Investigating Word Pair Correlations

Jackson Crook

May 10, 2015

1 Introduction

The idea for this project came when I first heard about a mathematical conjecture known as the Gaussian Unitary Ensemble Hypothesis. The essential claim of the GUE hypothesis is that the relative spacings of eigenvalues of a very large random matrix of a certain type (that is, a Gaussian Unitary Ensemble) should resemble, or be underlying identical, to the relative spacings of zeros of the Riemann zeta function, which is used to predict prime numbers.

Since the 50s, this same particular spacing pattern has also been known to describe the scattering resonances of very heavy nuclei (though what exactly a scattering resonance is is not relevant to this project), and, recently, to describe several other phenomena that can be represented as events on a one-dimensional spectrum, such as birds resting on power lines, bus arrival times, and parallel parked cars.

This led me to wonder about the spacing of words, to see if they too could be described in this way. As it turns out, the technique used in the GUE hypothesis that allows for the quantitative description of these one-dimensional spacing patterns is known as the pair correlation function, commonly written as g(r).

The pair correlation function describes how the density of certain events, or, let's say, particles in solution, varies as a function of distance from a reference particle. To do this, $g(r)$ gives the probability of finding a particle at a distance r from another particle. The formal, general definition of $g(r)$ restricts r to continuous variables in any number of dimensions and involves some calculus. As a result, due to the discrete, combinatorial nature of language, a slightly altered interpretation of the pair correlation function is necessary.

2 Word Pair Correlation Function

The "word pair correlation function" can be defined for a single word type, or across all words.

Across all words, let's define it as the probability a pair of tokens, spaced r distance apart, are the same word. So if you picked a position t at random, what's the chance the same word is at $t + r$? (Restriction: you sample t only from positions $\{1..N - r\}$ where N is the number of tokens in the corpus.) This is:

$$g(r) = p(w_{t+r} = w_t)$$

For a specific word v, there are two possible ways to define it. We should standardize on one.

$$g_{cond}(r, v) = p(w_{t+r} = v \mid w_t = v)$$

versus

$$g_{joint}(r, v) = p(w_{t+r} = v,\ w_t = v)$$

the joint version (the latter) has much smaller numerical values than the first, and they're especially smaller for rare words. For this reason, when doing single word pair correlations, we'll use the conditional version, as its values will be easier to interpret.

Finally, $g(r)$ could be defined for a set of words — for example, to analyze a single paircorr function curve for all names, or all verbs, etc. In fact, the word set paircorr has two versions, a $set - inclusive$ version and a $set - exclusive$ version.

Let the set of words be A. For the set-inclusive version, the paircorr would look like this:

$$g(r, A) = p(w_{t+r} \in A \mid w_t \in A)$$

This is equivalent to rewriting all words in the set A to a special symbol, and computing the word paircorr for that symbol (specifically g_{cond}).

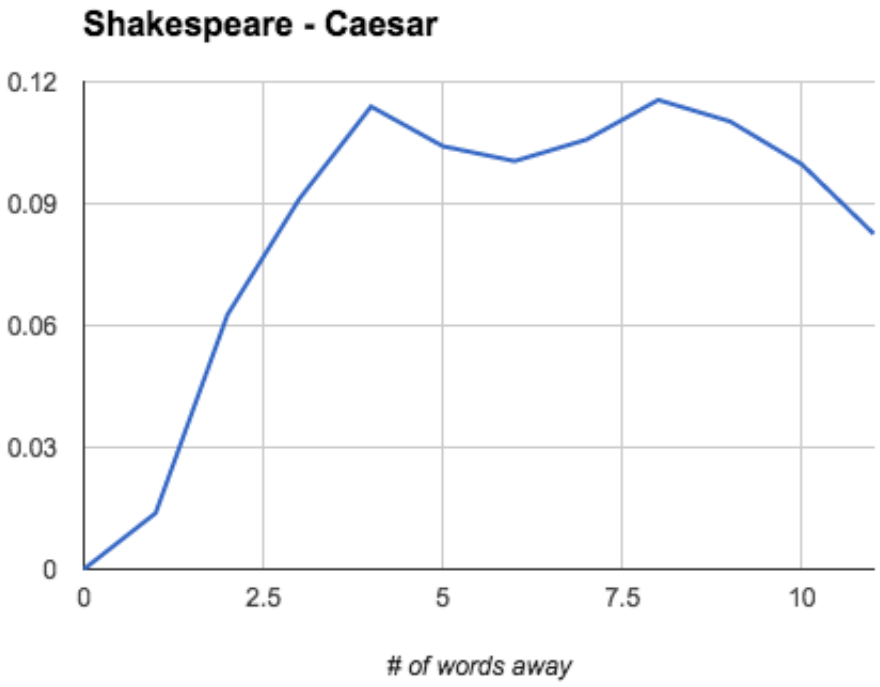
For the set-exclusive version, the paircorr would look like this:

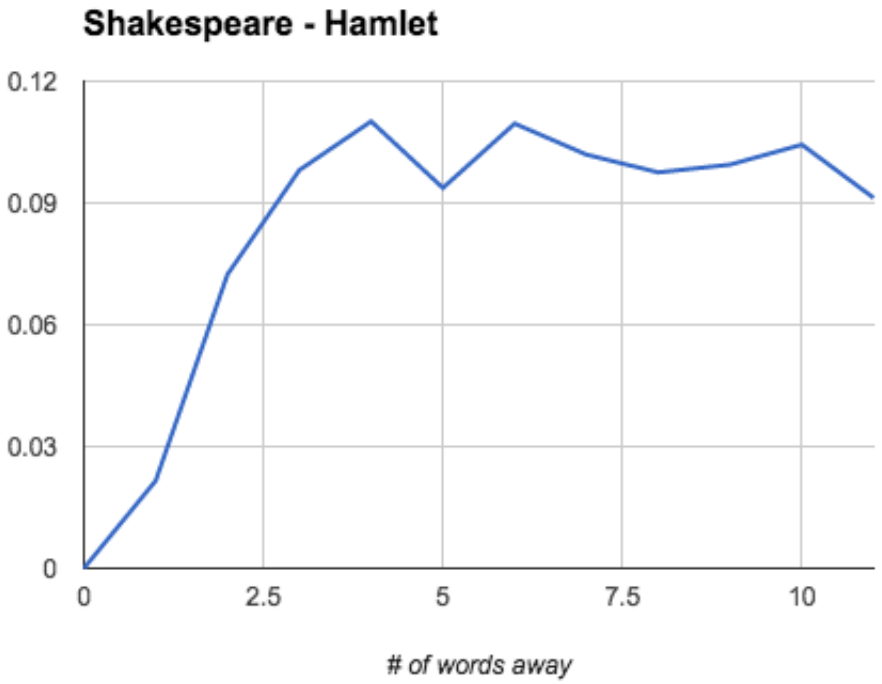
$$g(r) = p(w_{t+r} = w_t \mid w_t \in A)$$

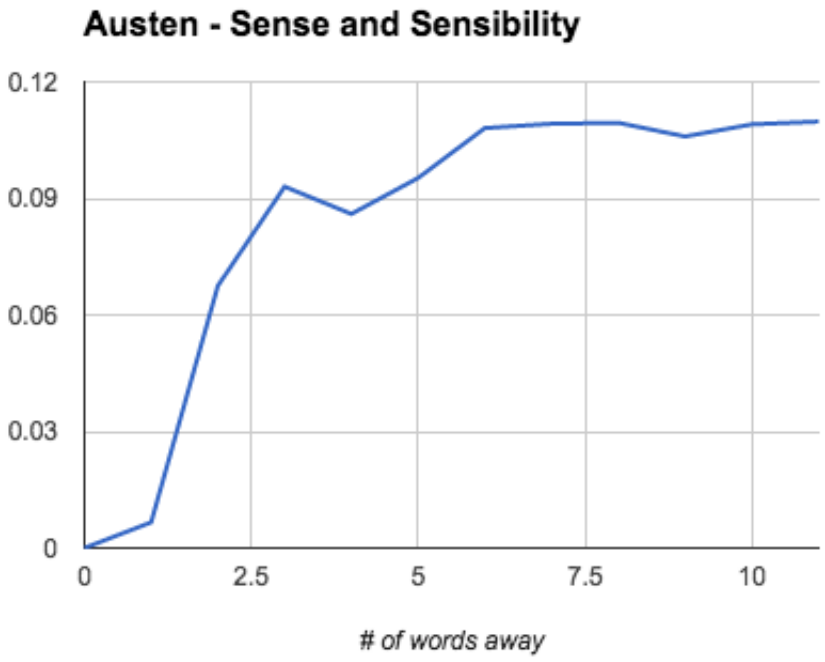
This is equivalent to rewriting all words not in the set A to a special symbol, and ignoring that symbol when evaluating the pair correlation.

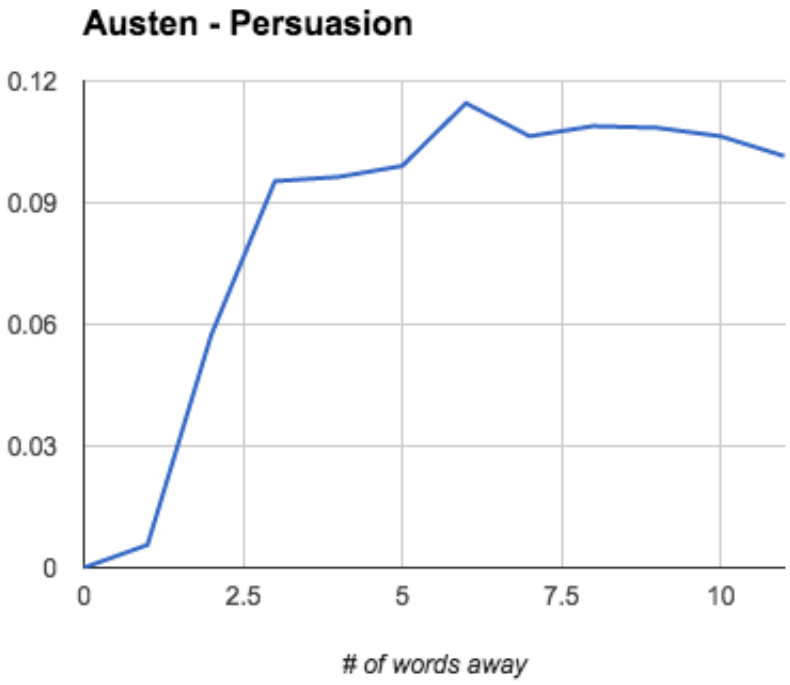
3 Pair Correlations of Famous Texts

To get a feel for the pair correlation, I took the all-word paircorr for a number of well known texts. To make the curves easier to compare, they have been normalized so the area under the graph sums to one. This is not how a true pair correlation is normalized, and this normalization will vary depending on the max value of r in consideration; however, in the context of linguistics as apposed to math or physics, the typical normalization technique (which I will not go into) does not work well, so I just decided on this.

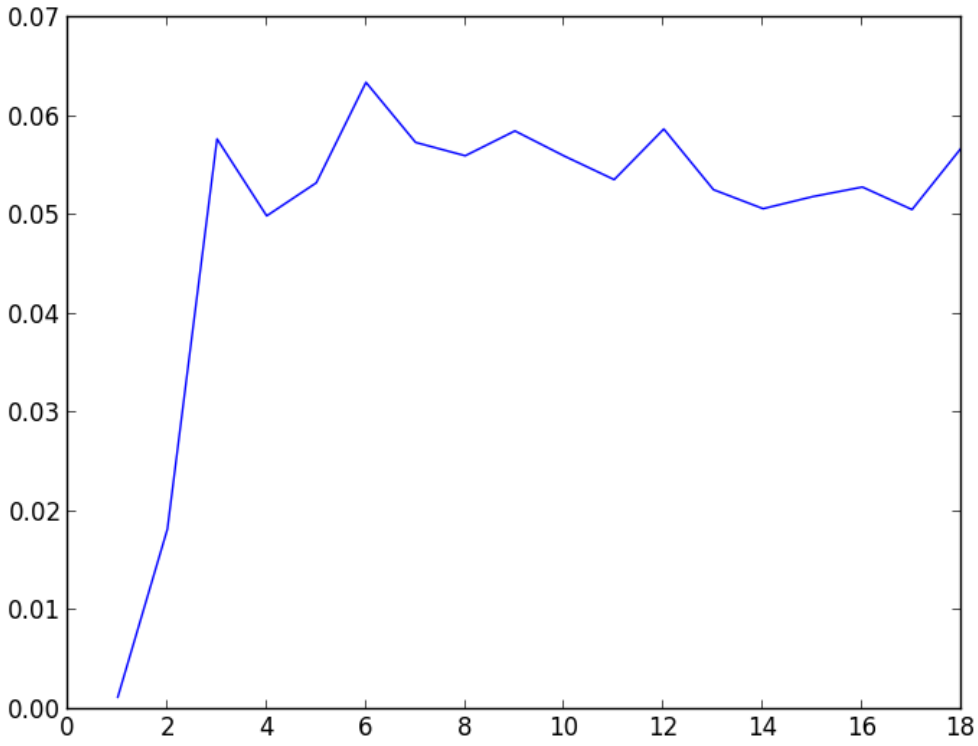

Shakespeare - Caesar
0.12
0.09
0.06
0.03
0
0
2.5
5
7.5
10
# of words away


Shakespeare - Hamlet
0.12
0.09
0.06
0.03
0
0
2.5
5
7.5
10
# of words away

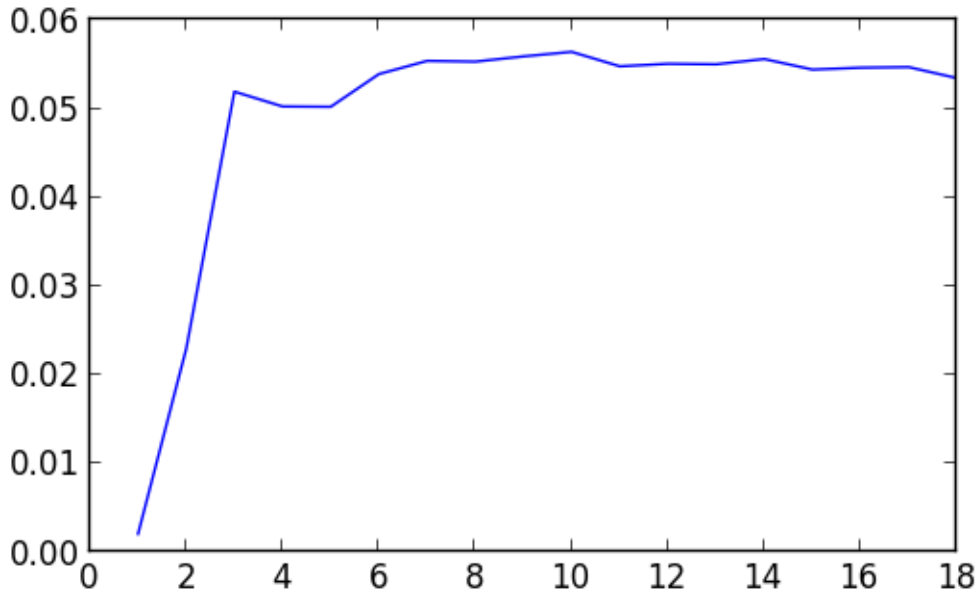

Austen - Sense and Sensibility
0.12
0.09
0.06
0.03
0
0
2.5
5
7.5
10
# of words away


Austen - Persuasion
0.12
0.09
0.06
0.03
0
0
2.5
5
7.5
10
# of words away

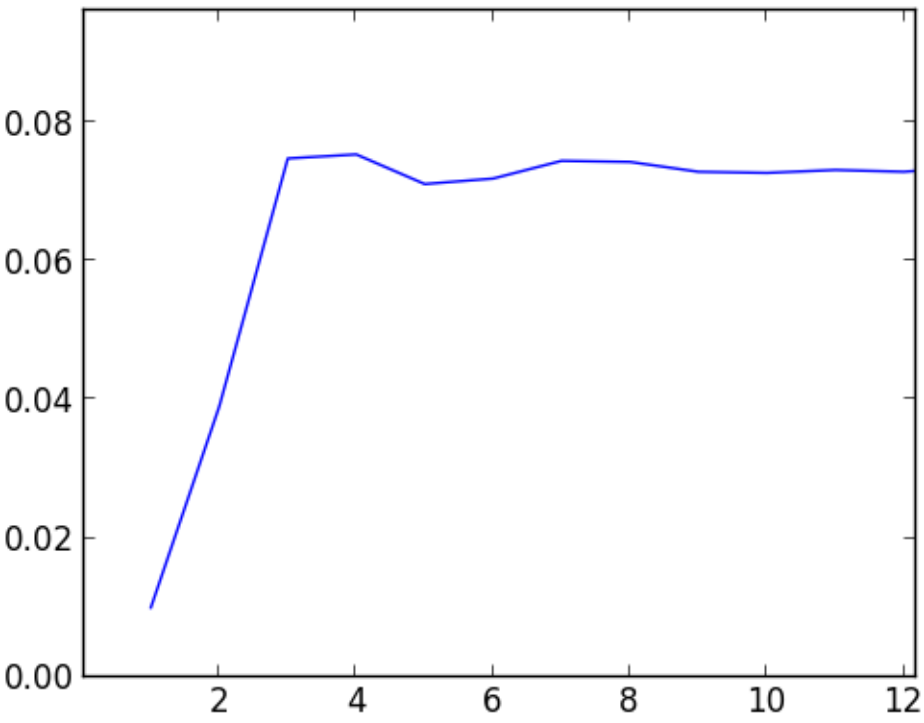
The King James Bible:



The Brown Corpus:



Moby Dick:



However, the pair correlations for these texts don't meaning anything without something to compare them to. We need a null hypothesis of what these pair correlations should look like. For that, let's look at what the pair correlations would look like for Markov language models.

4 Pair Correlation of Markov Models

Consider the case $r = 1$. $g(1)$ can be rewritten in a history conditional form as follows.

$$g(1) = p(w_{t+1} = w_t) \tag{1}$$
$$= \sum_v p(w_{t+1} = v,\ w_t = v) \tag{2}$$
$$= \sum_v p(w_{t+1} = v \mid w_t = v)\, p(w_t = v) \tag{3}$$

Step 2 decomposes the probability into the sum over all words that are possible. Step 3 is definition of conditional probability.

A zeroth order (independent unigrams) assumption yields the following equivalency:

$$p(w_{t+1} = v \mid w_t = v) = p(w_{t+1} = v) = p(w_t = v)$$

With this, we can simplify step 3 and fully derive $g(r)$ under the zeroth order assumption.

$$\sum_v p(w_{t+1} = v \mid w_t = v)\, p(w_t = v) \tag{4}$$
$$= \sum_v p(w_{t+1} = v)\, p(w_t = v) \tag{5}$$
$$= \sum_v [p(w_t = v)]^2 \tag{6}$$
$$= g(r) \tag{7}$$

Now let's consider a first-order (bigram) model. For $r = 1$, $g(r)$ is already defined for each word type, by definition of a first-order model. So:

$$p(w_{t+1} = w_t) = \sum_v p(w_{t+1} = v \mid w_t = v) * p(w_t = v)$$

Now g(2):

$$p(w_{t+2} = w_t) = \sum_v p(w_{t+2} = v \mid w_t = v) * p(w_t = v)$$

The conditional probability $p(w_{t+2} = v, w_t = v)$ can be put into terms that are defined in the model as follows:

$$p(w_{t+2} = v \mid w_t = v) \tag{8}$$
$$= \sum_u p(w_{t+2} = v, w_{t+1} = u \mid w_t = v) \tag{9}$$
$$= \sum_u p(w_{t+2} = v \mid w_{t+1} = u)\, p(w_{t+1} = u \mid w_t = v) \tag{10}$$
$$\tag{11}$$

When we combine 10 with 4, we get $g(2)$ under a first order model:

$$\sum_v [\sum_u p(w_{t+2} = v \mid w_{t+1} = u) * p(w_{t+1} = u \mid w_t = v)] * p(w_t = v)$$

Here, the gap between w_t and w_{t+2} has been 'bridged' by the non-specific token u . By bridging this way twice, this expression can be extended to $r = 3$:

$$\sum_v [\sum_u [\sum_q [p(w_{t+3} = v \mid w_{t+2} = q) * p(w_{t+2} = q \mid w_{t+1} = u)] p(w_{t+1} = u \mid w_t = v)] p(w_t = v)$$

and all other r (with a slight change in notation):

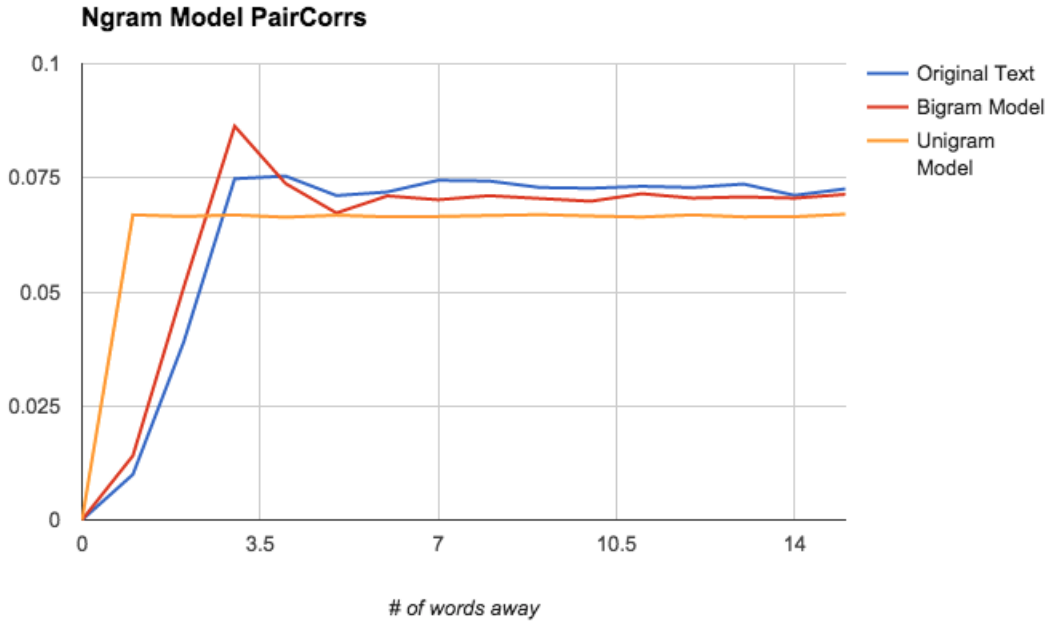
$$g(r) = \sum_{v_0} [\sum_{v_1} [...[\sum_{v_{r-1}} [p(w_{t+r} = v_0 \mid w_{t+r-1} = v_{r-1}) p(w_{t+r-1} = v_{r-1} \mid w_{t+r-2} = v_{r-2})] \tag{12}$$

$$* p(w_{t+r-2} = v_{r-2} \mid w_{t+r-3} = v_{r-3})] p(w_{t+r-3} = v_{r-3} \mid w_{t+r-4} = v_{r-4})]...] p(w_t = v_0) \tag{13}$$

For clarification, here is g(r) for a first-order model represented in pseudocode.

$V = set(words\ in\ book)$
$g(r) = 0$
    $for\ w_0\ in\ V:$
        $for\ w_1\ in\ V:$
        ...
            $for\ w_{r-1}\ in\ V:$
                $g(r) +=\ p(w_0) * p(w_1 \mid w_0) ... * p(w_r \mid w_{r-1}) * p(w_0 \mid w_{r-1})$
$return\ g(r)$

So what do these pair correlations actually look like? Computing the actual values described above proved to be incredibly time consuming because of those nested for-loops. So, to get an approximation, I used Moby Dick as a corpus to generate 10,000,000 words under a zeroth order model and 1,000,000 words under the first order model and calculated g(r) for both:

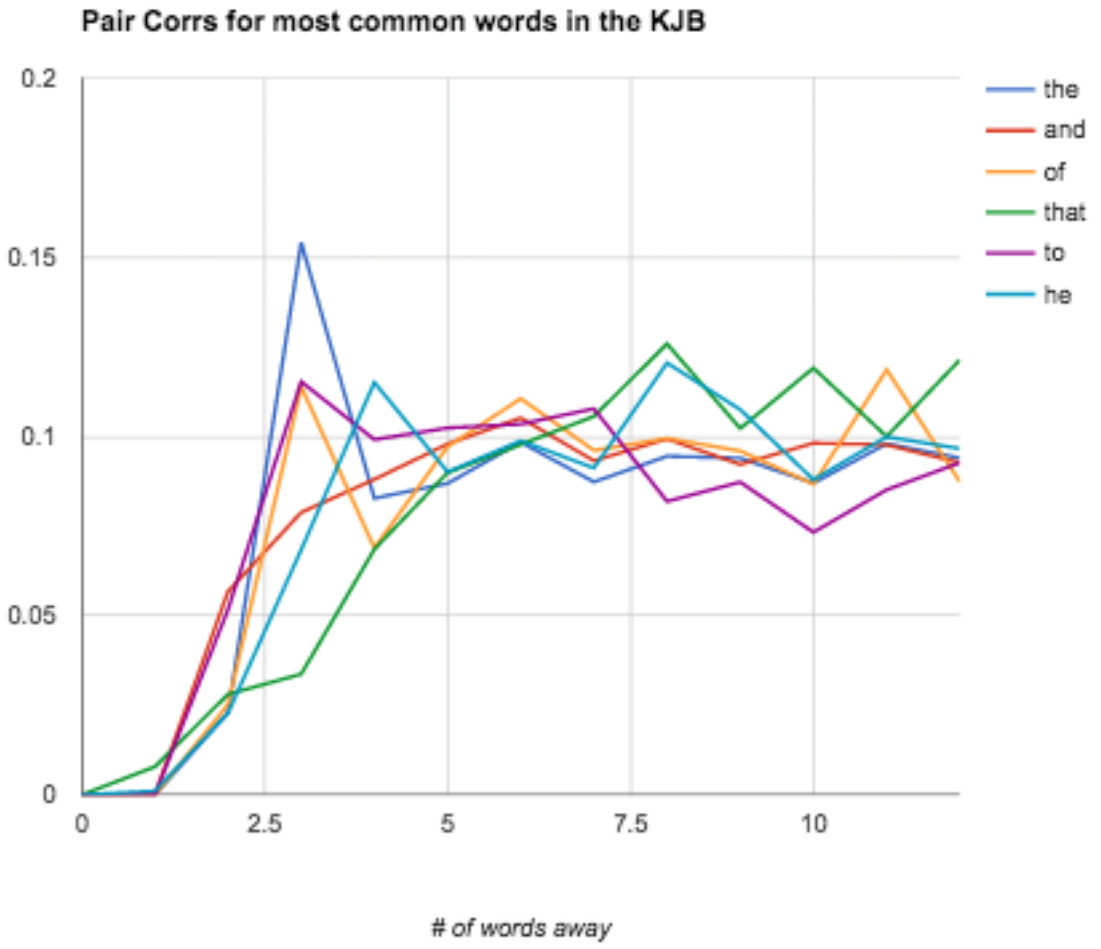


As one would expect, the unigram model pair corr is simply constant. (Note: the graph is a little misleading; g(r) should actually be zero until r = 1.)

For the bigram model, there are low values at r = 1 and r = 2, with a distinct peak at r = 3. This could result from the high probability of prepositions to select words that in turn have a low probability of selecting prepositions. For example, ”in the” or ”with a” are common, while ”a with” or ”a in” aren’t grammatical. The peak could result from nested prepositional phrases, such as ”in the bag in the box in the kitchen,” where spacings of three occur for the prepositions and articles involved.

5 Pair Correlation by Part of Speech

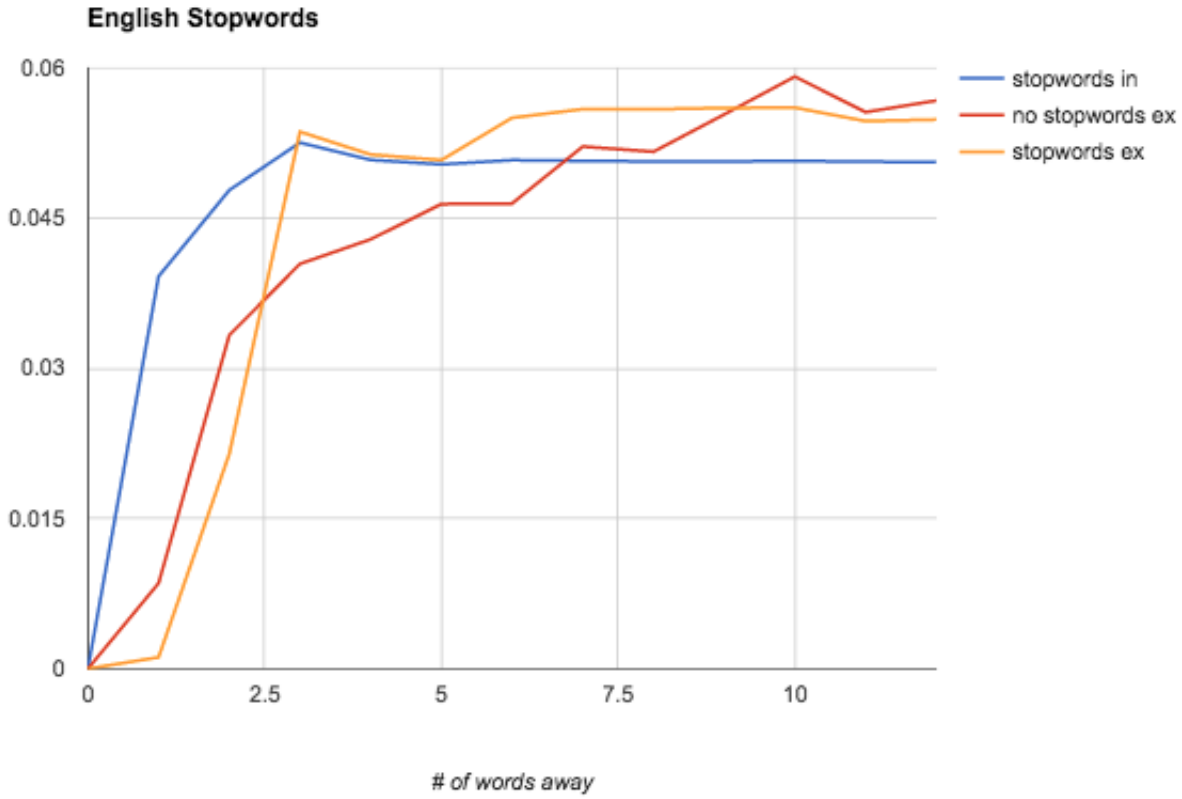
I also looked at the pair correlations of some of the most common words, to see if there were any evident similarities. Here are the 5 most common words in the King James Bible:



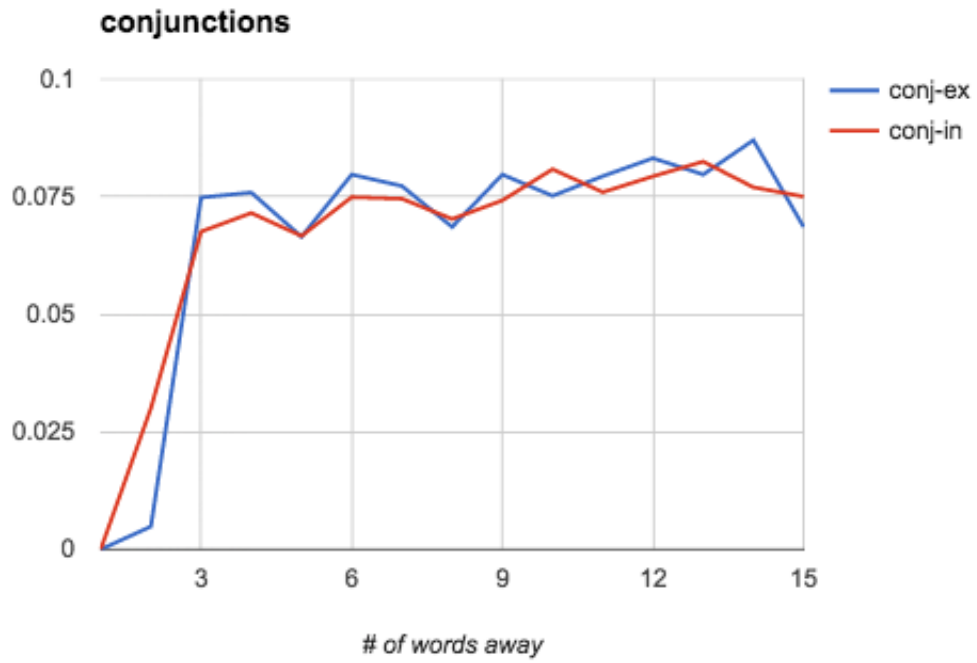
This data actually comes from late fall 2014, when I had just figured out the single word pair correlation. The data has been normalized to let the area = 1 however, to be consistent with the rest of this report. Notice how the only word with a non-zero value for g(1) is ’that;’ this makes sense, given constructions like this: ”I don’t like that that kid keeps walking on my lawn.”

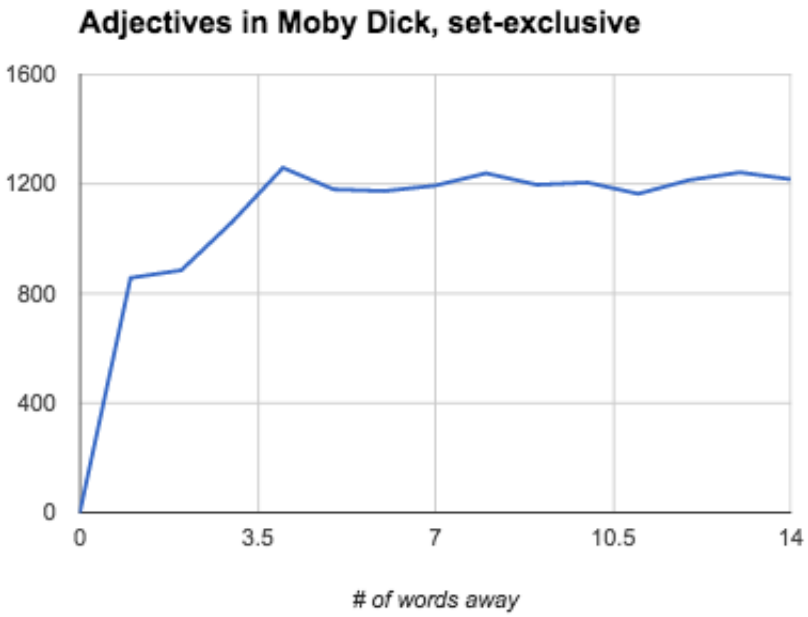
Additionally, I took word-set pair correlations for some parts of speech with typically low type/token ratios, as well as for all stop words. 'In' and 'ex' refer to the set-inclusive/set-exclusive word-set pair correlations. It's worth pointing out that that the set-exclusive stop words pair correlation for the Brown Corpus (below) looks nearly identical to Brown's all-word pair corr. This reflects the fact that the most common words in a corpus have the greatest impact on its all-word pair correlation. 'no stop words' refers to the set of all non-stopwords.

Stop words, the Brown Corpus:

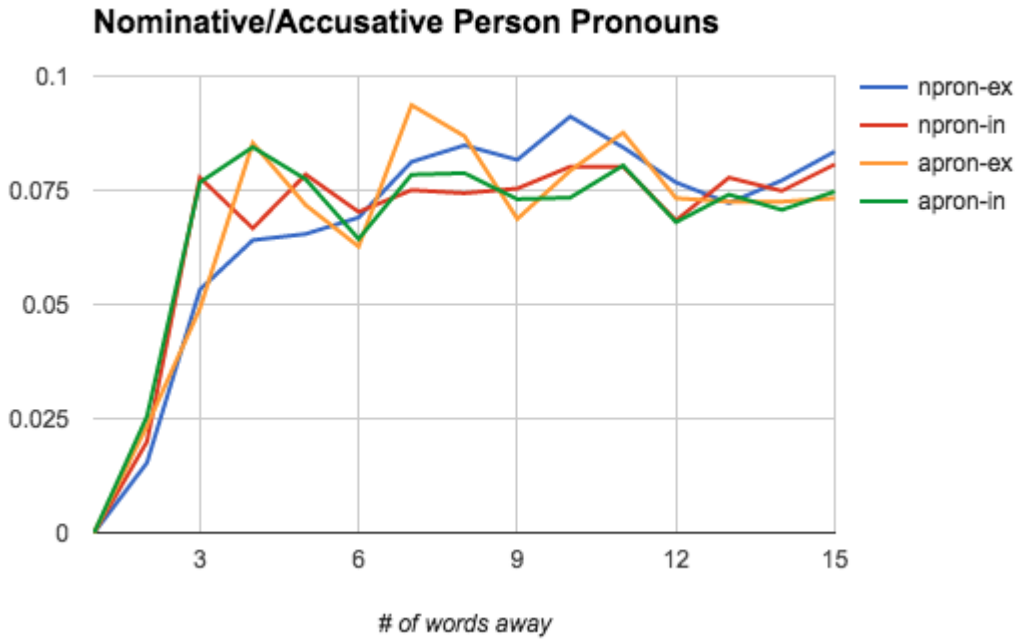


For the rest of these graphs, I used Moby Dick as a corpus. (Note: for graphs with only one curve, the data has not been normalized.)

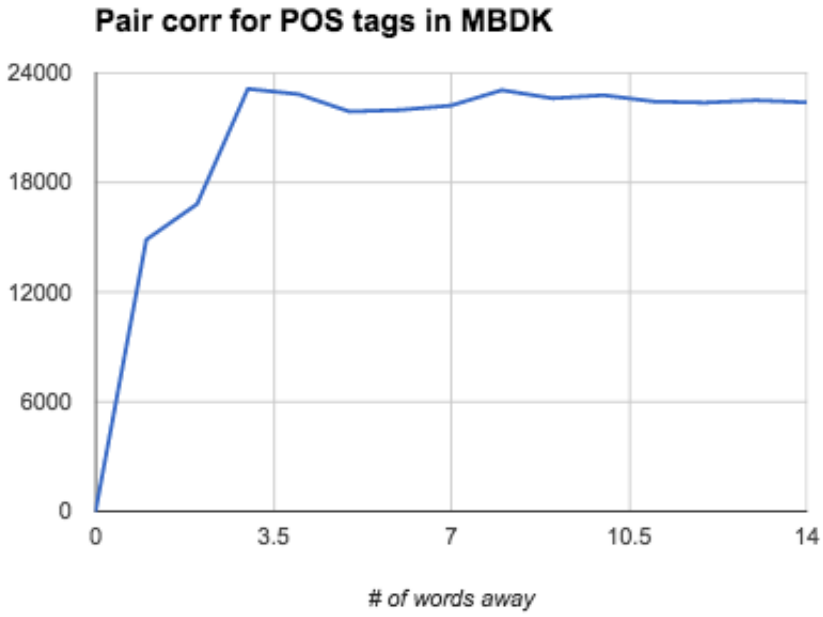

conjunctions
0.1
0.075
0.05
0.025
0
3
6
9
12
15
conj-ex
conj-in
# of words away


Adjectives in Moby Dick, set-exclusive
1600
1200
800
400
0
0
3.5
7
10.5
14
# of words away

Personal Pronouns:



Here as well is the all word pair correlation for POS tags in Moby Dick. That is, the pair correlation for Moby Dick if you were to replace all words with their Penn Treebank style tag. To tag Moby Dick, I used the nltk built-in pos-tagger.



While these graphs are interesting, not much can be said from inspection alone; the accusative pronouns and conjunctions appear mildly periodic between the first and third peaks, but that's about it. There are a few techniques however that could yield more meaningful analysis in the future.

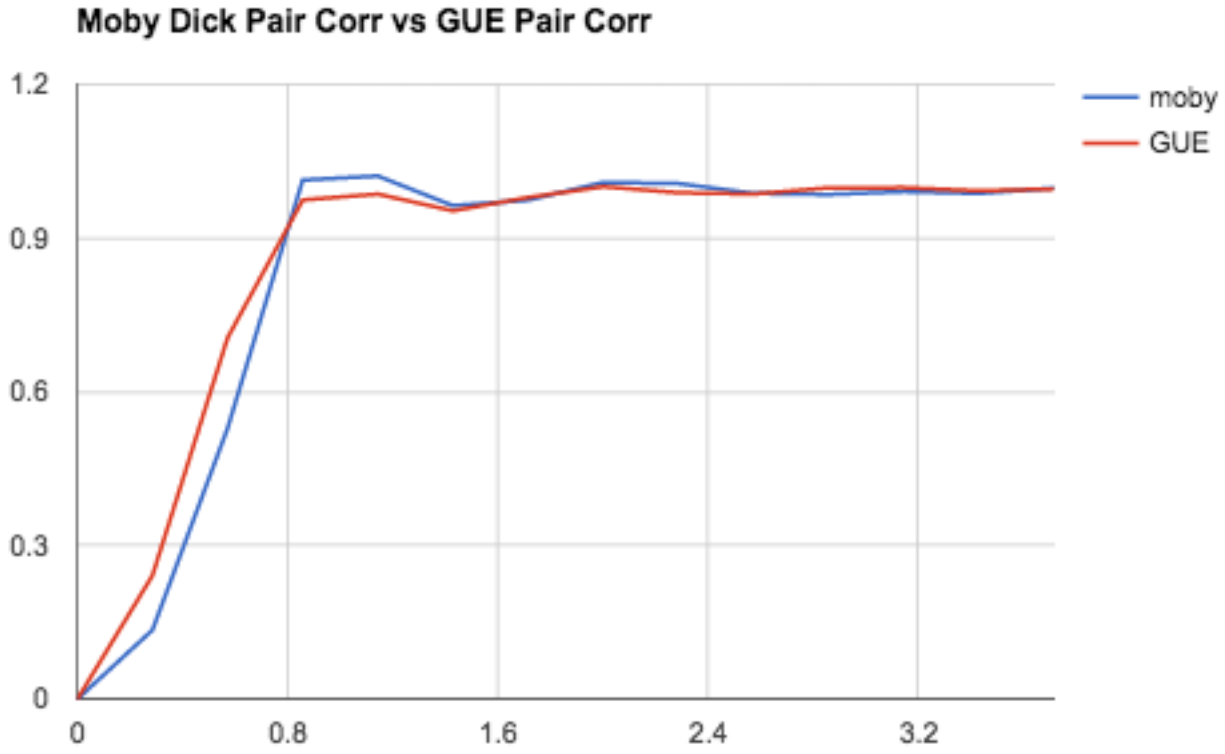
6 Future Plans

To make sense of these pair correlations, there needs to be a way to compare them to each other and group them according to similarity. A good way to do this would be to implement a clustering algorithm that can be applied to pair correlation data. Once the pair correlations are clusterable, a number of interesting investigations would be possible. One could take the paicorrs for a number of high frequency words and group them, to see if the groupings reveal any insight. Like, for instance, part of speech. Likewise, one could take the paircorrs of books, and see if the groupings align with author, genre, time period, etc.

Something else that remains an issue is the normalization of the Pair Corr over word frequency. It would be nice if there were a way to get meaningful pair correlation data for uncommon words, which at the moment hardly have any effect on the all word-pair correlation. If a word like, say, "illusion," never appears within 15 words of itself, then it has no impact on the pair corr. Yet, it still might have a distinct spacing pattern, just on a larger scale.

7 A Note About the Moby Dick Pair Corr

The all-word pair corr for Moby Dick shares a striking similarity to the pair corr associated with the GUE hypothesis. According to the GUE hypothesis, the pair correlation function for GUE eigenvalues and the Riemann zeta function zero's should look like $y = 1 - sin^2(\pi x)/(\pi x)^2$, when the interval being analyzed has been normalized to a density of 1. The normalization guidelines for this comparison don't translate well from math to language. Regardless, to approximate it, I've adjusted the pair corr for Moby Dick so it's first peak is at 1. Here are the two curves overlayed:



It is interesting that Moby Dick fits so well to this curve while all the other texts I've looked at don't. However, I'm not sure how to evaluate the significance of this fit.