KWIC concordances, word senses and geometric sums of stationary random variables

Jason H. Stover

created 1 April 2004
updated Sunday, 17 October 2004

Contents

Chapter 1

Abstract

This paper presents a measure of "distance" between phrases with the same central word and examines the probabilistic behavior of this distance. It is defined between phrases formed with the method of KWIC concordancing from the field of linguistics. The distance measure is formed by first replacing rare words with a single, common artificial word, then matching words between two phrases with a moving window, then using the fraction of non-matched words in a geometric series of powers of some $\theta \in (0, 1)$. The data suggest the continuity of the limiting distribution of the series for some θ, window lengths and replacement rates. A generalized version of a theorem by Garsia [2] states conditions under which this distribution is singular. For the same window length and θ, two particular words with different meanings are shown to have different distributions of their respective distance measures. A difference in distribution functions for different words may therefore imply a difference of meanings between two words.

Chapter 2

Introduction

This paper defines a measure of distance between phrases which may imply differences or similarity in meaning of the central words in the phrases. The phrases are formed by extracting samples of text and forming KWIC concordances, a tool linguists use to compare the uses of words by examining their contexts. The distance measure is formed by creating a geometric sum of the form $Y_n = \sum_{i=-n}^{n} \theta^i X_i$, where $\theta \in (0,1)$ and the X_i are the proportion of non-matched words between the phrases within a moving window. Y_n is not strictly a metric since it may violate the triangle inequality, but it does give a sense of nearness between phrases. After replacing a certain number of rare words in the text, theoretical and empirical results suggest the Y_n will have a nonsingular distribution function which still shows some relation to the original texts. A theorem by Garsia [2] can be extended to give sufficient conditions for the singularity of the distribution of $Y = \lim_{n \to \infty}$. In section 4, the distribution of Y_n is examined for two words, "fruit" and "door," with data taken from 158 novels.

The question of when Y has a density is related to the field of Bernoulli convolutions [6]. In this area, the questions usually relate to the singularity or nonsingularity of the distribution of a geometric series of the form $\sum_{i=0}^{\infty} \theta^i X_i$ where $\theta < 1$ and the X_i are independent identically distributed on $\{-1, 1\}$ or $\{0, 1\}$. Most of the literature in this field describes conditions for which Y has a density and what types of sets support its distribution function. Unlike the problems previously treated in this field, the X_i here are *dependent* and take values on a finite set of order k with probabilities $p_1, ..., p_k$. For independent X_i, Garsia [2] proved a condition under which Y must have a singular distribution function, and his theorem is extended with minor modifications in section 5 to include a case in which the X_i's are stationary and ergodic. Hill and Blanco [3], and Sugiyama and Huzii [9] showed that for independent X_i's there are some values of θ for which Y has a continuous density given by polynomial splines. In section 4, we will see our data match these results though they are dependent.

Let F be the distribution function of Y, and F_n that of Y_n. To maximize the relation of Y_n with the original text, it is desirable to alter as little text as possible while attempting to create a nonsingular F. To achieve a nonsingular distribution function for the data shown here, some rare words must be replaced. This will obviously remove some relevant semantic content from the text. It may be possible to create a variant of Y_n which alters the text little or not at all, and still gives a nonsingular F. This improvement is suggested in the final section. The suggestion stems from the fact that, from the perspective of the linguistic community, the presentation of the data in this paper would be considered naive, failing to account for such features as parts of speech, tenses of verbs, plurality of nouns, and identification of proper nouns. I hope the novelty of the statistical approach compensates for this shortcoming.

Section two defines Y by permuting, replacing and matching words in concordances. Section three examines the behavior of Y for the words “fruit” and “door” in 158 novels. Section four presents a generalization of Garsia's theorem which partly explains the behavior seen in the data.

Chapter 3

Distances between phrases

To compute the distance between two phrases, a linguistic technique called *Key Word in Context* (KWIC) concordancing is used to align phrases with common central words. The distance between two phrases is then found by an algorithm presented below. The apparent continuous distribution function of this distance measure allows probabilistic comparison between phrases, which in turn allows one to assess similarity of word use among such phrases.

In this procedure, the rarest words must be replaced to reveal a non-degenerate probability distribution of the distances. If we do not replace these words, the distances between phrases will be either very large or very small with probability one. Such dichotomous distances, which correspond to phrases that either match exactly or almost nowhere, cannot tell us about the range of similarity between phrases, so replacement of rare words is necessary. There is a danger of removing too many words, and if this happens, the probability distribution again becomes singular, with most of its mass around 0. We will see that we can replace sufficiently many rare words to give a continuous distribution without replacing enough words to remove all content of the original phrase.

Before defining the distance between phrases, we shall see an example of the KWIC concordances on which Y is based. In this method, phrases with identical central words are aligned to compare the uses of the central word. The following example shows concordances of for the word “fruit,” taken from the novels *A Christmas Carol*, *A Portrait of the Artist as a Young Man*, *A Study in Scarlet*, *American Notes* and *An Old-Fashioned Girl*:

...sausages, oysters, pies, puddings,	fruit	and punch all vanished...
...as easily as a	fruit	is divested of its...
...to eat of the	fruit	of the forbidden tree....
...ate of the forbidden	fruit	they would become as...
...or rust stains or	fruit	stains or what are...
...time corrupts the whole	fruit	Will you come with...
...savoury cold meats, and	fruit,	and wine, we started...
...islands where every known	fruit	vegetable and flower is...
...the green and purple	fruit	lay all about us...
...we never saw the	fruit	that Nelly didn't look...

The concordances above reveal much about the meaning of the word "fruit." We can see this word is surrounded by words related to food, eating, or plants, staining and bright colors. From this, one who did not know what "fruit" means could surmise that a fruit is a food produced by a plant. It may be green or purple, and may stain. One might induce from the middle phrase that the word can be used to describe metaphorically something desirable and forbidden. With knowledge of the surrounding words, one could infer a lot from examining these contexts.

Much linguistic literature suggests that humans interpret meanings of words by the contexts in which they are used ([4], [8], [5]).

If context determines the meaning of a word, then a measurement of distance between contexts of that word should have certain properties that relate to its meaning. Among these properties is the probabilistic behavior of the distance, which should give an idea of how far apart one can expect concordances to be. Moreover, if two different words typically are surrounded by different patterns of contextual words, the statistical behavior of the distances between their respective concordances should be different. Any soundly-defined measure of distance will be 0 between identical copies of a phrase, and will increase with a rise in the proportion of non-matching words between the two phrases. In addition, a measure of distance between two phrases should place more weight around the central word, since words closer to the central word are more likely to relate to its meaning.

To state distinctly whether we are referring to a lexicon or a corpus (i.e., a collection of text), define *token* to be particular occurrence of a word in the corpus. *Word* hereafter refers to an element of the lexicon. For example, the phrase "the Sun and the Moon" contains four words but five tokens. We can think of a word as a possible value in the sample space and a token as an observed value.

The distance between two phrases is measured as follows: First, denote a phrase by

$$W^*_{1,-n}, W^*_{1,(-n+1)}, ..., W^*_{1,-1}, W^*_{1,0}, W^*_{1,1}, ..., W^*_{1,(n-1)}, W^*_{1,n}.$$

First replace all tokens representing "rare" words with a common pseudo-token to give a new sequence of words

$$W_{1,-n}, W_{1,(-n+1)}, ..., W_{1,-1}, W_{1,0}, W_{1,1}, ..., W_{1,(n-1)}, W_{1,n}.$$

Denote a second phrase, after replacing these same rare words, by

$$W_{2,-n}, W_{2,(-n+1)}, ..., W_{2,-1}, W_{2,0}, W_{2,1}, ..., W_{2,(n-1)}, W_{2,n},$$

where both phrases are chosen so that $W_{1,0} = W_{2,0}$, i.e., the middle words match. In our example above,

$$W_{1,-n}, W_{1,(-n+1)}, ..., W_{1,-1}, W_{1,0}, W_{1,1}, ..., W_{1,(n-1)}, W_{1,n}$$

might be "sausages, oysters, pies, puddings, fruit and punch all vanished," in which case $n = 3$. Our second phrase could be any other phrase from the example. For the first phrase, define the set

$$\mathcal{S}^L_{1,-n} = \{W_{1,-n}, W_{1,(-n+1)}, ..., W_{1,(-n+L-1)}\}.$$

Let $S^L_{1,-n}$ be the set of distinct elements of $\mathcal{S}^L_{1,-n}$ (i.e., with repeated values removed). Define $\mathcal{S}^L_{2,-n}$ and $S^L_{2,-n}$ for the second phrase similarly. Define X_{-n} to be the fraction of non-matching elements from $\mathcal{S}^L_{1,-n}$ and $\mathcal{S}^L_{2,-n}$:

$$X_{-n} = \frac{|S^L_{1,-n} \bigtriangledown S^L_{2,-n}|}{2L}. \tag{3.1}$$

Then shift the window forward one word, defining

$$\mathcal{S}^L_{1,-n+1} = \{W_{1,-n+1}, W_{1,(-n+2)}, ..., W_{1,(-n+L)}\}.$$

Again, define $S^L_{1,-n+1}$, $\mathcal{S}^L_{2,-n+1}$ and $S^L_{2,-n+1}$ as for the previous window, and define

$$X_{-n+1} = \frac{|S^L_{1,-n+1} \bigtriangledown S^L_{2,-n+1}|}{2L}.$$

Continue this process until we have a sequence

$$X_{-n}, X_{-n+1}, ..., X_0, X_1, ..., X_{n-L}$$

of $2n - L + 1$ random variables that record the proportion of unmatched tokens in windows of length L. For convenience, choose L to be odd and re-label this sequence

$$X_{-n+(L-1)/2}, ..., X_0, ..., X_{n-(L-1)/2}.$$

Then define the distance between the two phrases to be

$$Y_n = \sum_{i=-n+(L-1)/2}^{n+(L-1)/2} \theta^{|i|} X_i \tag{3.2}$$

where $0 < \theta < 1$. Notice that Y_n is not a metric since it may violate the triangle inequality. Hereafter, assume the X_i's form a stationary ergodic sequence. Since $\theta \in (0, 1)$ and the X_i's take a only finite number of values between 0 and 1, $\lim_{n\to\infty} Y_n = Y$ almost surely for some random variable Y. Let F_n be the distribution function of Y_n, and F be the distribution function for Y. We will show empirical evidence that for some choices of rare word replacement and θ, F is nonsingular. We will also present a generalized version of a theorem by Garsia [2] that gives a sufficient condition for the singularity of F.

If no words were be replaced Y would be singular since so many of the tokens represent rare words. Because of the window used, common words will often match. Most of the non-matches are caused by the appearance of infrequently-used words. The large number of low-probability words in a lexicon is known as Zipf's Law [4], which states that the probability of an appearance of a word is proportional to the reciprocal of its rank, i.e., if the word w_i is the r_i^{th} most commonly used word, $\Pr(\text{see } w_i) \propto 1/r_i$. While Zipf's Law does not perfectly describe the distributions of words [4], it is a close enough approximation to tell us that there is a large proportion of the lexicon whose individual members are used rarely, but in sum these words constitute a large proportion of tokens, thereby causing many non-matching tokens, even among phrases with similar meaning. Dropping these words forces more matches, reducing the distance. At the same time, we want to retain any common words, especially context-dependent ones, since they are likely to cause a match in semantically similar phrases. There is no rule presented here for replacing rare words. The words chosen for replacement were chosen to give an apparent density function for Y.

Chapter 4

Example: Fruit vs. Door

This section explores Y_n via an example using data from 158 novels whose copyrights have expired. All were either written in English or are English translations. A list of all the novels used appears in the appendix. They were obtained from www.bibliomania.org.

The data were created from the corpora as follows. All punctuation was removed from the novels' text. Plural forms of nouns were treated as distinct words, as were different tenses of the same verb. The possessive modification *'s* was treated as a distinct word. (For a treatment of the question of what is or is not a word, see [4]). All phrases containing "fruit" and "door" were extracted, and concordances were formed with either "fruit" or "door" as the central token, surrounded by the leading and trailing eleven tokens, i.e. $n = 11$ in (3.2). The window length L was chosen to be 5. All phrases were $2n + 1 = 23$ tokens long.

The following concordances illustrate how the definition of distance in (3.2) is used for these novels. The first excerpts show two phrases, the first from *A Double-Barrel Detective Story* by Mark Twain and the second from *Notre-Dame de Paris* by Victor Hugo.

> ...with a gripsack handy, with a change in it and my *door* ajar. For I suspected that the bird would take wing now...

> ... the wild boar in his lair, pressed tumultuously round the great *door*, disfigured now and injured by the great battering ram. But...

After replacing the rare words among all the novels, the two phrases appear this way:

> ...with a -1 -1 with a change in it and my *door* -1. For I -1 that the -1 would take -1 now....
>
> ...the -1 -1 in his -1 -1 -1 round the great *door*, -1 now and -1 by the great -1 -1 But...

The distance between these two phrases is about 5.08, close to the sample mean for the "door" phrases.

The distance between the following phrases was less than 1.8, closer to the minimum for the "door" phrases. The phrases were taken from *Dr. Jekyll and Mr. Hyde* by Robert Louis Stevenson and *Mr. Sponges' Sporting Tour* by Surtees.

> ...inseparable friends. On the 12th, and again on the 14th, the *door* was shut against the lawyer. 'The doctor was confined to the...
>
> ...hanging out of the windows, flirting and chatting and ogling, the *door* was shut, the blinds were down, the shutters closed, and...

Most of these words are replaced with -1, which causes more matches and a corresponding smaller distance. Also notice the common phrase "the door was shut" in both excerpts. This explains why the "fruit" concordances have a smaller mean: There are more rare words and fewer repeated phrases surrounding "fruit" than surrounding "door," causing more matches after the rare words have been dropped.

There were 655 phrases with "fruit" as the central token and 12647 phrases with "door" as the central token. Since computing all possible pairwise distances among the "door" concordances would result in 159 million values, the data were sampled to give 200326 distances computed for the phrases centered on "door." All $\binom{655}{2} = 214185$ distances centered on "fruit" were computed. The coefficient θ was 0.8. The "rare" words were defined to be those least-used words which accounted for a fraction of 0.25 of all tokens from the 159 novels. These 0.25 of the tokens were accounted for by about 0.97 of the 16160 distinct words represented in the corpus.

In addition, to check the distribution of the distances for phrases centered on different words, the central tokens were replaced with "fruitdoor" for a randomly selected 16460 phrases from both "fruit" and "door" concordances.

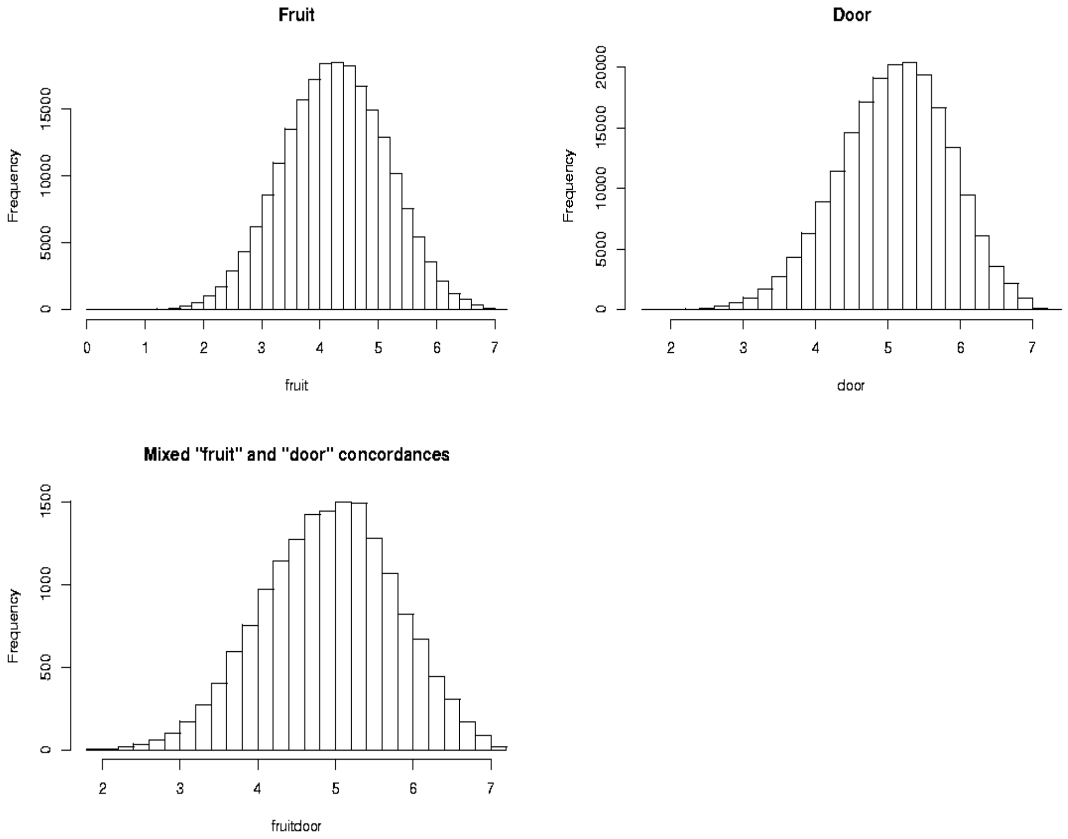

Figure 4.1: Histograms for Y_{11}, $\theta = 0.8$, $L = 5$, $n = 11$.

Distances between these phrases were also computed. The histograms for the three types of distances are shown in Figure 4.1.

The replacement of the rare words may have the following interpretation. If no words are replaced, F will be singular, placing all its mass at high values, since few tokens will match. If most words are replaced, the distribution will again be singular, this time with mass close to 0, since most tokens will match. There is a proportion of words which, if replaced, will give a nonsingular F. There are some words authors must use frequently (e.g., "a,""an," "of," etc.). Other, rare words are more topic dependent ("taste," "peel," etc.). Some words may depend weakly on the topic and appear frequently (e.g., "through" as in *through the door*). Semantic information for humans is contained both in the rare words, most of which relate to the phrase's topic by virtue of their presence in the phrase, and the "glue" among those words: The common words in the phrase tell us about the relation of the central word to other concepts. Prepositions tells us about placement with respect to other objects ("*through* the door" or "piece *of* fruit"), specifiers tell us whether the central word is a specific instance of an object (*the* door) or an unspecified member of a class of objects (*a* door). Removal of the rare words is therefore removal of important content words showing the topic at hand, leaving words that, when matching tokens from other phrases, indicate similarity in the relationships around the central token.

The histogram for the "fruit" concordances, surprisingly, looks normally distributed. It does share several features with the normal distribution, including approximate symmetry about the quartiles and rate of decay in the tails. Kolmogorov's D was 0.0085 for a test with a null hypothesis of normally distributed data. This value, though small, is large enough to reject the hypothesis for such a large sample size. The deviation from normality is due to a slight right-skew in the data. Nonetheless, the closeness to normality raises the question: Are there values of θ, L, and the number of words replaced that will give a normally distributed Y? The answer is not known, but likely to be negative: In [3] and [9], it was shown that when the X_i's in the sum are independent, the density of Y is a polynomial spline. In [3], this spline density resembles a normal density for some values of θ. We have no rigorous result showing that our dependent X_i's give a spline density for Y, so the statement that the data are never normally distributed is a conjecture.

The histograms for "fruit" and "door" have obvious differences in center and shape. The histogram for "door" has a larger center and notable right skew, whereas the histogram for "fruit" is more symmetric with a smaller center. This is seen when checking any relevant statistics: the sample mean for the "fruit" concordances is 4.26 while that for "door" is 5.08 (the large sample size precludes the need or relevance of mentioning that these differences are significant). The third central moment for "fruit" is $\frac{1}{214185}\sum_{i=1}^{214185}(Y_{11,i} - \bar{Y}_n)^3 = -0.03$ while that for "door" is -0.11.

The value of θ has an important effect on Y_n. If chosen too large, i.e. close to 1, then all words in a phrase will be weighted approximately equally, and our sequence will not depend much more on the central tokens than the outlying ones. On the other hand, if θ is chosen too close to 0, then F_n will be singular, in which case we will not see a range of values with different probabilities.

In Section 5, the following result relating θ to the the match probabilities of the tokens will be shown: For a certain class of stationary X_i's which take a finite number of values with probabilities $p_1, ..., p_m$, $Y = \lim_{n\to\infty} Y_n$ has a singular distribution function if the X_i's have entropy less than $\log(1/\theta)$. This theorem was proved by Garsia [2] for independent X_i's, and can be generalized with a slight modification.

Let $F_{n,\theta}(x)$ be the distribution function of Y_n for a specified θ. As the value of θ decreases, $F_{n,\theta}$ will move from a nonsingular distribution placing positive probability at high values of Y_n to a singular distribution. This fact is partly explained by Garsia's theorem, since for a large θ, $\log(1/\theta)$ is smaller than the entropy of the X_i. On the other hand, if θ is small, the entropy of the X_i will fall below the bound given by Garsia's theorem, and $F_\theta = \lim_{n\to\infty} F_{n,\theta}$ will be singular. We can see $\log(1/\theta)$ overcome the entropy of X_i to give a singular distribution in the histograms shown in Figure 4.2. So interplay between θ and the X_i gives us two competing features of the data: To satisfy our notion that the tokens close to the center of the phrase are more important, θ should be small, but to give Y a nonsingular distribution, θ should be large. The value of θ for the histograms in Figure 4.1 was chosen as a compromise between these two features.

Garsia's theorem also tells us the histogram of Y_n should become concentrated around a few values when the X_i's in the sum are concentrated on only a few of their possible values. This happens whenever the chance of a match between phrases is either too high or too small. So replacing rare words is necessary to increase the entropy of the X_i's, thereby allowing Y to have a nonsingular distribution function. There is no known converse to Garsia's theorem for dependent X_i, i.e., we cannot say with certainty that higher entropy among the X_i's will give a nonsingular distribution function

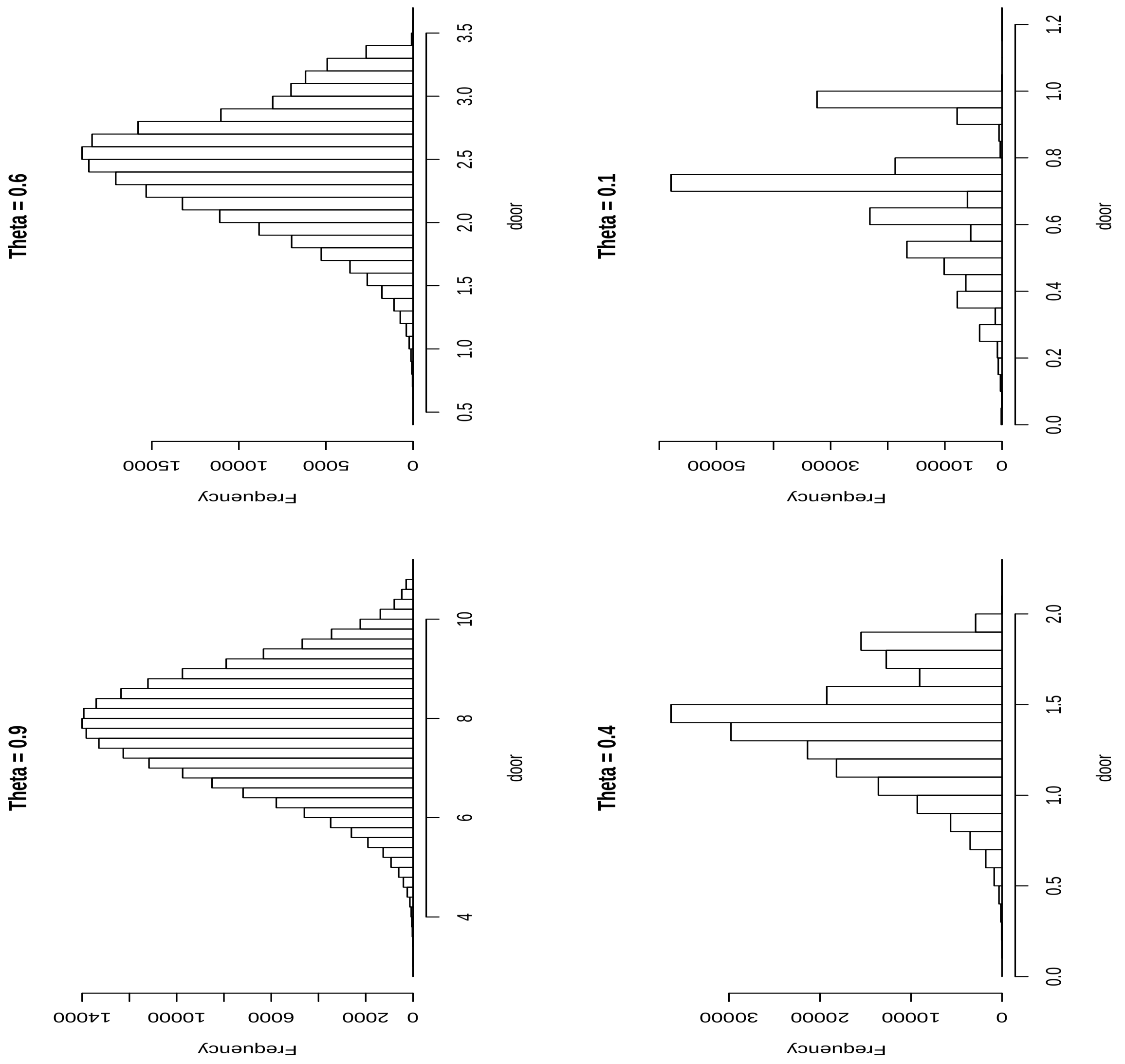

Figure 4.2: Histograms for Y_{11} for different values of θ. As θ decreases, the histograms appear more like those from a singular distribution function.

Transition probability matrix

	Fruit									
	0	1	2	3	4	5	6	7	8	10
0	0.265	0.388	0.258	0.075	0.010	0.004	0.000	0.000	0.000	0.000
1	0.063	0.260	0.371	0.227	0.063	0.012	0.003	0.000	0.000	0.000
2	0.013	0.107	0.275	0.338	0.202	0.049	0.013	0.001	0.002	0.000
3	0.002	0.029	0.141	0.291	0.312	0.166	0.045	0.003	0.010	0.000
4	0.000	0.005	0.051	0.172	0.302	0.284	0.144	0.017	0.021	0.003
5	0.000	0.001	0.010	0.072	0.204	0.320	0.261	0.078	0.045	0.010
6	0.000	0.000	0.003	0.021	0.104	0.239	0.341	0.169	0.094	0.029
7	0.000	0.000	0.000	0.002	0.017	0.131	0.290	0.330	0.140	0.090
8	0.000	0.000	0.001	0.010	0.032	0.083	0.194	0.135	0.409	0.136
10	0.000	0.000	0.000	0.000	0.006	0.025	0.079	0.135	0.206	0.548

Table 4.1: The estimated one-step transition probability matrix for the X_i's in the "fruit" distances. Notice that for $L = 5$, X_i cannot be 9.

Stationary distribution

Fruit									
0	1	2	3	4	5	6	7	8	10
0.002	0.011	0.037	0.086	0.149	0.199	0.212	0.0.118	0.110	0.076
Door									
0.00002	0.00050	0.00364	0.0297	0.07968	0.15592	0.44081	0.27492	0.60123	0.58096

Table 4.2: The estimated stationary distribution for the transition probability matrix of the X_i's. The entropies for the two distribution are 2.020 and 1.934 respectively, both greater than $\log(1/\theta) = 0.223$, the bound below which Y has a singular distribution function according to Corollary 5.0.2

for Y. We are speculating on the basis of empirical evidence.

To compare the entropy of the X_i's to $\log(1/\theta)$, the one-step transition probability matrix for the X_i's was computed for the "fruit" concordances and is shown in Table 4.1. There is criticism in the linguistic community of the appropriateness of Markov chains as models for human language [4], but they are regarded as a useful model in some cases. Viewing our sequence as a Markov chain, Table 4.2 shows the estimated stationary distributions for the X_i's in the "fruit" and "door" concordances. The estimated entropies of the X_i's were 2.020 and 1.93 for "fruit" and "door" respectively, both well above the lower bound of $\log(1/\theta) = 0.223$ given by Garsia's theorem.

There is a concrete relationship between he higher entropy of the X_i's in for the "fruit" concordances, the greater scatter in its histogram, and our scheme of replacing rare words. Table 4.3 shows the breakdown of re-

Proportion of matches by replacement

	"Fruit"		"Door"	
	Match	Non-match	Match	Non-match
Replaced	0.319	0.031	0.176	0.056
Not replaced	0.111	0.538	0.139	0.629

Table 4.3: The proportion of non-matching words was lower among "fruit" concordances than "door" concordances. This resulted in higher average distances between "door" concordances, as shown in Figure 4.1

placements by matching. The higher entropy for the terms of the "fruit" concordances results from there being more tokens replaced in these phrases than in the "door" phrases. This causes more matches among the X_i for the "fruit" concordances, which raises their entropy and in turn causes a wider spread in the distribution of the Y_n. Also notice that more matches in the "fruit" phrases gives them a lower mean distance than that of the "door" phrases.

The fact that there are more replaced values in the "fruit" concordances than in the "door" concordances tells us the word "fruit" is often surrounded by rare words more often than is the word "door." So the difference in the distribution functions is caused at least partly by a richer variety of context words for "fruit" than for "door."

Do the resulting probability distributions relate to semantics in the phrases? Or have the semantic features of the phrases been erased during replacement of rare words, leaving a set of random variables that have little to do with the language? The answer to this question tells us whether the method allows us to see semantic similarity of words, or only to study an interesting, but semantically irrelevant, aspect of the randomness of language.

Though these are philosophical questions, a view of the match types suggests both are partly true. Certainly if too many words are replaced, almost all matches will occur because of the replaced words, erasing the effect of semantic similarity. But in the example, $0.111/(.111+.319) = .258$ of the matches for the "fruit" concordances were made with non-replaced words, so much of the similarity between phrases is accounted for by matching among the original tokens. Though replacing rare words removes some or most original meaning, it does not remove all of it.

The third histogram in Figure 4.1 shows another interesting feature of the data. This histogram was created by replacing the words "fruit" and "door" with the artificial word “fruitdoor,” and measuring the distances between the two types of phrases. Since the middle words of any pair now match because of the bogus word “fruitdoor,” we can see how the distance function behaves when comparing the contexts of two different words. This method is used to test word-sense disambiguation methods [4]. One might hope that, for distances measured between phrases with different central tokens, the distribution of the Y_n would differ from a distribution of Y_n between phrases with identical central tokens. If they were different, we could detect this difference by examining the distribution function of the distances between the two types of phrases. One manifestation of this difference in distributions we might hope for is a higher mean of the distribution containing phrases of mixed type. The third histogram shows a distribution that is different from the other two, but its mean is not higher than both. Let $\bar{Y_f}$ be the sample mean for the “fruit” distances and $\bar{Y_d}$ be the sample mean for the “door” distances. The sample mean for the “fruitdoor” distances was 4.91, slightly larger than $.5(\bar{Y_f} + \bar{Y_d})$. Nevertheless, this distribution does differ from both the “fruit” and “door” distributions, which could be caused by a semantic difference between the two words via their different contextual words. There is no doubt that much of the difference in distribution is caused by more replacements in the “fruit” phrases. The question of whether this indicates different meanings between the two words depends what we mean by “mean,” and that question is still debated.

Chapter 5

Garsia's Theorem

This section presents a theorem stating sufficient conditions under which the distribution of $Y = \lim_{n\to\infty} Y_n$ is singular. The theorem was proved for independent random variables by Garsia [2]. Let $Z_n = Y - Y_n$, and let $F_n(x)$ be the distribution function of Y_n. Lemma 5.0.1, coupled with the lack of assumed independence of the X_i's in Lemma 5.0.3, allows us to generalize Garsia's main theorem to include a class of X_i's which are stationary and ergodic. Unlike the definition of Y_n in Section 3, in this section assume Y_n is a one-sided sequence with initial value Y_0 chosen from a stationary distribution.

Garsia's theorem gives a sufficient condition for the singularity of $F(x)$. Unfortunately, there is no known necessary condition for stationary X_i. Research toward this result lies in the field of Bernoulli convolutions [6]. There are some known circumstances in which Y has a density when the X_i's are independent [9],[3],[7], but our X_i's are dependent. Despite the lack of a necessary condition, knowing when the data cannot have a density function is instructive when choosing the number of words to replace to produce the Y_n.

For completeness, the proofs of all of Garsia's original theorems which rely on Lemma 5.0.1 are reproduced here.

Lemma 5.0.1 *Let $Z_n = Y - Y_n$. For any $x, y > 0$,*

$$F_n(x+y) - F_n(x) \le 2\left[F(x+2y) - F(x-y)\right].$$

Proof: First notice

$$\Pr\{x < Y_n < x+y, |Z_n| \le y\} \le \Pr\{x-y < Y \le x+2y\} \tag{5.1}$$

Since Y_n takes values on a finite set,

$$\begin{aligned} \Pr\{x < Y_n &\leq x+y, |Z_n| \leq y\} = \\ &\sum_{x_i \in (x,x+y]} \Pr(Y_n = x_i)\Pr(|Z_n| \leq y | Y_n = x_i) \end{aligned} \tag{5.2}$$

where the x_i's in the sum denote possible values of Y_n in the interval $(x, x+y]$. If n is large enough, because $\theta < 1$ and Y_n is ergodic, takes values on a finite set and $|Z_n| \to 0$, we must have $s = \inf_i \{\Pr(|Z_n| \leq y | Y_n = x_i)\} > 0$, so using 5.2

$$\begin{aligned} \Pr\{x < Y_n \leq x+y, |Z_n| &\leq y\} \\ &= \sum_{x_i \in (x,x+y]} \Pr(Y_n = x_i)\Pr(|Z_n| \leq y | Y_n = x_i) \\ &\geq s[F_n(x+y) - F_n(x)] \end{aligned} \tag{5.3}$$

Eventually, $\mathrm{E}(Z_n^2)/y^2 < 1/2$, so using the Chebycheff inequality and the definition of s, we see

$$\Pr(|Z_n| > y) \leq \frac{\mathrm{E}(Z_n^2)}{y^2} < 1/2 \Rightarrow \frac{1}{s} < 2.$$

Combining this, 5.1 and 5.3, we have

$$F_n(x+y) - F_n(x) \leq 2\left[F(x+2y) - F(x-y)\right].\Box$$

The following theorems are modified versions of those proved by Garsia. Their proofs have been modified to account for our stationary X_i's.

Lemma 5.0.2 *(Garsia [1]) If y_n is any sequence decreasing to 0, then $F(x)$ has a singularity only if the following condition is satisfied:*

Condition S. There exists a $\gamma > 0$ such that for any integer n_0 and $\epsilon > 0$, there is a set of integers S such that for some $n > n_0$,

$$\begin{aligned} \sum_{k \in S}[F_n(ky_n + y_n) - F_n(ky_n)] &> \gamma \\ |S| &\leq \epsilon/y_n \end{aligned} \tag{5.4, 5.5}$$

On the other hand, if $y_n \to 0$ slowly enough that

$$\begin{aligned} \liminf_{n\to\infty} y_n^2/\mathrm{E}(Z_n^2) &> 0 \qquad (5.6)\\ \liminf_{n\to\infty}\inf_{x_i}\{\Pr(|Z_n| < y_n|Y_n = x_i)\} &> 0 \qquad (5.7)\end{aligned}$$

then condition S sufficient to guarantee the singularity of $F(x)$.

Proof: If $F(x)$ is singular, there exists a $\gamma > 0$ such that for any ϵ there is a disjoint finite union of open intervals $I = \cup_i(a_i, b_i)$ such that $\int_I dF(x) > \gamma$, $\sum_i(b_i - a_i) < \epsilon$. Since $F(x)$ is a uniform limit of $F_n(x)$ (by Scheffe's Theorem), if we define $k^a_{i,n} = \sup\{k : ky_n \le a_i\}$ and $k^b_{i,n} = \inf\{k : ky_n \ge b_i\}$ we have $k^a_{i,n}y_n \to a_i$ and $k^b_{i,n}y_n \to b_i$. Therefore, when n is large enough, the intervals $(k^a_{i,n}y_n, k^b_{i,n}y_n)$ are disjoint and if we let $S = \cup_i\{k : k^a_{i,n} \le k \le k^b_{i,n} - 1\}$, (5.4) follows and $\epsilon > \sum_i(b_i - a_i) > \sum_{i\in S}(ky_n + y_n) - ky_n = |S|y_n$ gives (5.5).

Suppose now condition S is satisfied and that (5.6) is true. Then we may assume there is an integer m such that $m^2y_n^2 \ge 2\mathrm{E}(Z_n^2)$. If (5.7) is true, by Lemma 5.0.1, for any x,

$$F_n(x + my_n) - F_n(x) \le 2[F(x + 2my_n) - F(x - my_n)] \qquad (5.8)$$

Given a set S of integers satisfying condition S, define $S^{-m}, S^{-m+1}, ..., S^{2m}$ by setting $S^j = \{p : k + j = p, k \in S\}$. From (5.4) and (5.8) we have

$$\begin{aligned} \gamma &< \sum_{k\in S}[F_n(ky_n + my_n) - F_n(ky_n)]\\ &\le 2\sum_{k\in S}[F(ky_n + 2my_n) - F(ky_n - my_n)]\\ &= 2\sum_{k\in S}[F((k + 2m - 1)y_n + y_n) - F((k + 2m - 1)y_n)\\ &\quad +F((k + 2m - 2)y_n + y_n) - F((k + 2m - 2)y_n)\\ &\quad ...\\ &\quad +F((k - m)y_n + y_n) - F((k - m)y_n)]\\ &\le 2\sum_{j=-m}^{2m}\sum_{i\in S^j}[F(iy_n + y_n) - F(iy_n)]\end{aligned}$$

This implies that for at least one j we have $\sum_{k\in S^j}[F(k)-F(ky_n)>\gamma/6m$. Because of (5.5) we have

$$\sum_{k\in S^j} y_n < \epsilon$$

which implies $F(x)$ is not absolutely continuous with respect to Lebesgue measure. □

Corollary 5.0.1 *(Garsia [2]) If y_n tends to 0 slowly enough that*

$$\begin{aligned} \liminf_{n\to\infty} y_n^2/\mathrm{E}(Z_n^2) &> 0 &(5.9)\\ \liminf_{n\to\infty}\inf_{x_i}\{\Pr(|Z_n|<y_n|Y_n=x_i)\} &> 0 &(5.10)\end{aligned}$$

then a necessary and sufficient condition for the singularity of $F(x)$ is that for an M so large that the quantity

$$\lambda_n = \sum_{|ky_n|\le M}[F_n(ky_n+y_n)-F_n(ky_n)] \qquad (5.11)$$

is bounded away from 0, the probability distributions

$$S_n = \left\{\frac{F_n(ky_n+y_n)-F_n(ky_n)}{\lambda_n} : |ky_n|\le M\right\} \qquad (5.12)$$

form a singular sequence.

Proof: If (5.11) implies (5.12), then both statements of condition S in Lemma 5.0.2 are satisfied by choosing S to be a small subset of $\{k : |ky_n| \le M\}$ when M is large enough. Lemma 5.0.2 then implies the singularity of $F(x)$. Conversely, if $F(x)$ is singular, then condition S in Lemma 5.0.2 is satisfied, and $\{k : |ky_n| \le M\}$ will contain the set S eventually if $M > \sup_i k_{i,n}^b$, where $k_{i,n}^b$ is defined in the proof of Lemma 5.0.2. This gives a non-empty subset of singular elements of S_n whenever $n > M$ and (5.11) is true. □

The next lemma does not rest on the assumption of independence, hence requires no modification to apply to our situation. See [1] for a proof.

Lemma 5.0.3 *(Garsia [1]) Let $\mathcal{R}_n$ be the set of possible values of Y_n. $F(x)$ is singular unless*

$$\lim_{n\to\infty} -\frac{\sum_{i\in\mathcal{R}_n} p_n(i)\log p_n(i)}{\log|\mathcal{R}_n|} = 1$$

Theorem 5.0.1 *(Garsia [2]) If $\{y_n\}$ is a sequence of positive numbers tending to 0 such that*

$$\begin{aligned}\liminf_{n\to\infty} y_n^2/\mathrm{E}(Z_n^2) &> 0 \qquad (5.13)\\ \liminf_{n\to\infty}\inf_{x_i}\{\Pr(|Z_n| < y_n|Y_n = x_i)\} &> 0. \qquad (5.14)\end{aligned}$$

If

$$\liminf_{n\to\infty}\frac{H_1+H_2+...+H_n}{\log(1/y_n)} < 1 \qquad (5.15)$$

Then $F(x)$ is singular.

Proof: Assume (5.13), (5.14) and (5.15) are true. Let Ω be the measure space where Y_n and Y are defined as the product of the measure spaces $\Omega_1, \Omega_2, ..., \Omega_n$ where the X_i's are defined. The equivalence relation

$$\omega'' \sim \omega' \text{ if and only if } X_i(\omega') = X_i(\omega''), i = 1, ..., n$$

generates a partition which is finer than the partition generated by the relation

$$\omega'' \sim \omega' \text{ if and only if } Y_n(\omega') = Y_n(\omega''), i = 1, ..., n \qquad (5.16)$$

Let D_n be the entropy of the distribution of Y_n. By the properties of entropy,

$$D_n \le H_1 + H_2 + H_3 + ... + H_n \qquad (5.17)$$

Suppose Y_n takes values $y_{n,1}, y_{n,2}, ..., y_{n,N(n)}$ with probabilities

$$p_n(1), p_n(2), ..., p_n(N(n)).$$

For a given integer M, partition the indices $1, 2, ..., N(n)$ into two sets S' and S'' as follows: S' is the set of all i such that $y_{n,i} \le M$ and S'' the complement. Let

$$Q'_n = \sum_{i\in S'} p_n(i)$$

$$D'_n = -\sum_{i\in S'}\frac{p_n(i)}{Q'_i}\log\frac{p_n(i)}{Q'_i}$$

$$Q''_n = \sum_{i\in S''} p_n(i)$$

$$D_n' = -\sum_{i \in S''} \frac{p_n(i)}{Q_i''} \log \frac{p_n(i)}{Q_i''}.$$

Then

$$Q_n' D_n' + Q_n'' D_n'' = D_n + Q_n' \log Q_n' + Q_n'' \log Q_n'' \leq D_n \tag{5.18}$$

When M is sufficiently large, since $\mathrm{E}(Y^2) < \infty$, we can guarantee that Q_n' remains arbitrarily close to 1. Since $D_n'' \geq 0$, by (5.17), (5.18) and (5.15), we can choose M large enough that

$$\liminf \frac{D_n'}{\log(1/y_n)} < 1. \tag{5.19}$$

On $\{\omega \in \Omega : Y_n \leq M\}$, define a partition by the equivalence relation $\omega_1 \sim \omega_2$ if and only if $Y_n(\omega_1)$ and $Y_n(\omega_2)$ belong to the same interval $[ky_n + y_n, ky_n)$. Since this partition is coarser than the one induced by (5.16), the entropy E_n' for this partition must satisfy

$$E_n' \leq D_n'. \tag{5.20}$$

Let λ_n is defined as in (5.11), and let

$$E_n = -\sum_{|ky_n| \leq M} \left[\frac{F_n(ky_n + y_n) - F_n(ky_n)}{\lambda_n} \log \left(\frac{F_n(ky_n + y_n) - F_n(ky_n)}{\lambda_n} \right) \right], \tag{5.21}$$

then $E_n' \approx E_n$. Combining this relation with (5.19) gives

$$\liminf \frac{E_n}{\log(1/y_n)} < 1.$$

Lemma 5.0.3 then implies

$$\left\{ \frac{F_n(ky_n + y_n) - F_n(ky_n)}{\lambda_n} \right\}$$

forms a singular sequence, so Corollary 5.0.1 implies the singularity of $F(x)$. □

Corollary 5.0.2 *(Garsia [2]) Assume the X_i are stationary and ergodic, and Y_0 is drawn from a stationary distribution. If the entropy of the X_i's is less than* $\log(1/\theta)$*, $F(x)$ is singular.*

PROOF: Let $y_n = \theta^n$ in Theorem 5.0.1. Since $\mathrm{E}(Z_n^2) < \theta^{2n}/(1-\theta)^2$,

$$\liminf_{n\to\infty} y_n^2/E(Z_n^2) \geq \frac{\theta^{2n}}{\theta^{2n+2}/(1-\theta)^2} > 0$$

so assumption (5.13) of Theorem 5.0.1 is met. The stationarity and ergodicity of the X_i gives $\inf_{x_i} \Pr(|Z_n| < y_n | Y_n = x_i) = \Pr(|Z_0| < 1 | Y_0 = x_i) > \alpha$ for some positive constant α, so assumption (5.14) of Theorem 5.0.1 is met, and $F(x)$ is singular. □

Chapter 6

Conclusion

The distance Y_n presented in this paper provides an interesting view of the statistical behavior of KWIC concordances. The data suggest it can be formed so as to have a continuous limiting distribution. The generalized version of Garsia's theorem gives a sufficient condition for the singularity of this limiting distribution. Y_n provides some insight into the similarity of the uses of words by showing relationships among their context words.

A higher match rate among the unaltered tokens would raise the entropy of the X_i, thereby reducing the proportion of replaced words necessary to give a nonsingular F. A method which allows partial matching could be employed to this end. The match rate for verbs could be increased by allowing partial matching between two tokens if those tokens have the same infinitive and match either tense or conjugation, i.e. "has" and "had" have the same infinitive ("to have"), but different tenses. Also, pronouns could be divided among first, second and third person subjective and objective cases, giving partial matches among words such as "they" and "them" or "her" and "me."

Appendix A

Novels

A Christmas Carol
A Double-Barrel Detective Story
A Portrait of the Artist as a Young Man
A Sentimental Journey through France and Italy
A Study in Scarlet
Alice's Adventure in Wonderland
American Notes
An Old Fashioned Girl
Anna Karenina
Around the World in 80 Days
Barchester Towers
Barnaby Rudge
Billy Budd
Black Beauty
Bleak House
Brave New World
Bruno's Revenge and other Stories
Confessions of an English Opium-Eater
Crime and Punishment
David Copperfield
Dead Souls
Dombey and Son
Dr. Jekyll and Mr. Hyde
Dracula
Dubliners
Eight Cousins
Emma
Erewhon
Far from the Madding Crowd
Frankenstein
Good Wives
Great Expectations
Guide to Fiction
Gulliver's Travels
Hard Times
His Last Bow
Huckleberry Finn
Ivanhoe
Jane Eyre
Joseph Andrews
Jude the Obscure
Kidnapped
Kim
King Solomon's Mines
Lady Chatterleys Lover
Lady Susan
Lavengro
Little Dorrit
Little Women
Lord Jim
Lorna Doone
Madame Bovary
Mansfield Park
Martin Chuzzlewit
Martin Eden
Middle March
Mill on the Floss
Moby Dick
Moll Flanders
Moonfleet
Moonstone
Mr. Midshipman Easy
Mr Sponges Sporting Tour
Nicholas Nickleby
Northanger Abbey
Nostromo
Notre-Dame de Paris
Of Human Bondage
Oliver Twist
Omoo
Our Mutual Friend
Persuasion
Peter Pan
Peter Pan in Kensington Gardens
Phantom of the Opera
Pollyanna
Pride and Prejudice
Prince Otto
Rob Roy
Robinson Crusoe
Sense and Sensibility
She
Shirley
Silas Marner
Sons and Lovers
Stalky and Company
Stories from the Bible
Sylvie and Bruno
Sylvie and Bruno Concluded
Tale of Two Cities
Tales from Shakespeare
Tales of Mystery and Imagination
Tess of the d'Urbervilles
The Adventures of Sherlock Holmes
The Adventures of Tom Sawyer
The Age of Innocence
The Aspern Papers

The Brothers Karamazov
The Call of the Wild
The Castle of Otranto
The Dynamiter
The Expedition of Humphry Clinker
The Heart of Darkness
The History of Rasselas Prince of Abyssinia
The Hound of the Baskervilles
The Jungle Book
The Last of the Mohicans
The Life and Opinions of Tristram Shandy Gent
The Man Upstairs
The Mayor of Casterbridge
The Memoirs of Sherlock Holmes
The Old Curiosity Shop
The Pickwick Papers
The Picture of Dorian Gray
The Pilgrims Progress
The Portrait of a Lady
The Prairie
The Prisoner of Zenda
The Rainbow
The Red Badge of Courage
The Scarlet Letter
The Scarlet Pimpernel
The Sea-Wolf
The Secret Agent
The Sign of Four
The Tenant of Wildfell Hall
The Three Musketeers
The Turn of the Screw
The Valley of Fear
The Vicar of Wakefield
The Virginian
The Warden
The Water Babies
The Way of All Flesh
The Werewolf
The Woman in White
Three Men in a Boat
Through the Looking Glass
Tom Browns School Days
Tom Jones
Tommy and Co.
Treasure Island
Typee
Ulysses
Uncle Toms Cabin
Under Western Eyes
Valperga
Vanity Fair
Vathek an Arabian Tale
Villette
War and Peace
Washington Square
Westward Ho!
What Katy Did Next
White Fang
Wives and Daughters
Women in Love
Wuthering Heights

Appendix B

Source Code

The entire source code may be downloaded as http://lisp-p.org/conc/conc.cpio.bz2 or as a tar-ball from http://lisp-p.org/conc/conc.tar.gz.

The individual source code files may be browsed at http://lisp-p.org/conc/src/. The files are:

permissions	**links**	**size (octets)**	**modification time**	**filename**
-rw-r–r–	1	1778900	Oct 3 15:24	concordances.lisp
-rw-r–r–	1	157517	Oct 3 15:24	concordances.txt
-rw-r–r–	1	2287	Oct 3 15:24	dependence.pl
-rw-r–r–	1	1020	Oct 3 15:24	door_by_corpus.pl
-rw-r–r–	1	975	Oct 3 15:24	door_r_macro.pl
-rw-r–r–	1	1413	Oct 3 15:24	fixlisp.pl
-rw-r–r–	1	947	Oct 3 15:24	fruitdoor_r_macro.pl
-rw-r–r–	1	978	Oct 3 15:24	fruit_r_macro.pl
-rw-r–r–	1	3819	Oct 3 15:24	get-concordance.pl
-rw-r–r–	1	992	Oct 3 15:24	get-concordances.sh
-rw-r–r–	1	2947	Oct 3 15:24	get-sentences.pl
-rw-r–r–	1	7371	Oct 3 15:24	process-data.pl
-rw-r–r–	1	1450	Oct 3 15:24	p-val-hash.pl
-rw-r–r–	1	1121	Oct 3 15:24	r_macro.pl
-rw-r–r–	1	1046	Oct 3 15:24	runit.sh
-rw-r–r–	1	1340	Oct 3 15:24	sample.pl
-rw-r–r–	1	1190578	Oct 3 15:24	wordhash_pvalues.lisp
-rw-r–r–	1	19665	Oct 3 15:24	wordsense.lisp

Appendix C

Change Log

2004-Oct-17 Finally, Figure 4.1 looks right. I'm not convinced it'll look right everywhere, so I added a fail-safe comment to the caption that directs the human reader to download the PNG or Post Script files if the figure appears incorrect. This figure was a real pain in the ass.

Appendix D

Other File Formats

- This document is available in multi-file HTML format at http://lisp-p.org/conc/.
- This document is available in PostScript format at http://lisp-p.org/conc/conc.ps. (It's easy to print PostScript files, even from Microsloth Winders. [**?**, gms:psw]

Bibliography

[1] A.M. Garsia. Arithmetic properties of bernoulli convolutions. *Transactionf of the American Mathematical Society*, 102(3):409–432, 1962.

[2] A.M. Garsia. Entropy and singularity of infinite convolutions. *Pacific Journal of Mathematics*, 13:1159–1169, 1963.

[3] F.S. Hill and M.A. Blanco. Random geometric series and intersymbol interference. *IEEE Transactions on Information Theory*, 19(3), May 1973.

[4] C.D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing.* The MIT Press, Cambridge, Massachussetts, 1999.

[5] E. Margolis and S. Laurence, editors. *Concepts: Core Readings.* MIT Press, 1999.

[6] Y. Peres, W. Schlag, and B. Solomyak. Sixty years of bernoulli convolutions. In C. Bandt, S. Graf, and M. Zaehle, editors, *Fractal Geometry and Stochastics II*, volume 46 of *Progress in Probability*, pages 39–65. Birkhauser, 2000.

[7] Y. Peres and B. Solomyak. Absolute continuity of bernoulli convolutions, a simple proof. *Math. Research Letters*, 3:231–236, 1996.

[8] H. Schütze. Disambiguation and connectionism. In Y. Ravin and C. Leacock, editors, *Polysemy: Theoretical and Computational Approaches.* Oxford University Press, New York, 2000.

[9] H. Sugiyama and A. Huzii. On the distribution function of a random power series with bernoulli variables as coefficients. *IEEE Transactions on Information Theory*, 41(6), November 1995.