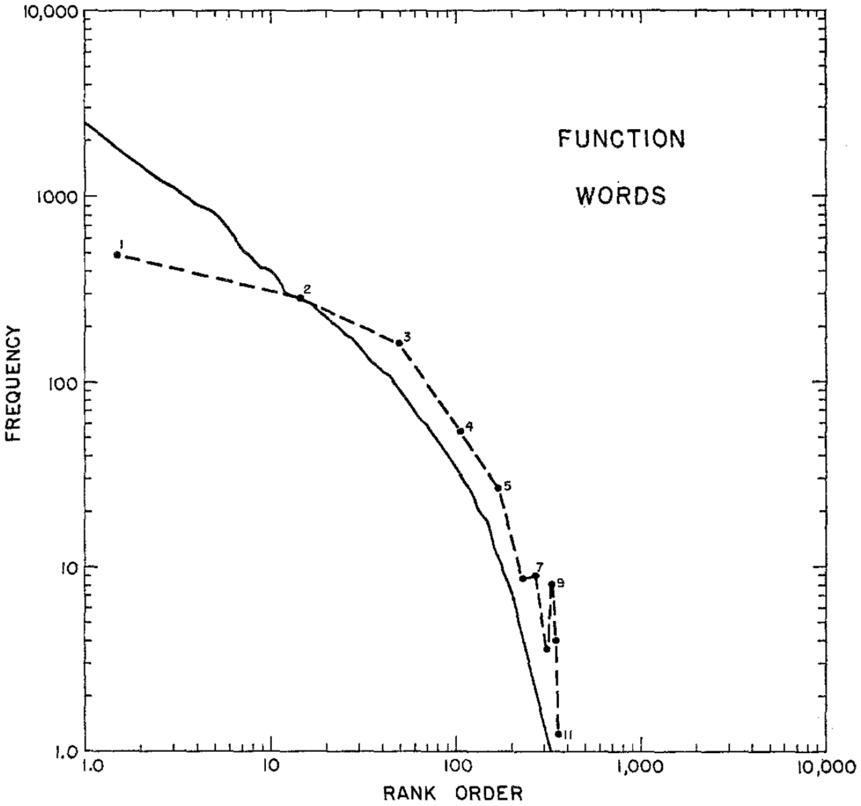


Figure 9: The length-frequency relation for function words in written English is shown by the dashed curve. The solid line is the rank-frequency relation [5].

function is plotted by Miller in figure 8. Obviously, this model gives a preference for shorter words, and underestimates the value of mid-length words. Considering that the most average word length in English based on frequency is 4.8, this model is clearly insufficient. To attempt to understand this, Miller separates out "function words" and "content words", as defined in section 3.2, to attempt to find a reasonable length rank-frequency relation. The average length of a function word was 3.13 letters, while the average length of a content word was 6.47.

Miller then plotted both the rank-frequency and the length-frequency, showing that while rank- and length-frequency curves are similar within the word categories, they are significantly different between function and content words. Function words, in general, have a much higher average frequency than content words, and it drops off significantly.

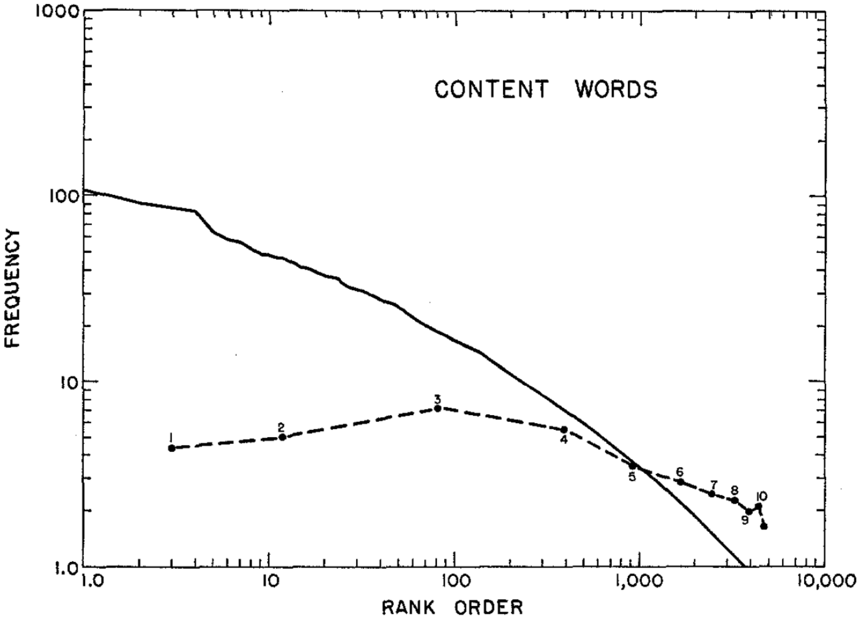
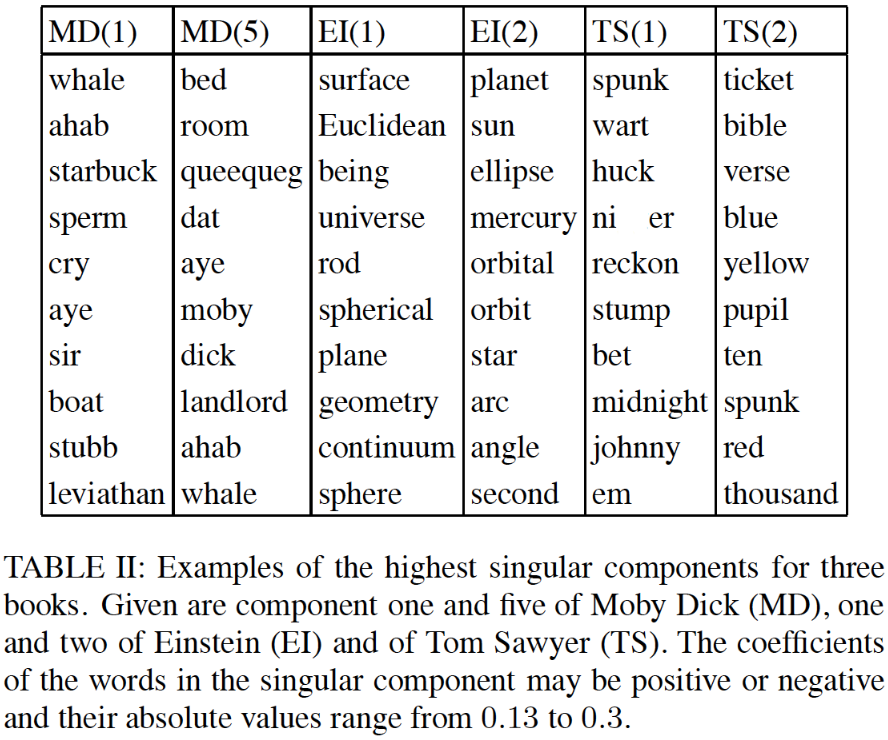


Figure 10: The length-frequency relation for content words in written English is shown by the dashed curve. The solid line is the rank-frequency relation [5].

The drop off in length-frequency curve tells us that there are not many content words above a certain length, and the rank-frequency curve informs us that there is a body of function words that are used commonly, and not many that are rarely used. In general, this curve does not have the linear shape implied by Zipf's law [5].

Meanwhile, both content words plots have flatter slopes, indicating a much more democratic use among content words. The drop off in rank-frequency does indicate that there are content words that are much more popular than others. However, note that this drop off in popularity is well beyond an order of magnitude when compared to the drop off in function words. Furthermore, the length-frequenct plot does not drop off significantly with length, telling us that popularity of content words does not depend on length. Again, this curve deviates from the Zipf power law expected. When taken together, the two plots do give back the Zipf law, but it is certainly interesting to be able to decompose this rank-frequency relation based on an intuitive understanding of the purpose of English words. This project is currently in the process of reproducing this work with larger corpa, as Miller's results depend on a sample of only 20,000 words.

The idea of separating out words based on usage in a more mathematical way was



MD(1)	MD(5)	EI(1)	EI(2)	TS(1)	TS(2)
whale	bed	surface	planet	spunk	ticket
ahab	room	Euclidean	sun	wart	bible
starbuck	queequeg	being	ellipse	huck	verse
sperm	dat	universe	mercury	ni er	blue
cry	aye	rod	orbital	reckon	yellow
aye	moby	spherical	orbit	stump	pupil
sir	dick	plane	star	bet	ten
boat	landlord	geometry	arc	midnight	spunk
stubb	ahab	continuum	angle	johnny	red
leviathan	whale	sphere	second	em	thousand

TABLE II: Examples of the highest singular components for three books. Given are component one and five of Moby Dick (MD), one and two of Einstein (EI) and of Tom Sawyer (TS). The coefficients of the words in the singular component may be positive or negative and their absolute values range from 0.13 to 0.3.

Figure 11: A reproduction of Table II from Alvarez et al.'s paper *Hierarchical structures induce long-range dynamical correlations in written texts* [4].

done by Alvarez et al., followed by the use of auto-correlation functions to find hierarchical structure in novels. Reproduction of this work is currently underway, but some results from that paper are reproduced here to give an example of results of word-based analysis.

In figure 11 a sampling of the concept vectors $\mathbf{v}_j$ from various works are listed. The rankings given are related to the singular value, assuming that a high singular value is related to importance in the text. As should be clear, these vectors pull out the most important thematic elements from the text being studied. For example, the first concept vector from *Moby Dick* immediately presents the main characters, location of the story and the tone of the story. Similarly, the second concept vector from *The Adventures of Tom Sawyer* brings to mind the short story of the tickets and bibles from the classic childhood novel.

The log-log plot of the results of the autocorrelation function as applied to *The Adven-*

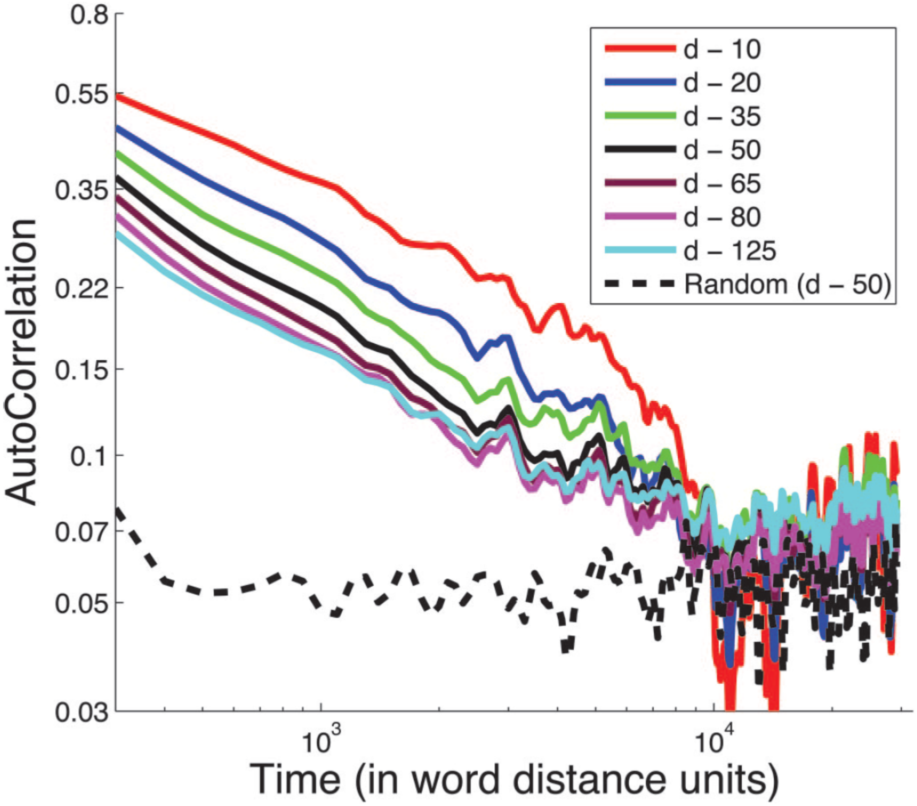


Figure 12: A log-log plot from Alvarez et al. of the autocorrelation function for *The Adventure of Tom Sawyer* for different values of d. A randomized version of the book is examined in the black dashed line [4].

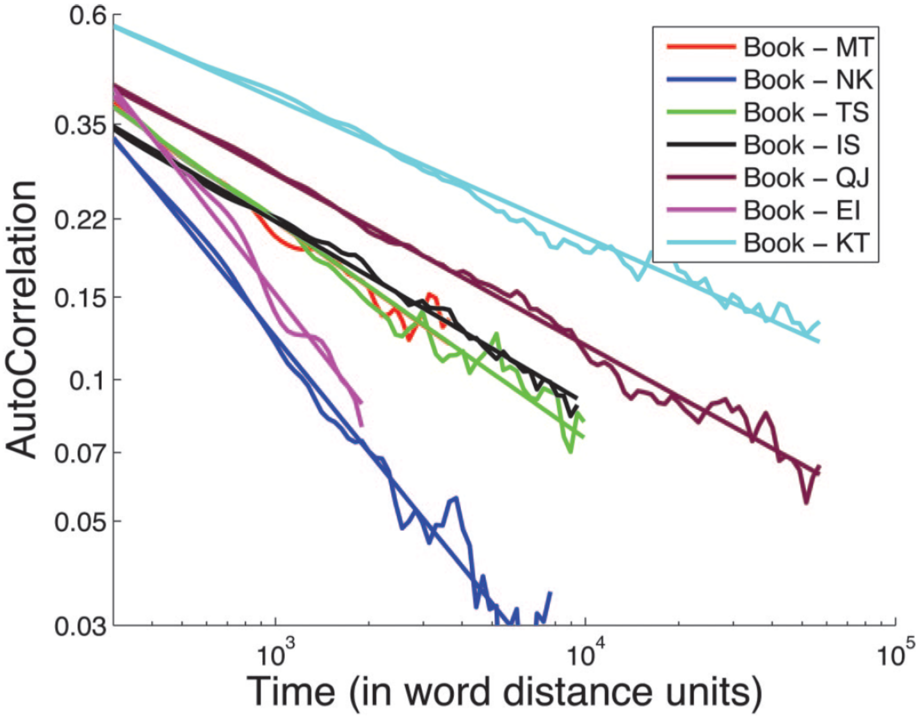


Figure 13: A comparison of the autocorrelation functions for the books analyzed by Alvarez et al. [4].

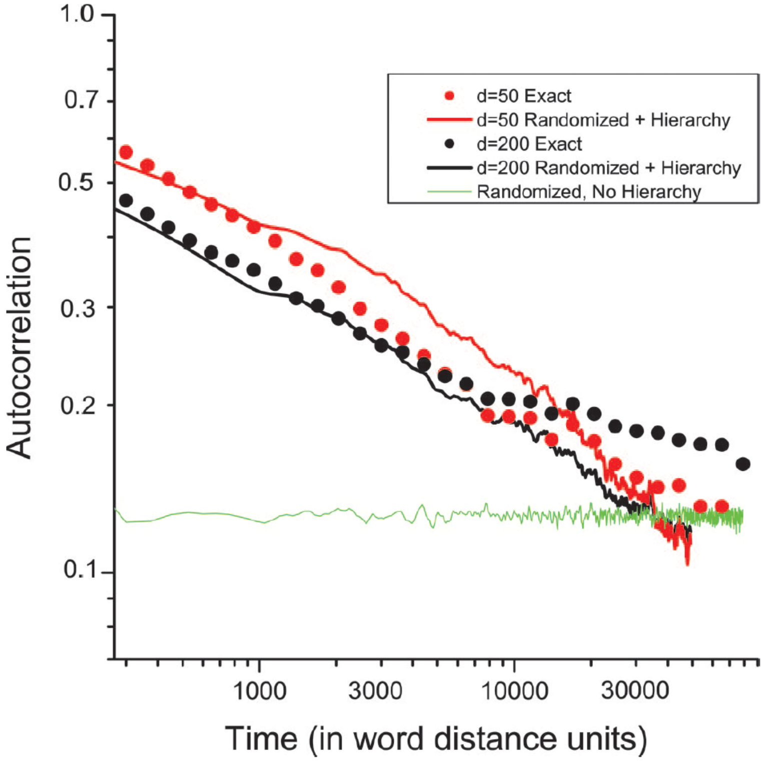


Figure 14: Comparison of the autocorrelation fuctions of the *Critique of Pure Reason* by Kant, original, randomized while maintaining hierarchy, and completely randomized [4].

tures of Tom Sawyer is shown in figure 12. The parameter d, which represents how many concept vectors are considered in the concept vector space, is varied from 10 to 125. A comparison is made with a randomized version of the text, which clearly indicates that the autocorrelation function captures the structure present in the novel. For short distances the correlation is about half of the maximum value, and remains above the level of noise on the order of $1,000$ words.

As d is decreased, the autocorrelation function approaches a straight line on the log-log plot, which indicates a power law relationship between correlation and word distance. The range over this behavior is evident depends on the source, which will have some inherent noise. To estimate this, the autocorrelation function of a randomized version of the book is also shown. The average value of this function is determined by properties of the specific sample.

In figure 13 different autocorrelation functions from various novels are compared, showing no particular relationship between power law exponent and genre. The books analyzed, in order, are *The Metamorphosis* by Kafka, *Naked Lunch* by Burroughs, *The Adventures of Tom Sawyer* by Twain, *The Illiad* by Homer, *Don Quixote* by Cervantes, *Relativity* by Einstein, and *Critique of Pure Reason* by Kant.

The existence of power laws is often used to infer hierarchical structure in the source data. It seems reasonable to imagine a novel in particular as having this type of structure, considering that is is separated into sections, which are further separated into chapters and futher into paragraphs, ect. Alavarez et al. investigated whether these conclusions were meaningful by running the autocorrelation function on three different texts: the original, the original with words randomized within the existing hierarchical structure of the book, and the book completely randomized. The results of this work are shown in figure 14. Since destroying the hierarchical structure of the book also eliminates the power law, it is reasonable to claim that this power law originates from the this structure.

5 Conclusion

Although our analysis of the Peter Norvig data sets was less successful than originally hoped, overall the methods for studying written English in an information theoretic context are robust and promising. We have proposed throughout this report ideas for moving forward with this work, including changing the database used for sampling probabilities of N-grams to achieve better bounds on the entropy of English. Furthermore, current work is underway on accomplishing recreation of the results outlined in section 3.2.

An additional area of study that is of interest is examining written English on the syntactical level. Such methods, known as grammars, have been discusses by linguists and formal language theorists since being introduced by Noam Chomksy in the late 1950s [10]. This area holds great promise for the implementation of ϵ-machines, which may be layered with the Shannon Markovian English approximation to produce more realistic samples. The author is currently experimenting with this, by layering statistical models at various levels of English to improve predictive capability.

References

[1] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27, October 1948.

[2] James P. Crutchfield and David P. Feldman. Regularities unseen, randonmess observed: Levels of entropy convergence. *Chaos*, 13(1), March 2003.

[3] C. E. Shannon. Prediction and entropy of printed english. *Bell System Technical Journal*, 27, October 1948.

[4] E. Alvarez-Lacelle, B. Dorow, J. P. Eckmann, and E. Moses. Hierarchical structures induce long-range dynamical correlations in written texts. *Proceedings of the National Academy of Sciences*, 103(21):7956–7961, May 2006.

[5] G. A. Miller, E. B Newman, and E. A. Freidman. Length-frequency statistics for written english. *Information and Control*, 1, 1958.

[6] G. A. Miller. Some effects of intermittent silence. *The American Journal of Psychology*, 70, July 1957.

[7] URL `http://norvig.com/mayzner.html`.

[8] URL `http://googlebooks.byu.edu`.

[9] G. K. Zipf. The psychobiology of language. 1935.

[10] N. Chomsky. Three models for the description of language. *IEE Transactions on Information Theory*, 1956.