School of Electrical & Electronic Engineering



Group 32: Code Cracking: Who Murdered The Somerton Man?



Final Seminar Supervisors: Professor Derek Abbott and Dr. Matthew Berryman Students: Nicholas Gencarelli and Jikai Yang

Outline of Presentation

- Introduction
 - History and Background
 - Project Aims and Motivation
- Approach and Methodology
- Previous Studies
- Specific Tasks
 - 1. Statistical Frequency Analysis of Letters
 - 2. N-Gram Search
 - Rubaiyat Omar Khayyam as One-Time pad
 - 4. Statistical Frequency of Letters Reanalysis

- Project Management
 - Task Allocation
 - Budgeting
 - Risk Assessment
 - Gantt Chart
 - Communication
- Relationship to possible career paths
- References
- Questions

Introduction: History and Background



Figure 1: The Herald Sun

Who:

An unidentified man

Time of death:

Found at 6:30am on the 1st of December 1948

Cause of death:

Unknown

Location of death:

Found at Somerton Beach, South Australia





Introduction: History and Background

- Why did the public show significant interest in this case?
 - The man lacked any identification
 - The cause of death was unknown, potentially an undetectable poison
 - The case occurred during the heightened tension of the Cold War
 - A page from a book with some letters written on it was found in his pocket
 - These letters have been a mystery until today



Introduction: Existing Evidence

- A brown suitcase
 - Found at the Adelaide Railway Station
 - A package of thread was found inside the suitcase that was used to repair the lining in the trousers the Somerton Man was wearing



Figure 2: The Herald Sun

Introduction: Existing Evidence



Figure 3: The Herald Sun

- A scrap of paper
 - Found in the man's trouser pocket
 - "Tamam Shud" translating to "it is finished" in Persian



Introduction: Existing Evidence

 Letters written on a page out of *Rubaiyat of Omar Khayyam.*

> WRGOABABD MLIAOI WTBIMPANETP MLIABOAIAQC ITTMTSAMSTGAB



Figure 4: The Herald Sun





Introduction: Project Aims and Motivation

- In this project, our group will attempt to solve a suspected murder that took place in Adelaide in 1948.
- The ultimate aim is to decrypt the code in order to solve the mystery, but this may be somewhat unrealistic as the code has remained uncracked for many years.
- However, we will utilise computational techniques to attempt the decryption.





Approach and Methodology: *P*-Value

- In statistics, the *p*-value is a function of the observed sample results that is used for testing a statistical hypothesis.
- We will use *p*-value to do some simple tests to determine if English was the language in the original message.



Approach and Methodology: Chi-Squared test

- Chi-square is a statistical test commonly used to compare observed data with expected data
- Since it is a non-parametric test, it doesn't care about distribution of samples
- 'Goodness of Fit' test, meaning it measures how closely one group of data is related to another

 $(observed - expected)^2$

expected

 We will use chi-squared test to calculate and analyse our task 1 Statistical Frequency Analysis of Letters

English		Frequency	Code	Frequency		
A:	7935	0.108153	8	0.181818		
B:	3038	0.041408	5	0.113636		
_						



(8-(44×0.10815 44×0.108153



Approach and Methodology: N-Gram Model

- An *n*-gram model is a type of probabilistic language model for predicting the next item in a sequence in the form of a (*n* – 1)order.
- We will be using an n-gram database to find common groups of words for a variety of initialisms.

Sample	1-gram	2-gram	3-gram	<i>n</i> -gram		
Sequence	sequence	sequence	sequence	sequence		
MLIABOAIAQ C	M, L, I, A, B, O, A, I, A, Q, C	M, L, I, A, B, O, A, I, A, Q, C BO, OA, AI, IA, AQ, QC				



Approach and Methodology: One-Time pad

- In cryptography, a one-time pad is an encryption technique that cannot be cracked if used correctly.
- We should be able to verify and prove if the code is the beginning letter of a sequence of words, or is composed of whole words.



Figure 6: Wikipedia

Approach and Methodology: Universal Declaration of Human Rights

- In 1948, the United Nations • General Assembly adopted this bill of human rights to represent a common standard for all people and all nations.
- There are 444 different • language translations.



NOT TRUE PORT ment for all peoples and all nations, to the end the ery individual and every aroun of society keeping this Declaratio undantly in mind, shall strive by touching spect for these rights and fr

national, to secure their uni cound rights of men and gromen and hore hempelies and among the peoples of territories under their inside

it is essential to promote the develop

essential, if man is not to be compelled to have rec

oples of the United Nations have in the Charter ,

ected by the rule of lan

UNITED NATIONS

Figure 7: British Library



Approach and Methodology: Project Gutenberg

- PG is a volunteer effort to digitise and archive cultural works.
- First document posted in 1971.
- Over 50,000 books are available in plain text and other formats like HTML, PDF and EPUB as eBooks.
- It is all free to use and includes major languages in the world.





Previous Studies: Previous Professional Attempts

- Limited techniques
- Fixed assumptions
- Navy: "Neither a code nor a cipher"
- Department of Defense:
 - "There are insufficient symbols to provide a pattern."
 - "The symbols could be a complex substitute code or the meaningless response to a disturbed mind."
 - "It is not possible to provide a satisfactory answer."



Previous Studies: Honours Projects 2009-2013

- Letter frequency analysis
- Initial letter and sentence letter probabilities
- Probabilities of known cypher techniques
- One-time pads
- Web Crawler
- Text type analysis
- Pattern matching
- 3D generated reconstruction of bust (See Fig. 8)
- Mass spectrometer data hair analysis



Figure 8: Final Report 2012



Previous Studies: Honours Projects 2009-2013

Conclusions:

- Letters unlikely to be random
- Likely to be initialism
- Likely to use Rubaiyat as one-time pad
- Likely to be English
- Unlikely to be an initialism of a poem
- Rubaiyat not used as straight substitution one-time pad



Specific Tasks:

Task 1: Statistical Frequency Analysis of Letters

Aims:

- Use statistical methods to verify whether the most possible language of Somerton Code is English
- Critical review of previous results
- Universal Declaration of Human Rights as base text
- Find out how common letters are in each language





Specific Tasks:

Task 1: Statistical Frequency Analysis of Letters

Method:

- Initially tried ANOVA, but did not produce valid results since used mean and standard deviation of letter frequencies
- Instead attempted using non-parametric Chi-Squared testing method





Specific Tasks: Task 1: Statistical Frequency Analysis of Letters

• During our chi-squared value calculation, the count number of some letters from sample languages were 0. Using the chi-squared value formula:

 $\frac{(\text{observed} - \text{expected})^2}{\text{expected}}$

It resulted in expected values becoming 0. Since the denominator cannot be 0, we instead used a count of 1 instead of 0 for these letters. We also used 0.0001 instead of frequency 0 for the letters to double check our result.



Specific Tasks: Task 1: Statistical Frequency Analysis of Letters

Chi-Squared Values of Closest 20 Lnguages based on 2013 Squared Difference (5Ms & 1W in Code)



Closest 20 Languages by Squared Difference EUROPE



(Count from 0 to 1)



Specific Tasks:

Task 1: Statistical Frequency Analysis of Letters



(Count from 0 to 1)

Specific Tasks: Task 1: Statistical Frequency Analysis of Letters

Chi-Squared Values of Closest 20 Languages based on 2013 Squared Difference (5M 1W)



Closest 20 Languages by Squared Difference EUROPE



(Frequency 0.0001)

Specific Tasks:

Task 1: Statistical Frequency Analysis of Letters



Specific Tasks: Task 1: Statistical Frequency Analysis of Letters

Top 20 European Languages based on Estimated Number of Speakers (Average)



Top 21 European Languages based on Estimated Number of Speakers (Average)



(Count 0 to 1)



Specific Tasks: Task 1: Statistical Frequency Analysis of Letters

Top 20 European Languages based on Estimated Number of Speakers vs Thomas Hardy Sample(number 0 to 1)

Top 20 European Languages based on Estimated Number of Speakers vs Thomas Hardy Sample(frequency 0 to 0.001)



Specific Tasks:

Task 1: Statistical Frequency Analysis of Letters

Results for using Chi-squared test:

- Previous results using the Universal Declaration of Human rights as a base text were confirmed
- The Universal Declaration of Human rights as a base text was found to have too small a sample size to accurately reflect letter frequencies in languages
- English had the lowest chi-squared value for most of the calculations, which means the Somerton Code is most likely be English. However, we still could not get a reasonable p-value for any languages against the code, so there was potential for reanalysis



Specific Tasks: Task 2: N-Gram Search

- Previous studies suggest code is an initialism
- Use of N-gram database increasing search speed
- Maximum 5-gram groups
- All variants of ambiguous letters
- Output most likely grams
- Repeated for lower grams (n=5-1)
- Large Data Set, limited processing power / time
- Processed using 'Amazon Elastic Compute Cloud'



Specific Tasks: Task 2: N-Gram Search Top Results

- that such a man should
- of it was to be
- Is this the man that
- it was to be in
- of it were to be
- so a man sharpeneth the
- I thought to myself that
- may serve to give a
- and I am quite convinced
- one is willing to be



Specific Tasks: Task 2: N-Gram Search Window Results

would rather go on as but all by different mothers late in August or in was the best in my processes are not easy to people must live in a be of an intensity akin quite clear in the text matter thus submitted as may supposed to get a blood



- Investigation that the letters have been substituted for others using a one-time pad technique using the code and the Rubaiyat as the pad.
- We are proposing to use different letter positions as keys to decode.
- We want to find some meaningful decoding messages by using this method.



Figure 9. Derek Abbott's Wiki Project

Method:

- Somerton Man Code acted as the cipher text, the *plaintext* of the code was desired to be deciphered using a *key* and the Rubaiyat of Omar Khayyam acted as the *onetime pad*.
- We used a computer program language called Matlab to implement the deciphering technique.

- Decoded using letter position within each word rather than using numbers assigned to each letter in the alphabet
- Also decoded using last letter of each word
- Verification using own encoded message



Quick example:

- Assuming we have Code message AFM and we will use second letter to decode it.
- 2. The program will search the Rubaiyat from beginning to end, until it finds the first word that begin with A. This will then be decoded to the second letter in the same word which is W.
- 3. The program will repeat for rest letters in the Code.
- 4. The decoding message should be WOO.

AWAKE FOR MORNING ΤN THF: BOWL OF. >> multi ('AFM', 2) NIGHT HAS Message:AFM FLING Message:WOO THE STONE THAT PITTS THE STARS ΤO

Decoding using second letter (Rubaiyat without formatting)

>> multi ('WRGOABABD', 2)
Message:WRGOABABD
Message:hoafnrnar
>> multi ('WTBIMPANETP', 2)
Message:WIBIMPANETP
Message:hhineononhi
>> multi ('MLIABOAIAQC', 2)
Message:MLIABOAIAQC
Message:oornufnnnua
>> multi ('ITIMISAMSIGAB', 2)
Message:IITMISAMSIGAB
Message:rhhoaunaoaasa

Decoding using third letter (Rubaiyat without formatting)

>> multi ('WRGOABABD', 3) Message:WRGOABABD Message:hoafnrnar Message:esrh,hsls >> multi ('WIBIMPANEIP', 3) Message:WIBIMPANEIP Message: hhineononhi Message:eartmtdtdie >> multi ('MLIABOAIAQC', 3) Message: MLIABOAIAQC Message: oornufnnnua Message:r!adth,tdan >> multi ('ITTMTSAMSTGAB', 3) Message: ITIMISAMSIGAB Message: rhhoaunaoaasa Message:aaenipdhvsrhh

Decoding using second letter (Rubaiyat with formatting)

>> multi ('WRGOABABD', 2)
Message:WRGOABABD
Message:HERNNYNUI
>> multi ('WIBIMPANETP', 2)
Message:WIBIMPANETP
Message:HHENAUNONHA
>> multi ('MLIABOAIAQC', 2)
Message:OONWEPNRLU0
>> multi ('IIIMTSAMSIGAB', 2)
Message:IITMISAMSIGAB
Message:NHHYHINAOORLU

Decoding using third letter (Rubaiyat with formatting)

>> multi ('WRGOABABD', 3) Message: WRGOABABD Message:HERNNYNUI Message:EIOECODIV >> multi ('WIBIMPANEIP', 3) Message:WIBIMPANEIP Message: HHENAUNONHA Message: EEFSYIDGOIR >> multi ('MLIABOAIAQC', 3) Message: MLIABOAIAQC Message: OONWEPNRLUO Message:RGSDFEDALAR >> multi ('IIIIMISAMSIGAB', 3) Message: ITIMISAMSIGAB Message:NHHYHTNAOORLU Message: SEOYEUDNAKRHR

Last letter decoding (Rubaiyat without formatting)

>> multi ('WRGOABABD', 100)

Message:e, nfddAk!

>> multi ('WIBIMPANEIP', 100)

Message:etdInperser

>> multi ('MLIABOAIAQC', 100)

Message:g!I, en, mdle

>> multi ('IITMTSAMSTGAB', 100)
Message:IneStgdhlins,

Last letter decoding (Rubaiyat with formatting)

>> multi ('WRGOABABD', 100)
Message:NNDETYDTE
>> multi ('WTBIMPANETP', 100)
Message:NEENYSDOWKE
>> multi ('MLIABOAIAQC', 100)
Message:GONAENDMLLR
>> multi ('ITIMTSAMSTGAB', 100)
Message:NEEYEDDYLODLT

Verification:

Assuming we have origin message GUN, and use third letter to encode the message based on Rubaiyat. The encoded words from Rubaiyat should be NIGHT, FLUNG and HUNTER. The code should be NFH.

1	AWAKE	8	NIGHT	15	THE
2	FOR	9	HAS	16	STARS
3	MORNING	10	FLUNG	17	TO
4	IN	11	THE	18	FLIGHT
5	THE	12	STONE	19	AND
6	BOWL	13	THAT	20	LO
7	OF	14	PUIS	21	THE
8	NIGHT	15	THE	22	HUNTER

Result:

If we use NFH as code and choose third letter location as key, by using One-Time pad method. What is the output?



Specific Tasks:

Task 4: Statistical Frequency of Letters Reanalysis

- Increased sample size to counteract letters with frequency 0
- Used Project Gutenberg Novels
- Utilised 2013 group's initial letter count and decoding toolkit
- Compared to initial results
- English P-value benchmark
- Hypothesis test:
 - H₀: The group of letters are from the English language
 - H_1 : The group of letters are from another language



Specific Tasks: Task 4: Statistical Frequency of Letters Reanalysis



Specific Tasks: Task 4: Statistical Frequency of Letters Reanalysis

Initial analysis: **Reanalysis: Code vs Gutenberg European Code vs Declaration European Languages** Languages 4000 3500 18000 Value 16000 3000 Chi-squared value 14000 2500 12000 Chi-squared 10000 2000 8000 1500 6000 4000 1000 2000 500 0 English swedish Portuguese spanish Romanian Clech 0 e German Polish DUTCH HUNBarlan Italian French German Serbian Uzbek Spanish Greek Kurdish English Dutch Russian Italian Turkish Polish French Hungarian Ukrainian Swedish Belarusian Czech Kazakh Portuguese Romanian Languages Language Life Impact | The University of Adelaide

Specific Tasks: Task 4: Statistical Frequency of Letters Reanalysis

Comparison of Chi-Squared Values of Code and English Samples against Gutenberg **Base Text**

Comparison of P-Values of Code and English Samples against Gutenberg Base Text





Project Management: Task Allocation

Task	Allocation				
Proposal Seminar	Together				
Draft Research Proposal	Together				
Research Proposal and Progress Report	Individual				
Project Management	Nicholas				
Task 1: Statistical Frequency Analysis of Letters	Together				
Task 2: N-Gram Search	Nicholas				
Task 3: Rubaiyat of Omar Khayyam as One-Time Pad	Jikai				
Task 4: Statistical Frequency of Letters Re-Analysis	Together				
Exhibition Poster	Nicholas				
Final Seminar	Together				
Project Exhibition	Together				
Honours Thesis/Final Report	Individual				
Youtube Video	Jikai				



Project Management: Budget

Proposed Budget	
Item	Cost
Google N-Gram Database	\$150
Hard Drive	≈\$100
Total	≈\$250

Final Budget	
Item	Cost
Google N-Gram Database	\$0
Amazon EC2 I2 Storage + Processing	\$576.01
Total	≈\$576.01



Project Management: Risk Assessment

Risk	Likelihood	Severity	Risk Estimation	Reduction Strategy				
1. Inaccurate estimation of time and resources	Likely	Moderate	High	Flexible Schedule				
2. Member unable to complete work	to Slight Major		High	Assistance from other member and project supervisors				
3. Member leaves group	Rare	Major	Medium	Encouragement and assistance when needed				
4. Loss of data Rare		Major	Medium	Regular backups				
5. Illness/Absences Unlikely		Minor	Low	Try to stay healthy and inform members as soon as possible				
6. Group conflict	Unlikely	Minor	Low	Report to supervisor				
7. Misunderstanding project tasks	Slight	Minor	Low	Schedule meeting with supervisors				
8. Bugs in code	Likely	Moderate	High	Debugging and testing of code before completion				
9. Inability to decipher the Somerton Man Code	Almost Certain	Negligible	Medium	Complete work to the best of the group's ability				
43			Life Impact	The University of Adelaide				



					2015								
Nam	e	Project	Begin date	End date	March	April	May	June	July	August	September	October	November
6	v	Veekly Progress	3/2/15	10/26/15									
6	F	Project Management Outputs	3/20/15	3/30/15									
6	P	repare Proposal Seminar	3/20/15	3/30/15									
6	F	Proposal Seminar Presentation	3/31/15	3/31/15		*							
e) [Draft Research Proposal	3/31/15	4/16/15		-							
6	0	Draft Research Proposal Due	4/17/15	4/17/15		•							
6	R	Research Proposal and Progress Report	4/18/15	6/11/15									
0	R	Research Proposal and Progress Report Due	6/12/15	6/12/15				•					
Ψ.	1	Tasks:	4/18/15	10/16/15		-							
	1	Task 1: Statistical Frequency Analysis of Letters	4/18/15	5/31/15									
	l	Task 1 Verification	6/1/15	6/7/15									
	1	Task 1 Complete	6/7/15	6/7/15				٠					
	1	Task 2: Web Crawler Re-design	6/13/15	7/16/15	~			and the second second					
		 Task 2 Clean up and present results 	7/16/15	10/5/15									
		Task 2 Complete	10/6/15	10/6/15								•	
	1	Task 3: Rubaiyat of Omar Khayyam as One-Time Pad	6/13/15	9/16/15									
		Task 3 Test	9/16/15	10/6/15									
		Task 3 Complete	10/6/15	10/6/15								•	
		Task 4: Statistical Frequency of Letters Reanalysis	9/15/15	10/1/15								1	
		Task 4 Verification	10/5/15	10/16/15									
	4	Task 4 Complete	10/16/15	10/16/15								•	
6	F	inal Seminar Preparation	9/26/15	10/15/15									
6	F	inal Seminar Presentation	10/13/15	10/13/15								•	
Ψ 0	F	Project Exhibition	10/21/15	10/25/15									
		Exhibition Poster	10/21/15	10/23/15									
		Exhibition Day	10/26/15	10/26/15								٠	
		 Final Performance 	10/26/15	10/26/15								\$	
6	F	inal Report/Honours Thesis	6/13/15	10/21/15								and the second second	
c	F	inal Report/Honours Thesis Due	10/21/15	10/21/15								•	
6		Dump of final work	11/6/15	11/6/15									•
6	0	Treate Youtube Video	11/2/15	11/5/15									
6	S	iubmit Youtube Video	11/6/15	11/6/15									*

Life Impact | The University of Adelaide

Project Management: Gantt Chart Milestones

Met:

- Proposal Seminar Presentation
- Submission of Draft Research Proposal
- Submission of Research Proposal and Progress Report
- Completion of Tasks 1-4
- Final Seminar Presentation

Upcoming:

- Project Exhibition Poster
- Final Performance on Exhibition Day
- Submission of Final Report/Honours Thesis
- Dump of final work
- Creation of project Youtube Video

Project Management: Communication

- Regular face-to-face contact through meetings
- Regular communication via text message and email
- Use of collaborative software:
 - Google Drive
 - GitHub repository
- Project Wiki progress page



Relationship to possible career paths

- Techniques:
 - Software and programming skills
 - Information theory
 - Probability
 - Statistics
 - Encryption and decryption
 - Datamining
 - Database trawling

- Types of jobs:
 - Computer security
 - Communications
 - Digital forensics
 - Computational linguistics, eg. Used in Google Translate



Relationship to possible career paths

- Industries:
 - Software
 - E-finance
 - E-security
 - Telecommunications
 - Search engine
 - IT

- Job examples:
 - Google
 - ASIO
 - ASIS
 - ASD

Life Impact | The University of Adelaide

Future Work

Our Group:

- Increase English base text letter frequency sample size
- Increase number of 44 letter English benchmarks
- Use to find more accurate significance level

Future Groups:

- Extend to all popular European Languages
- Focus on English and explore genre
- Explore more gram combinations by changing window size and shift
- Perform Mass
 Spectrometer Data
 Analysis

Conclusions

- Code not created using Rubaiyat of Omar Khayam as one-time pad and the proposed key method
- Further analysis of N-gram search results to provide valid or useful decryptions of the code
- English is most likely language from which the Somerton Man code was written assuming it is an initialism





Text References

- Author Unknown 1948, "Dead Man Found Lying on Somerton Beach," *The News*, 1 December, p. 1.
- Author Unknown 1949, "Definite Clue in Somerton Mystery," *The Advertiser*, 18 January, p. 1.
- "The Universal Declaration of Human Rights," 1948, *United Nations,* viewed 28 March 2015, http://www.un.org/en/documents/udhr/.
- Project Gutenberg, Author Unknown 2015, viewed 12 October, available https://www.gutenberg.org
- *Project Gutenberg,* Author Unknown 2015, viewed 12 October, available ">https://en.wikipedia.org/wiki/Project_Gutenberg>
- Duffy, A. & Stratfold, T. 2012, "Final Report 2012," *Derek Abbott's Wiki Project*, viewed 28 March 2015, https://www.eleceng.adelaide.edu.au/personal/dabbott/wiki/index.php/Final_Report_2012.
- Griffith, L. & Varsos, P. 2013, "Semester B Final Report 2013 Cipher Cracking," *Derek Abbott's Wiki Project*, viewed 28 March 2015,
 ">https://www.eleceng.adelaide.edu.au/personal/dabbott/wiki/index.php/Semester_B_Final_Report_2013_-
- Maxwell, S. & Johnson, P. 2011, "Final Report 2011," *Derek Abbott's Wiki Project*, viewed 28 March 2015, https://www.eleceng.adelaide.edu.au/personal/dabbott/wiki/index.php/Final_Report_2011.
- Ramirez, K. & Lewis-Vassallo, M. 2010, "Final Report 2010," *Derek Abbott's Wiki Project*, viewed 28 March 2015, https://www.eleceng.adelaide.edu.au/personal/dabbott/wiki/index.php/Final_Report_2010.
- Somerton Beach Mystery 1978, YouTube, ABC, Sydney, viewed 28 March 2015, https://www.youtube.com/watch?v=ieczsZRQnu8>.
- Turnbull A. & Bihari D. 2009, "Final Report 2009: Who killed the Somerton man?" Derek Abbott's Wiki Project, viewed 28 March 2015,
 https://www.eleceng.adelaide.edu.au/personal/dabbott/wiki/index.php/Einal_report_2009; Who, killed the Somerton man?" Derek Abbott's Wiki Project, viewed 28 March 2015,



Image References

- Figures 1-4: Fettes, J 2013, 'Professor's 15-year search for answers seeks to crack the secret code to the death of the 'Somerton man' found on an Adelaide Beach', Herald Sun, viewed 28 March 2015, <http://www.heraldsun.com.au/news/law-order/portrait-may-hold-key-tosomerton-man-beach-mystery/story-fni0ffnk-1226674957043>.
- Figure 7: Universal Declaration of Human Rights 1948, British Library, viewed 28 March 2015, http://www.bl.uk/onlinegallery/takingliberties/staritems/645universaldeclaration humanrightspic.html>.
- Figure 8: Duffy, A. & Stratfold, T. 2012 "Final Report 2012," *Derek Abbott's Wiki Project*, viewed 28 March 2015, https://www.eleceng.adelaide.edu.au/personal/dabbott/wiki/index.php/Final_Report_2012>.
- Figure 9: Rubaiyat Omar Khayyam n.d., *Derek Abbott's Wiki Project*, viewed 28 March 2015, http://www.eleceng.adelaide.edu.au/personal/dabbott/tamanshud/W&T_rubaiyat_wells_copy.pdf.





Questions





End of Presentation

