11 The Distribution of the Inner Coincidence Index

First we calculate the exact mean value of the inner coincidence index $\varphi(a)$ for $a \in \Sigma^r$. Then we determine empirical values for mean value and variance for English, German, and random texts by simulation, as we did for κ.

The exact value of the variance leads to a somewhat more complicated calculation. We omit it.

Mean Value

We calculate the mean value of the letter frequencies $m_s(a)$ over $a \in \Sigma^r$ for each $s \in \Sigma$. Because of the symmetry in s all these values are identical, therefore we have

$$n \cdot \sum_{a \in \Sigma^r} m_s(a) = \sum_{s \in \Sigma} \sum_{a \in \Sigma^r} m_s(a) = \sum_{a \in \Sigma^r} \underbrace{\sum_{s \in \Sigma} m_s(a)}_{r} = r \cdot n^r$$

This gives the mean value

$$\frac{1}{n^r} \sum_{a \in \Sigma^r} m_s(a) = \frac{r}{n}$$

for each letter $s \in \Sigma$.

Next we calculate the mean value of $\kappa_q(a)$ over $a \in \Sigma^r$. We treat the indices of the letters of the texts a as elements of the cyclic additive group $\mathbb{Z}/n\mathbb{Z}$. Then we have

$$\begin{aligned}
\sum_{a \in \Sigma^r} \kappa_q(a) &= \sum_{a \in \Sigma^r} \frac{1}{r} \#\{j \in \mathbb{Z}/n\mathbb{Z} \mid a_{j+q} = a_j\} \\
&= \frac{1}{r} \sum_{j \in \mathbb{Z}/n\mathbb{Z}} \sum_{a \in \Sigma^r} \delta_{a_{j+q}, a_j} \\
&= \frac{1}{r} \sum_{j \in \mathbb{Z}/n\mathbb{Z}} \underbrace{\#\{a \in \Sigma^r \mid a_{j+q} = a_j\}}_{n^{r-1}} \\
&= n^{r-1}
\end{aligned}$$

because in the underbraced count for a we may choose $r - 1$ letters freely, and then the remaining letter is fixed. This gives the mean value

$$\frac{1}{n^r} \sum_{a \in \Sigma^r} \kappa_q(a) = \frac{1}{n}$$

for each $q = 1, \ldots, r - 1$.

Now for φ. We use the additivity of the mean value.

$$\begin{aligned}\frac{1}{n^r}\sum_{a\in\Sigma^r}\varphi(a) &= \frac{1}{r-1}\left[\frac{1}{n^r}\sum_{a\in\Sigma^r}\kappa_1(a)+\cdots+\frac{1}{n^r}\sum_{a\in\Sigma^r}\kappa_{r-1}(a)\right]\\ &= \frac{1}{r-1}\cdot(r-1)\cdot\frac{1}{n}=\frac{1}{n}\end{aligned}$$

We have shown:

Proposition 4 *The mean values of the q-th autocoincidence index for* $q = 1,\ldots,r-1$ *and of the inner coincidence index over* $a \in \Sigma^r$ *each are* $\frac{1}{n}$.

And for the letter frequencies we have:

Corollary 3 *The sum of the letter frequencies* $m_s(a)$ *over* $a \in \Sigma^r$ *is*

$$\sum_{a\in\Sigma^r} m_s(a) = r\cdot n^{r-1}$$

for all letters $s \in \Sigma$.

Corollary 4 *The sum of the squares* $m_s(a)^2$ *of the letter frequencies over* $a \in \Sigma^r$ *is*

$$\sum_{a\in\Sigma^r} m_s(a)^2 = r\cdot(n+r-1)\cdot n^{r-2}$$

for all letters $s \in \Sigma$.

Proof. By the Kappa-Phi Theorem we have

$$\sum_{t\in\Sigma}\left[\sum_{a\in\Sigma^r} m_s(a)^2 - \sum_{a\in\Sigma^r} m_s(a)\right] = r\cdot(r-1)\cdot\sum_{a\in\Sigma^r}\varphi(a) = r\cdot(r-1)\cdot n^{r-1}$$

Substituting the result of the previous corollary and using the symmetry of the sum of squares with respect to s we get

$$n\cdot\sum_{a\in\Sigma^r} m_s(a)^2 = \sum_{t\in\Sigma}\sum_{a\in\Sigma^r} m_s(a)^2 = r\cdot(r-1)\cdot n^{r-1}+rn\cdot n^{r-1} = r\cdot n^{r-1}\cdot(r-1+n)$$

Dividing by n we get the above formula. $\diamond$

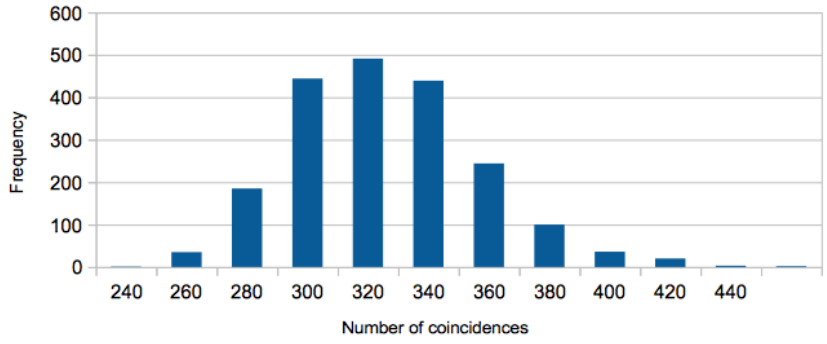

Figure 13: *Frequency of inner coincidence counts for 2000 English texts of 100 letters—to get φ values divide x-values by* 4950

Table 24: *Distribution of φ for 2000 English texts of 100 letters*

Minimum:	0.0481		
Median:	0.0634	Mean value:	0.0639
Maximum:	0.0913	Standard dev:	0.0063
1st quartile:	0.0594	5% quantile:	0.0549
3rd quartile:	0.0677	95% quantile:	0.0750

The Phi Distribution for English Texts

For empirically determining the distribution of the inner coincidence index $\varphi(a)$ we use the Perl program `phistat.pl` from `http://www.staff.uni-mainz.de/pommeren/Cryptology/Classic/Perl/`. For English texts (or text chunks) a, we again take a large English text—in this case the book *The Fighting Chance* by Robert W. Chambers from Project Gutenberg—and chop it into chunks $a, b, c, d, \ldots$ of r letters each. Then we count $\varphi(a)$, $\varphi(b)$, $\ldots$ and list the values in the first column of a spreadsheet. See the file `EnglPhi.xls` in `http://www.staff.uni-mainz.de/pommeren/Cryptology/Classic/Files/`. The text has 602536 letters. We take the first 262006 of them and consider the first 2000 pieces of 100 letters each. Table 24 and Figure 13 show some characteristics of the distribution.

The Phi Distribution for German Texts

We repeat this procedure for German texts, using *Scepter und Hammer* by Karl May. We already consumed its first 400000 letters for κ. Now we take the next 200000 letters—in fact we skip 801 letters in between—and form

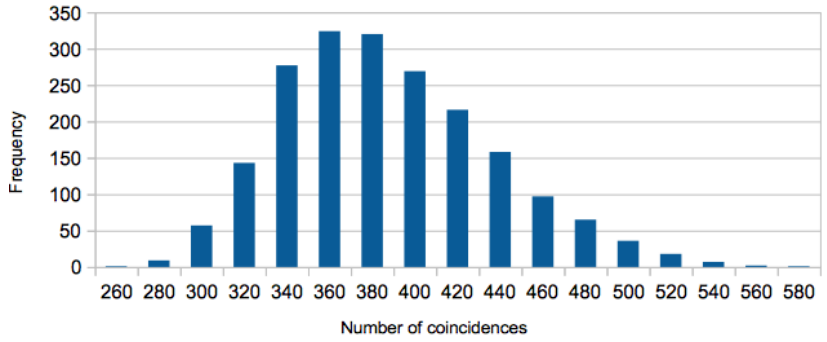

Figure 14: *Frequency of inner coincidence counts for 2000 German texts of 100 letters—to get φ values divide x-values by* 4950

Table 25: *Distribution of φ for 2000 German texts of 100 letters*

Minimum:	0.0517		
Median:	0.0752	Mean value:	0.0763
Maximum:	0.1152	Standard dev:	0.0099
1st quartile:	0.0689	5% quantile:	0.0618
3rd quartile:	0.0828	95% quantile:	0.0945

2000 text chunks with 100 letters each. The results are in Table 25 and Figure 14.

The Phi Distribution for Random Texts

And now the same procedure for random text. The results are in Table 26 and Figure 15.

Table 26: *Distribution of φ for 2000 random texts of 100 letters*

Minimum:	0.0331		
Median:	0.0398	Mean value:	0.0401
Maximum:	0.0525	Standard dev:	0.0028
1st quartile:	0.0382	5% quantile:	0.0360
3rd quartile:	0.0418	95% quantile:	0.0451

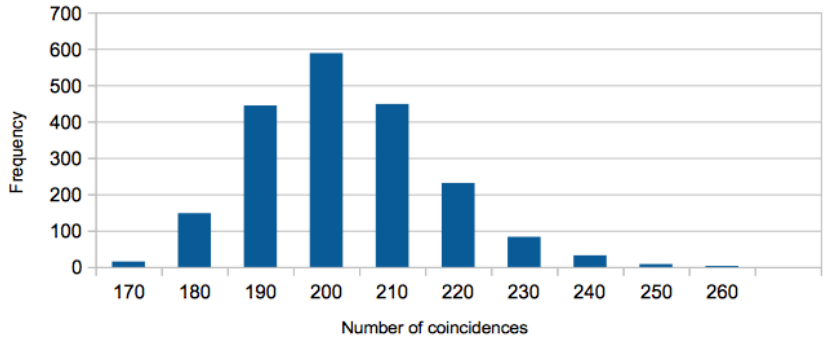

Figure 15: *Frequency of inner coincidence counts for 2000 random texts of 100 letters—to get φ values divide x-values by* 4950

Applications

To which questions from the introduction do these results apply?

We can decide whether a text is from a certain language. This includes texts that are monoalphabetically encrypted because φ is invariant under monoalphabetic substitution. And *we can recognize a monoalphabetically encrypted ciphertext.*

For both of these decision problems we calculate the coincidence index $\varphi(a)$ of our text a and decide "belongs to language" or "is monoalphabetically encrypted"—depending on our hypothesis—if $\varphi(a)$ reaches or surpasses the 95% quantile of φ for random texts of the same length—if we are willing to accept an error rate of the first kind of 5%.

For a text of 100 letters the threshold for φ is about 0.0451 by Table 26. Tables 24 and 25 show that English or German texts surpass this threshold with high probability: For both languages the test has a power of nearly 100%.

It makes sense to work with the more ambitious "significance level" of 1% = bound for the error of the first kind. For this we set the threshold to the 99% quantile of the φ distribution for random texts. Our experiment for texts of length 100 gives the empirical value of 0.0473, failing the empirical minimum for our 2000 English 100 letter texts, and sitting far below the empirical minimum for German. Therefore even at the 1%-level the test has a power of nearly 100%.

The Phi Distribution for 26 Letter Texts

Since the φ test performs so excellently for 100 letter texts we dare to look at 26 letter texts—a text length that occurs in the Meet-in-the-Middle attack against rotor machines.

Table 27: *Distribution of φ for 2000 English texts of 26 letters*

Minimum:	0.0227		
Median:	0.0585	Mean value:	0.0606
Maximum:	0.1385	Standard dev:	0.0154
1st quartile:	0.0492	5% quantile:	0.0400
3rd quartile:	0.0677	95% quantile:	0.0892

Table 28: *Distribution of φ for 2000 German texts of 26 letters*

Minimum:	0.0308		
Median:	0.0708	Mean value:	0.0725
Maximum:	0.1785	Standard dev:	0.0204
1st quartile:	0.0585	5% quantile:	0.0431
3rd quartile:	0.0831	95% quantile:	0.1108

Here we give the results as tables only.

The decision threshold on the 5%-level is 0.0585. For English texts the test has a power of only 50%, for German, near 75%. So we have a method to recognize monoalphabetic ciphertext that works fairly well for texts as short as 26 letters.

Table 29: *Distribution of φ for 2000 random texts of 26 letters*

Minimum:	0.0154		
Median:	0.0400	Mean value:	0.0401
Maximum:	0.0954	Standard dev:	0.0112
1st quartile:	0.0338	5% quantile:	0.0246
3rd quartile:	0.0462	95% quantile:	0.0585