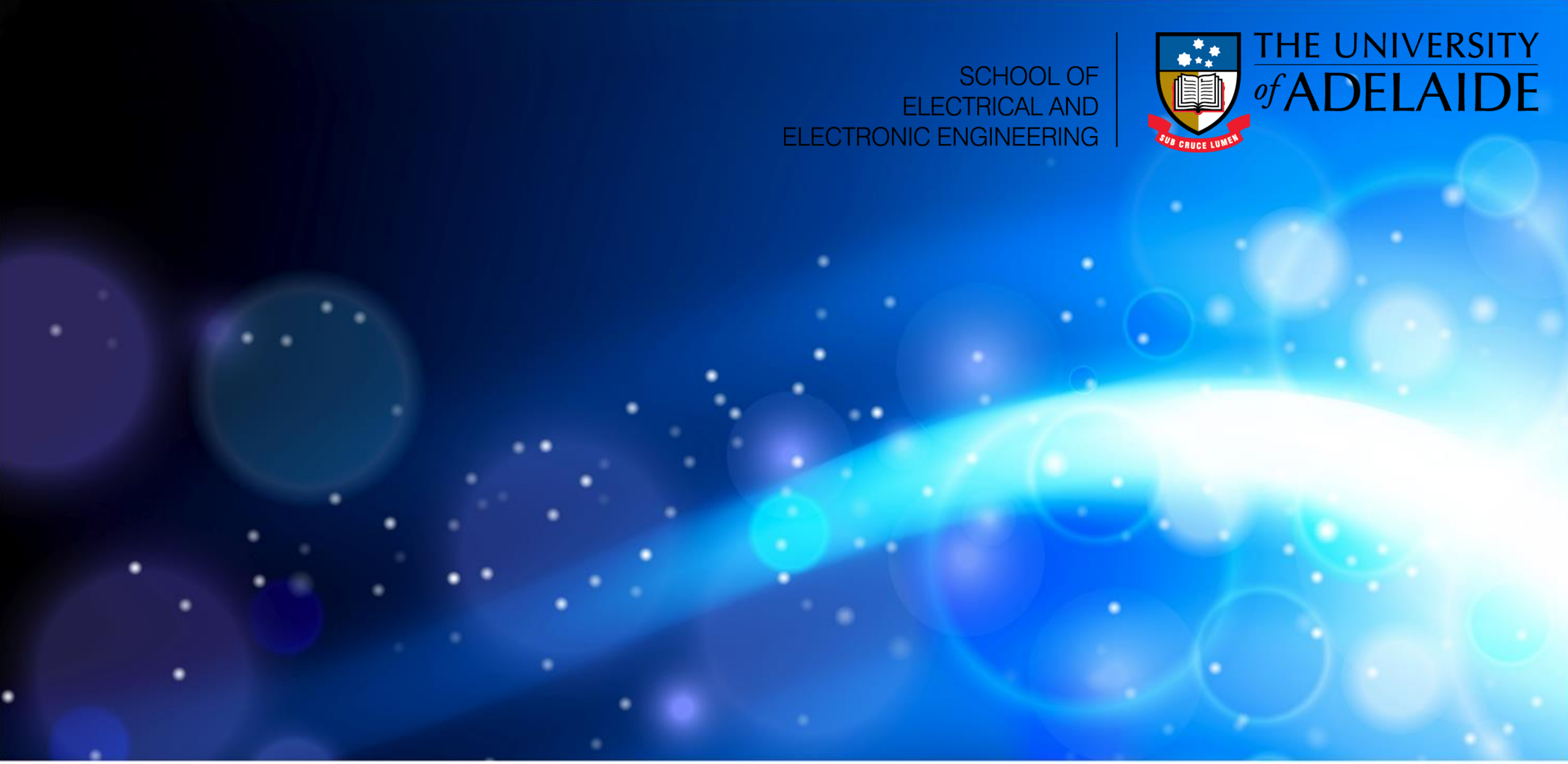


SCHOOL OF
ELECTRICAL AND
ELECTRONIC ENGINEERING

THE UNIVERSITY *of* ADELAIDE

SUB CRUCE LUMEN

CRICOS PROVIDER 00123M

Students: Peter Roush and Bryce Shi

Supervisors: Derek Abbott, Maryam Ebrahimpour and Brian Ng

Project 44: Cracking the Voynich Code

Final Seminar

Outline

- The Voynich Manuscript
    - Objectives, Background and Motivation
- Analysing the Manuscript
    - Techniques, Research, and Testing
- The Information Learnt
    - Results and Analysis
- Project Management
    - Team Roles, Milestones, and Budgeting
- Conclusions

THE UNIVERSITY of ADELAIDE

Background, Motivation, and Objectives

The Voynich Manuscript

A Brief History

- Voynich Manuscript
  - Found in an Italian Castle by Wilfred Voynich, a book collector
  - Pages and some references dated to the 15th Century
  - Author or authors unknown
  - Language unknown
  - Pictures have been inconclusively matched to plants in Europe and South America

- Electronic Transcriptions
  - At least two different languages or dialects
  - Hard to separate letters into a fixed alphabet
  - Interlinear Transcription File

Current Theories

- Early Language or Writing System
  - Early Welsh (Tim Ackerson)
  - Romanised Manchu Chinese (Zbigniew Banasik)
- Code
  - Fake cipher related to Arabic numerals (D'Imperio)
  - Cipher by Roger Bacon (William Newbold)
  - Cipher by Antonio Averlino (Nick Pelling)
  - Certain pages are key to unlocking the mystery (Mark Sullivan)
- Hoax
  - Written to scam money out of Rudolf II (Raphael Mnishovsky)
  - Written by Voynich for money and fame

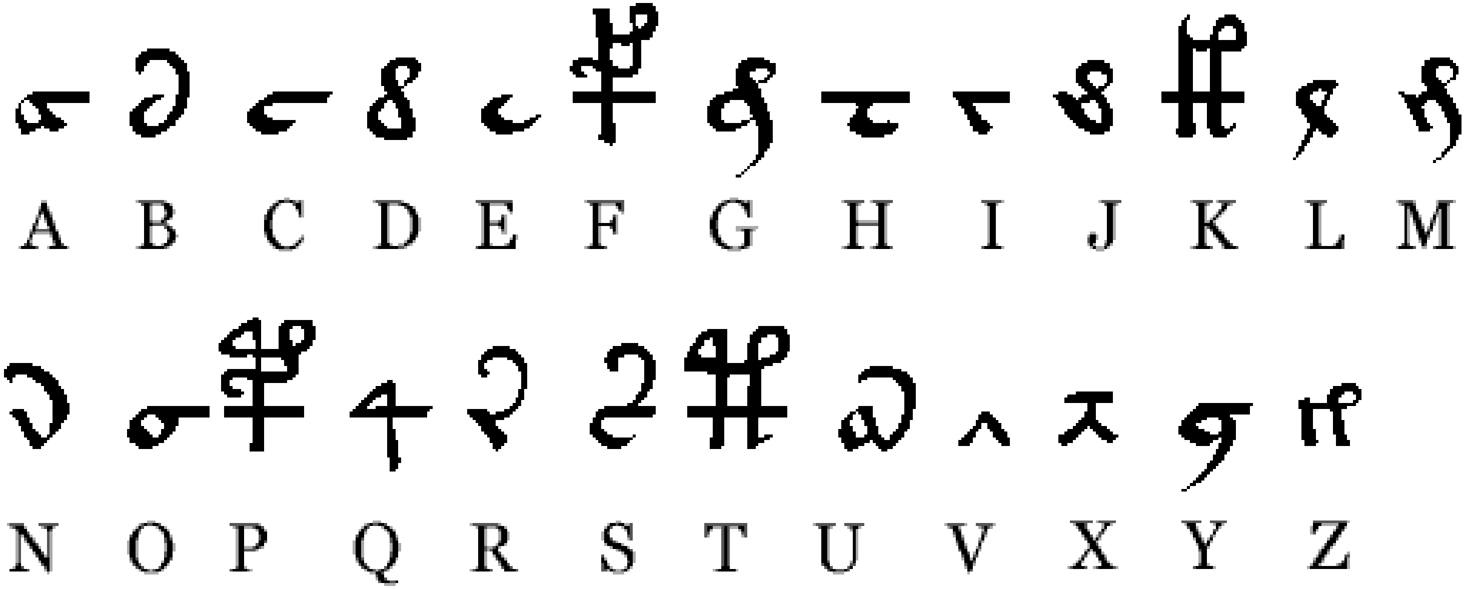
Voynich Manuscript

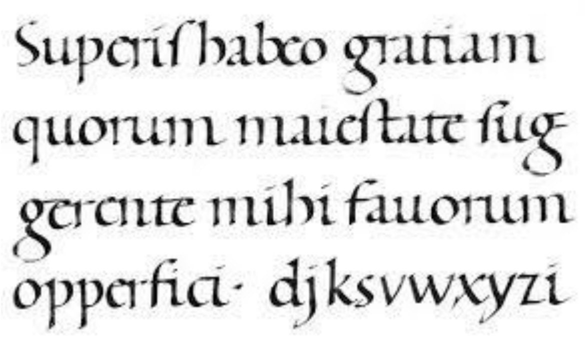
- **Part 1** (Herbal)
  129 pages
- **Part 2** (Astronomical)
  12 pages
- **Part 3** (Biological)
  20 pages
- **Part 4** (Cosmological)
  20 pages
- **Part 5** (Pharmaceutical)
  18 pages
- **Part 6** (Recipes)
  25 pages

Detailed chemical analysis can be found at Yale:
http://beinecke.library.yale.edu/sites/default/files/voynich_analysis.pdf

Pictures reproduced from Beinecke Library under the free public domain licence

Characters (EVA Alphabet)




Superiſ habeo gratiam
quorum maieſtate ſug
gerente mihi fauorum
opperfici· djksvwxyzi

Humanist miniscule writing (left)

Picture from http://www.afternight.com/runes/a-voynich.gif and dictionarytoday.tumblr.com

Objectives

- Develop Data Mining Techniques for the unknown language/code in the Voynich Manuscript.

- Compare linguistic features of the Voynich Manuscript and other languages.

- Determine whether the language in the Voynich manuscript is real, a code, or a hoax.

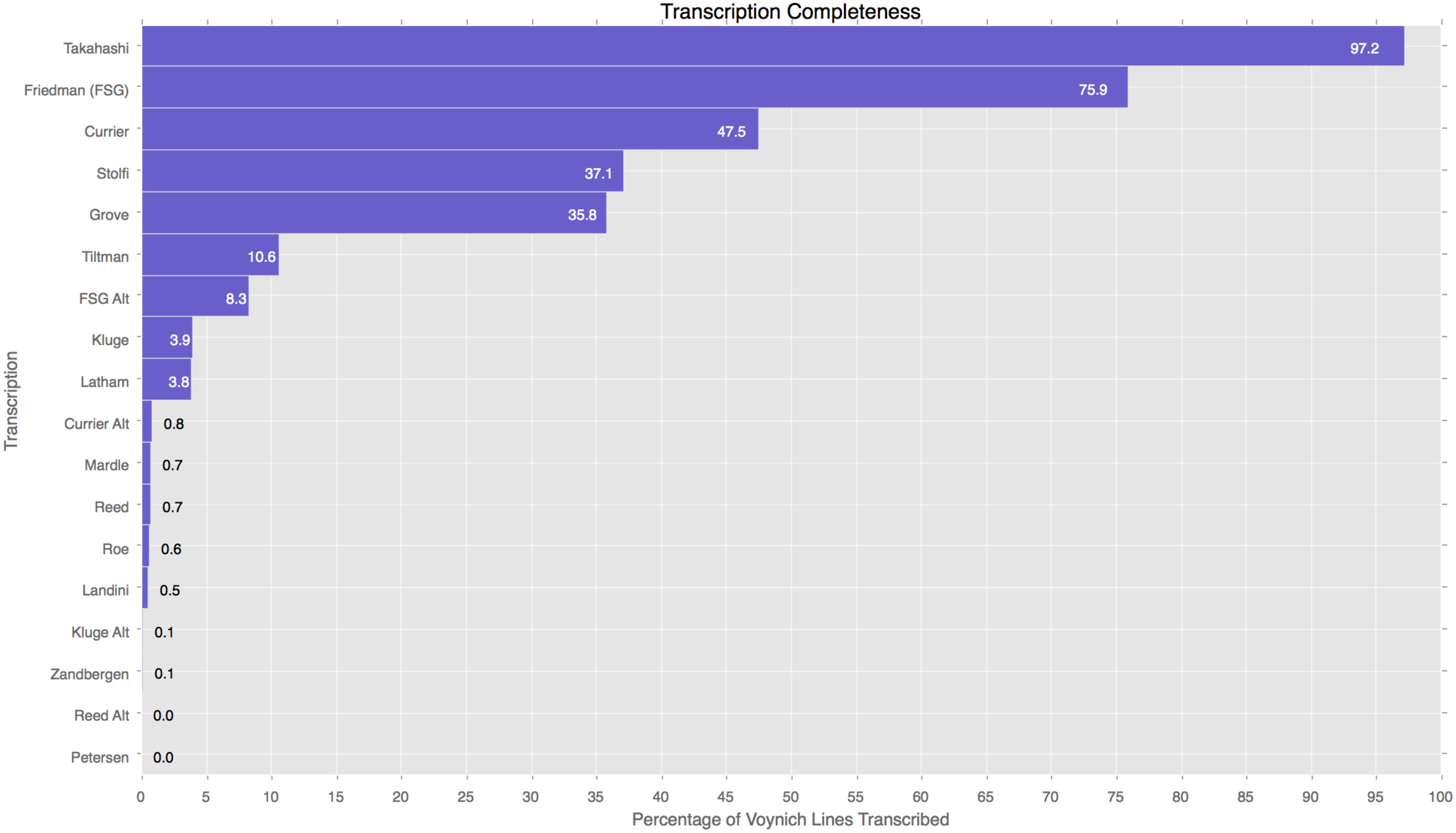
- Develop a code base and documentation to aid future projects.

THE UNIVERSITY of ADELAIDE

Research, Methods and Tools

Analysing The Manuscript

Electronic Transcriptions

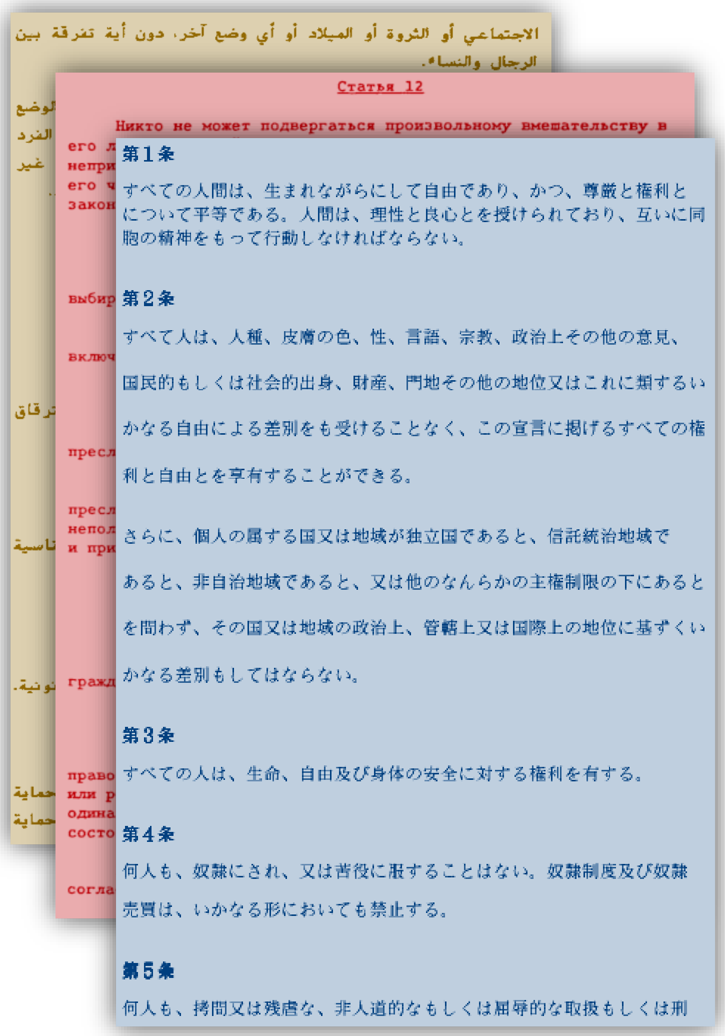


Testing Methodology

- Used Takahashi Transcription and EVA alphabet for all tests
- Handwritten text files for basic verification
- 10 Comparison Texts of similar length in selected languages
  - English (3 Texts)
  - Latin
  - Italian
  - Hungarian
  - Hebrew (Without vowel accents)
  - Chinese (Simplified Characters)
  - Chinese (Pinyin)

The UN Declaration of Human Rights

- 382 translated languages
- Allows greater selection of comparison languages.
- Translations contain an average of 1800 word tokens.



الاجتماعي أو الثروة أو الميلاد أو أي وضع آخر، دون أية تفرقة بين
الرجال والنساء.

Статья 12

Никто не может подвергаться произвольному вмешательству в

第１条

すべての人間は、生まれながらにして自由であり、かつ、尊厳と権利とについて平等である。人間は、理性と良心とを授けられており、互いに同胞の精神をもって行動しなければならない。

第２条

すべて人は、人種、皮膚の色、性、言語、宗教、政治上その他の意見、国民的もしくは社会的出身、財産、門地その他の地位又はこれに類するいかなる自由による差別をも受けることなく、この宣言に掲げるすべての権利と自由とを享有することができる。

さらに、個人の属する国又は地域が独立国であると、信託統治地域であると、非自治地域であると、又は他のなんらかの主権制限の下にあるとを問わず、その国又は地域の政治上、管轄上又は国際上の地位に基ずくいかなる差別もしてはならない。

第３条

すべての人は、生命、自由及び身体の安全に対する権利を有する。

第４条

何人も、奴隷にされ、又は苦役に服することはない。奴隷制度及び奴隷売買は、いかなる形においても禁止する。

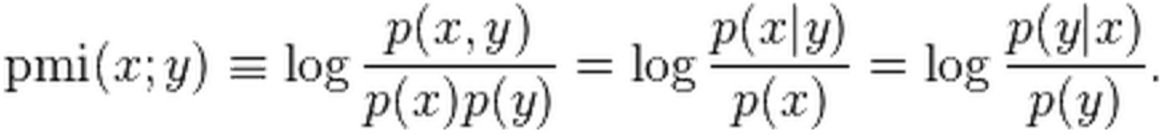
第５条

何人も、拷問又は残虐な、非人道的なもしくは屈辱的な取扱もしくは刑

Picture Reproduced From: www.boes.org (Public Domain)

Collocations

- A collocation is a word combination that occurs more often than would be expected by chance:
  - “Strong Tea”
  - “Friendly Footing”
  - “Saucer of Milk”
  - “Scotland Yard”
- Collocations indicate names and expressions in a language, and don’t translate well into other languages.



$$\operatorname{pmi}(x;y) \equiv \log\frac{p(x,y)}{p(x)p(y)} = \log\frac{p(x|y)}{p(x)} = \log\frac{p(y|x)}{p(y)}.$$

TF-IDF

- TF: Term Frequency
  - Proportional to the number of times a word is used in a document or section
- IDF: Inverse Document Frequency
  - Inversely Proportional to the number of documents or sections in which a word appears

- TF-IDF scores provide a way to find words relevant to a section, while ignoring words that are common across all sections.

Word Recurrence Interval (WRI)

- WRI is defined as the number of words in between successive occurrences of a keyword

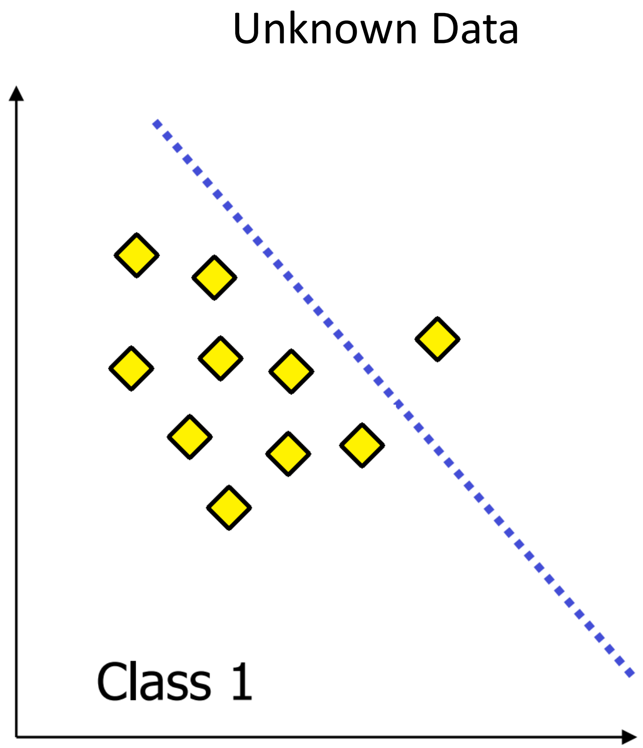
- Keyword being: I

  1 2 3 4 5 6 7 8 9 10 11

  I have six locks on my door all in a row. When I go out, I lock every other one. I figure no matter how long somebody stands there picking the locks, they are always locking three.

- Word Recurrence interval is: {0, 11, 2, 4}

Support Vector Machine (SVM)



- SVM is a binary classifier
- Defines a decision point from a set of training data which is split into two distinct classes
- Assigns new testing data into one of those classes based on the decision point.
- Can be used for authorship detection

Picture Modified From: Martin Law, 3/1/11, http://www.cise.ufl.edu/class/cis4930sp11dtm/notes/intro_svm_new.pdf
Reference: Ebrahimpour M, Putniņš TJ, Berryman MJ, Allison A, Ng BW-H, et al. (2013) Automated Authorship Attribution Using Advanced Signal Classification Techniques. PLoS ONE 8(2): e54998. doi:10.1371/journal.pone.0054998

Language Investigations (Herbal Book)

- **Language and grammar was lax at times**
  - Repeated letters skipped
  - Words abbreviated with symbols
- **Position dependent letters**
  - Two different interchangeable versions of letter ‘s’
- **Different authors, different substitutions**
  - Separate authors would substitute words with own symbols
- **Penmanship questionable**
  - Words sometimes written as one word sometimes split apart
- **Words continued on different lines**
  - Occasionally would have an indicator to show word had been split

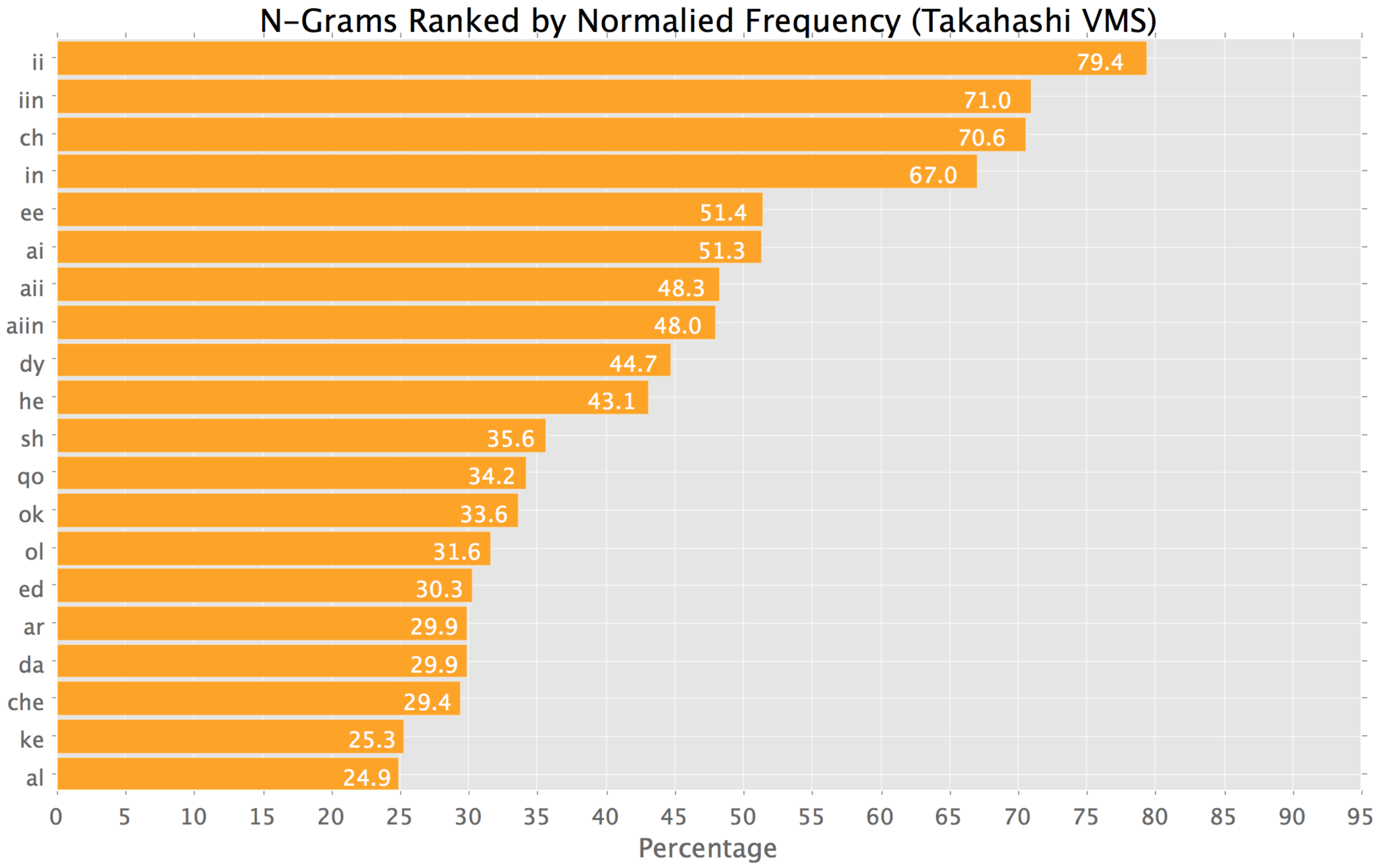
THE UNIVERSITY
of ADELAIDE

Results and Analysis

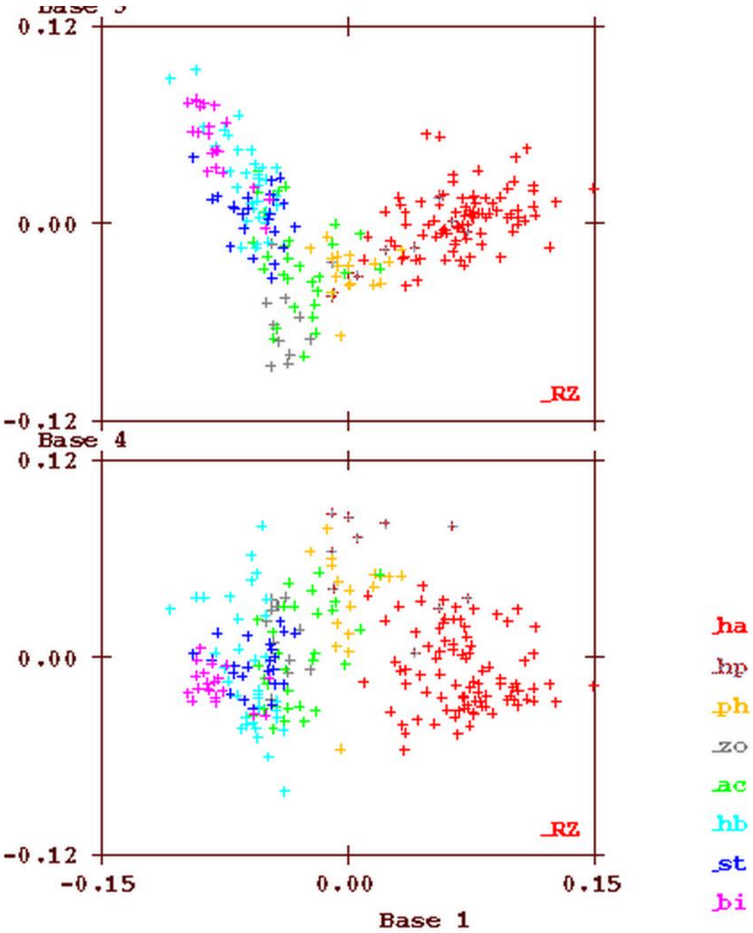
Information Learnt

Section	Currier Language	Pages	Tokens	Words	Words per Page	Full Alphabet Length	Common Alphabet Length
Cosmological	Unknown	20	3008	1521	150	27	24
Biological	B	20	6917	1549	346	21	18
Herbal A	A	97	7956	2492	82	32	21
Herbal B	B	32	3442	1349	108	23	20
Recipes	B	25	11417	3328	457	29	19
Pharma	A	18	2573	1139	143	21	19
Zodiac	Unknown	12	1331	808	111	20	19
Unclassified	Unknown	12	1276	708	106	28	24
Missing		20	0	0	0	0	0
Full Manuscript		**256**	**37945**	**8105**	**161**	**47**	**21**

Common Letter Combinations



Word and Illustration Relationships

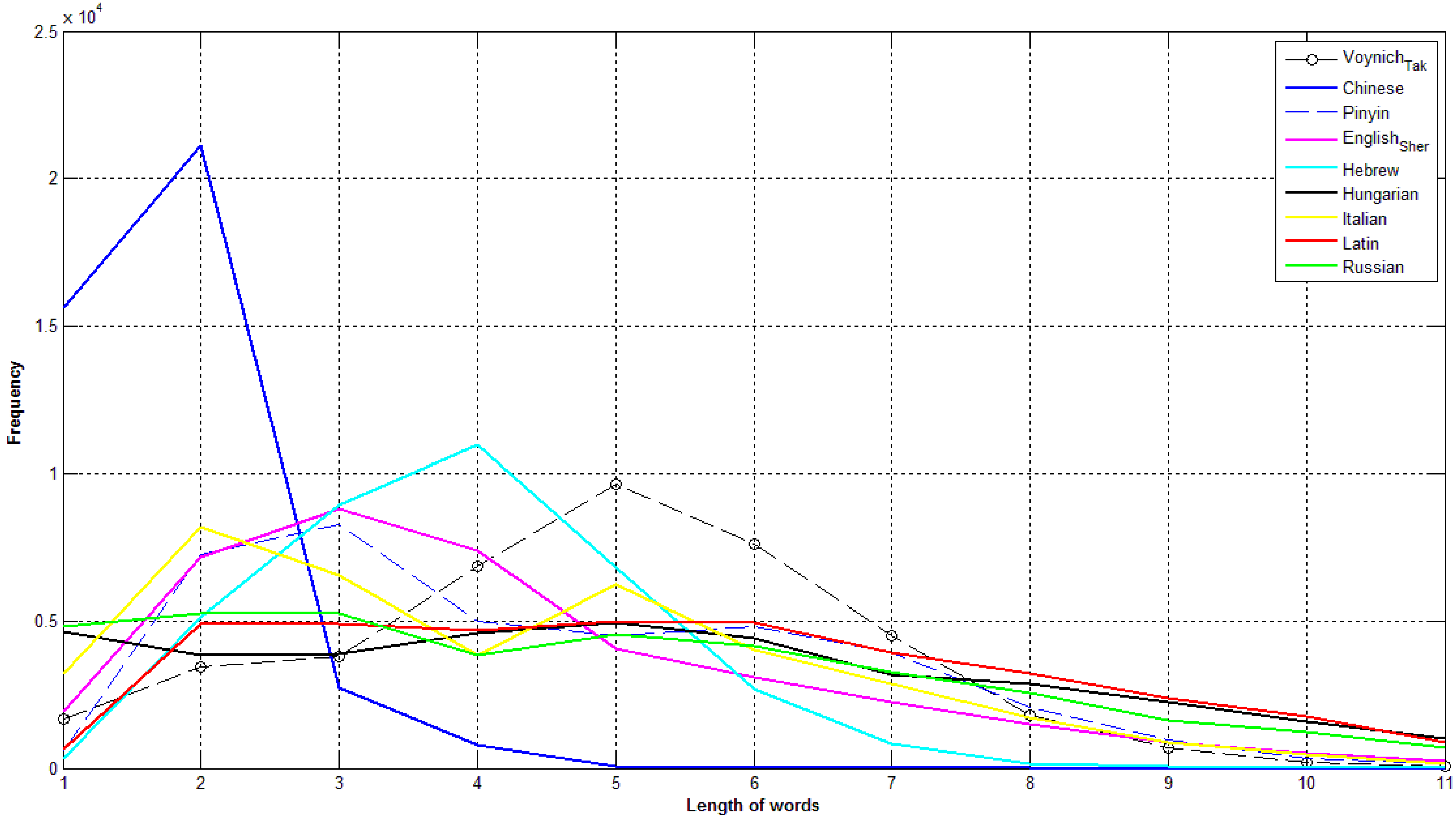


OFCC9 ZC89 8AN ZC9 SCFCC89 ROR 4OFAN SAE FAN ZOCF9 EPAN OR 9FCC9 EZC9 FAE 4OFCC89 FAR SC9 R AN ZOC89 4OFCC9 4OF9 FCCOR 8AR OFC9 SQ9 8SC89 SX9 EZC89
OFCC89 4OFC89 OESC89 2AE S89 OPC9 ESC89 4OFAE OFAR FCC89 2AN OPAR 2AM OPZC89 BSC8AR 8AE ZAE OPAN SAR AR ESC9 EOR AJ OEFCC9 SCQ9 ZX9 OPC89 8AT
ZCQ9 E SFAE 4OFC9 SCF9 SQC89 EPC89 OFAE OP9 4OFCS89 OPCC89 4OPCC9 4OPCC89 OFC89 OPAJ 2ZC89 2 O 8AM EFAR 8E ES89 FC9 SCC89 8AR9 ZCOR 4OFCOR AM
ZCC9 SE FCC9 4OPAN OFAN 4OPS89 OPCC9 SXC9 EFC89 4OFCCO AE SC8AN EFCC89 ZC8 SOE OEAN ZC8AE 4OFAM 9 SCOE EFC9 4OBSC9 OFS9 SCAE
AEOE ZCAR ORAM 4OXC9 RAM OFC8AE SCQC9 4OPCC2 ZC8AJ OFCCC89 4OXC89 SCXC9 OPCOE 2AEFCC89 PCC89 SXC89 PCC9 ZCFCC9 ZCCOE ZCAE
ON 4OFCAR SR 8AJ EAJ 8ZCC89 8SCC9 ZAR SQ*9 EOFCC89 ER SXAE 4OPSC89 ESCOE SC8AR ORAN 4OPSC9 AT OIF*9 4OSC9 OEFAE 8ZC9 PAR EFO
EFCC9 ZCO AEOJ FCCOE 4OFCSC89 O8AM ZCFAE 8SC8 4OVSC89 SO89 4OFCOE ESC8 SFCC9 8SCC9 EOFAJ 4OF OFCCOE SCBSC89 8OM ZCCX9 PC9
OPC8AR OPC2 4OFCS9 OPCO89 SCO89 4OFCO89 OPCC8 AK EVS9 SAM PAT CCC2 OJ EFAJ FCCC9 SCO2 8AK ZCCF EOP9 ZCFAN 4OFCCS9 OFCAE
EFCCO89 ZCO8 ZC8AN BSAE OPCAE OPCCAJ 4OAN 4O8 SCCO SCO SCOFCC9 AEAJ OFCCO EFCSC9 EFCSC89 SCCFAN OPCCC9 EFAT ESAE
OPCCOE SO FCSC89 OEFCCC89 4OFCO8 4OSC89 OPCO ZCCO8AR SCOJ OPARAE OAM OEFCCO EFCCOE EFAE EFCCC9 PC8AJ OFC8AN ZOF
4OFCCOE 4OFCCO2 OPCC8AN EFCCC89 SCAJ SCXC8 OSC9 FSOES8AR AIIB SCPAEZ9 4CCAE 4OFCCA2 SCOFC89 ZCCFZ9 SOFSC9 SX*9
9SCCO8AN 9FCC8AM RCCC9 OEA3 AIF*9 ZFAM 8ZCCO OPSC8C9 OPCOEAT 4OZCO 8SC8AR 9FCCCO2 BSC8C9 OPCO8O BS8AJ OFCO8AN ZCP
ZCOPAJ ZPAR OESCO89 ZCQC9 ESCOCFAJ 4OFCCO8OR SCQ89 4OPC8E EOC89 SC8C9 EFAK O*OR F98CC89 ZCCP SCOP989 2ZCO PSC8 FCOQC89
4OFCZC9 FCCZO89 OFCSC89 ESR 2O AEAE O*AR 2OAM OPCO8AM 4OPCC8AM PCC8AN ZCOFAR RF9 SPAR 4OTAN ZCOPSC89 ZFC9 4OFCAN
ZFCO89 8ZCCOPCC9 PCAR SOEFCCC89 ESCS89 4OCCCO SC8A FCCO8AE PCO OFCOJ 4OFCCO8 EFC8EFC9 PCC8 9SC8E BOEAE B9FCOR SCCV9
OBSC8AE EVSC89 ESCCOE OPCON SCAJAR 4OFCOFC89 OPCCOEFCC9 EAN 4OFCCAE 4OFCCE SC89PCOFAN EFC8AR EFCC8AN OFCAJ PCOEFC8AN
EPCCAE U OFCCC2C9 OESAE 4CCAR BOCOFCC9 PC8AN SCBSAJ 4OFCCOFAN FCCE BOEFCCO SCOFCAN I*AR *AN 9ZC OFZ89 ZFCC9 SXAM

Words and Illustration Relationships

Astrological	Biological	Cosmological	Pharma	Recipes	Herbal
osar	qol	v	daiin	qokeedy	Daiin
oteody	qolkeedy	ytaiin	okeol	qokaiin	chor
oteotey	qokedy	k	ctheol	lchedy	cthor
eody	qokain	{&169}	olchor	lkaiin	ctho
okalar	shedy	{&171}	qoor	lkain	qotchor
okeodaly	lchedy	x	shockhey	qokain	qotchy

Word Lengths and Frequency



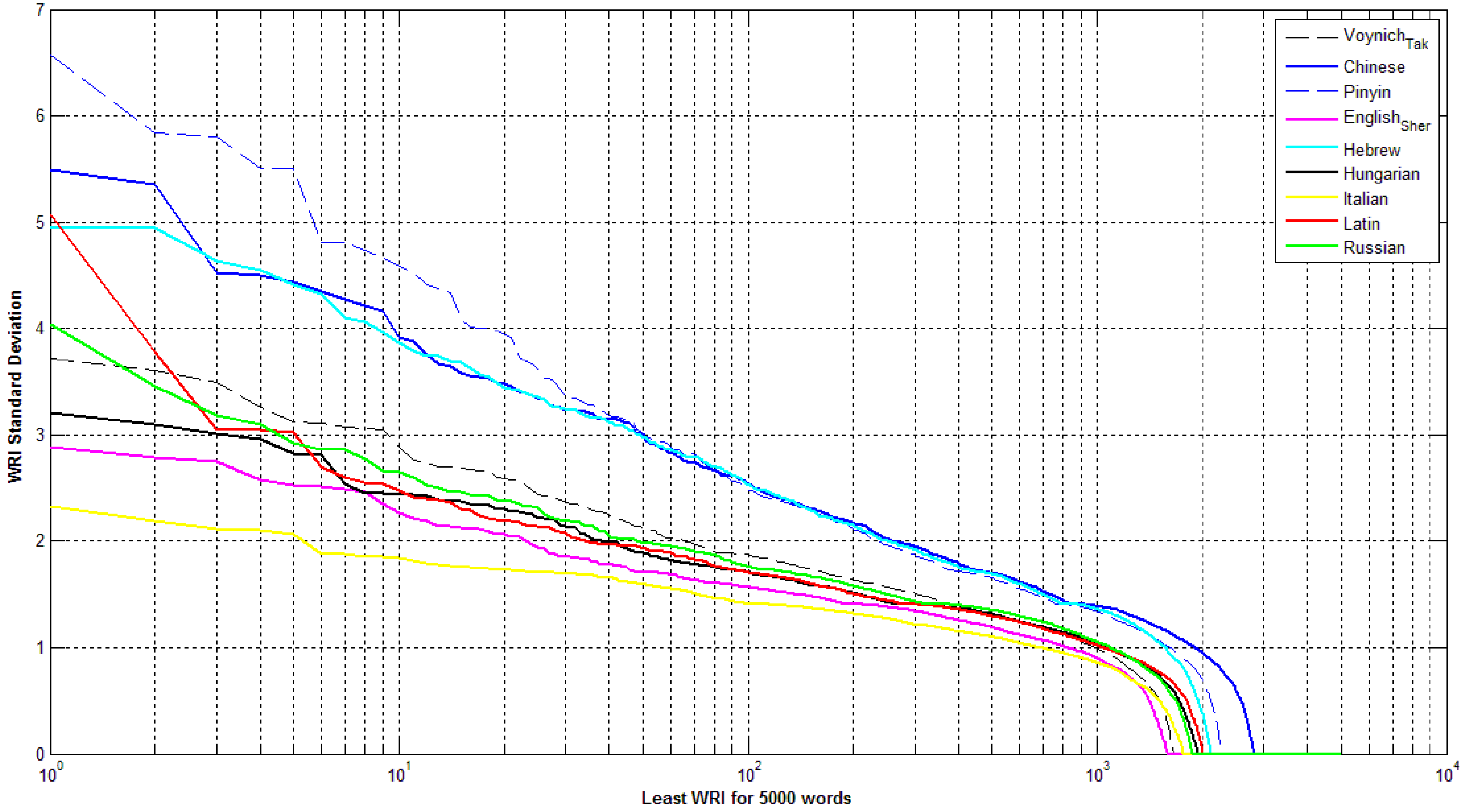
UDHR and Word Lengths

Text	Tolerance	Match	UDHR Match	Peak Length
Voynich	10%	45.45%	Arabic, Standard	**2**
Voynich	15%	54.54%	Arabic, Standard	**2**
Voynich	25%	63.63%	Malay (Arabic)	**4**
Voynich	40%	72.72%	Hebrew, Malay (Arabic), Guarayu, Arabic (Standard)	**4, 4** **5** **2**
Voynich	50%	81.81%	Arabic (Standard), Hausa (Niger), Hausa (Nigeria)	**2** **2** **2**

Voynich:

1	2	3	4	5	6	7	8	9	10	11
4.13%	8.52%	9.45%	17.01%	23.95%	18.84%	11.12%	4.49%	1.68%	0.52%	0.14%

WRI and Rank Plot

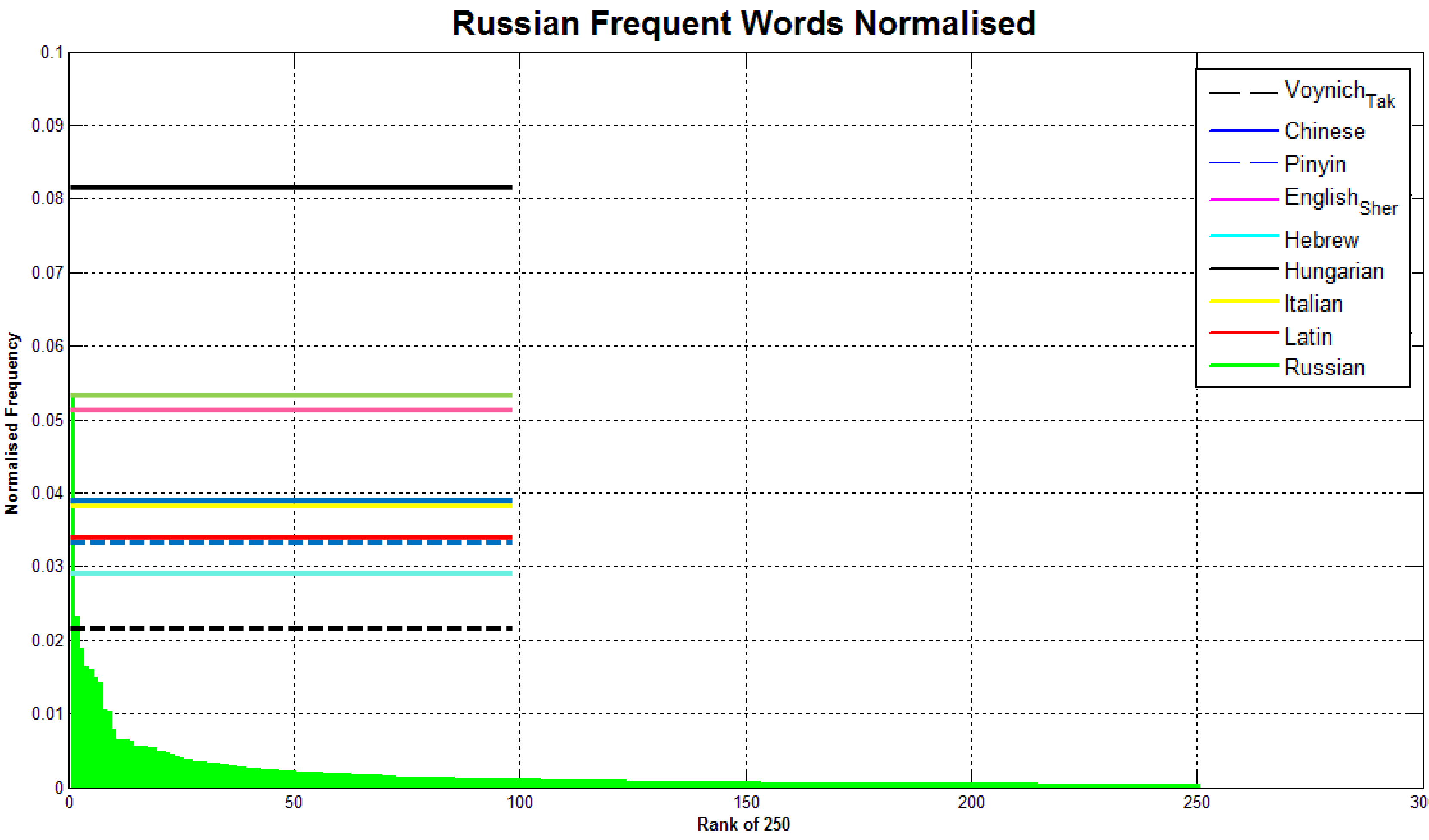


UDHR and WRI

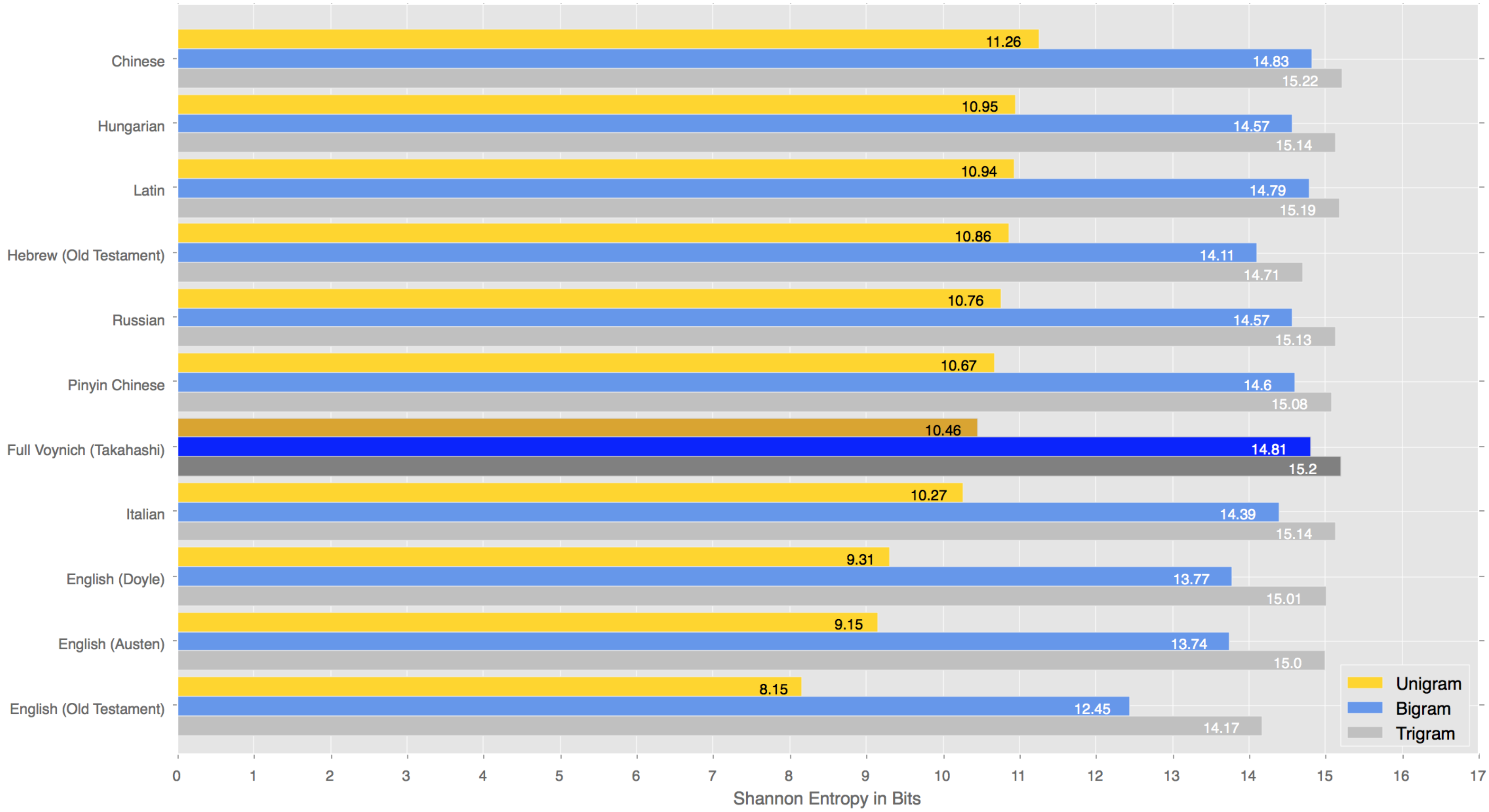
Name	Tolerance	Match	UDHR Match	Comments
Voynich Herbal A	10%	17%	Bosnian (Latin)	**f15r - f22v**
Voynich Herbal A	10%	12%	Jola-Fonyi	**f3r - f10v**
Voynich Biology B	10%	3%	Hmong (Southern Qiandong), Aceh	**f83r - f85r1**
Voynich Recipe B	10%	22%	Bosnian (Latin), Mapudungun	**f113r - f114r**
Herbal Book	10%	8%	Hmong, Southern Qiandong	**16th Century**

- Comparison text of ~1500 words
- Average UDHR text length is ~1800 words
- Top 100 data points

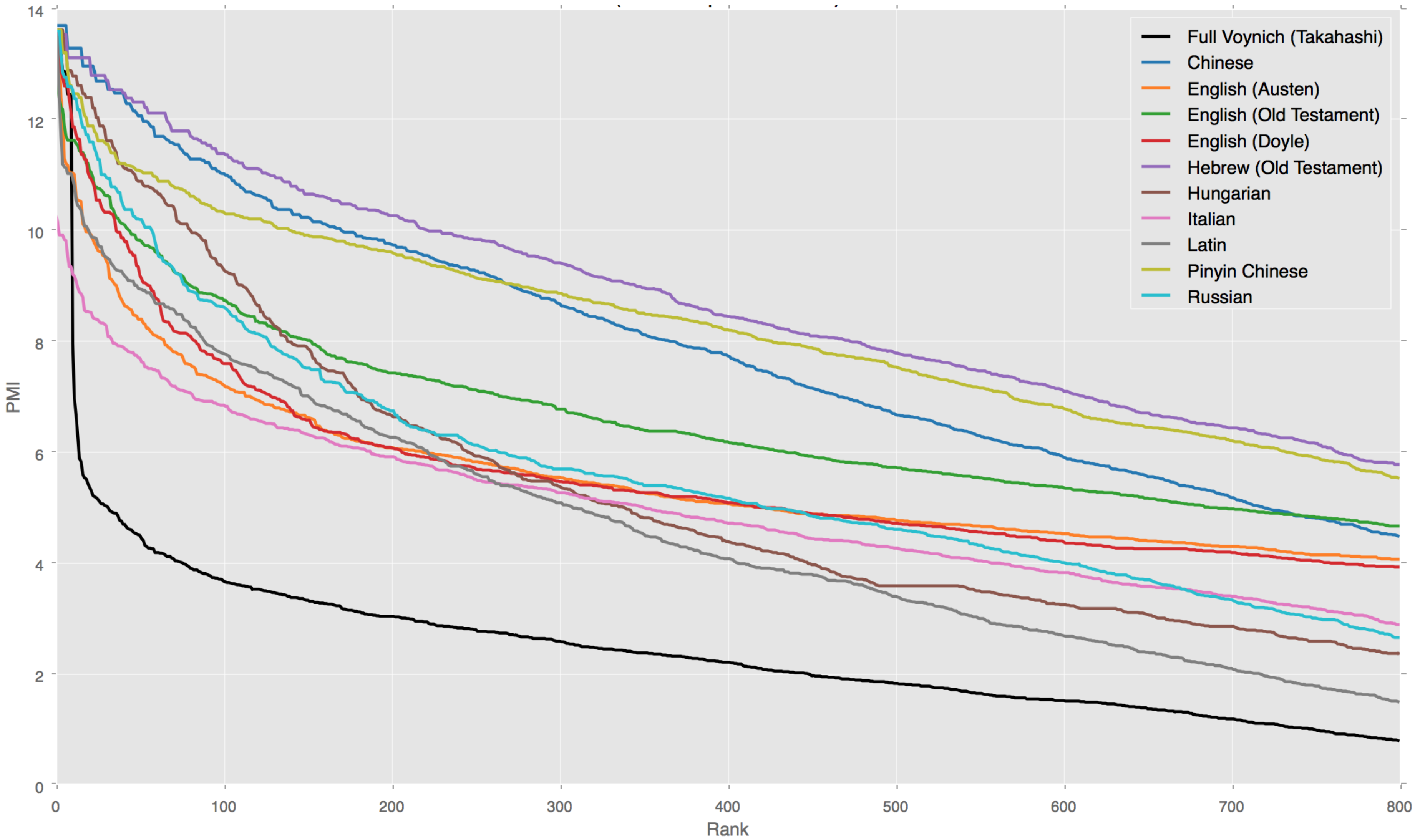
Word Frequency and Zipf's Law



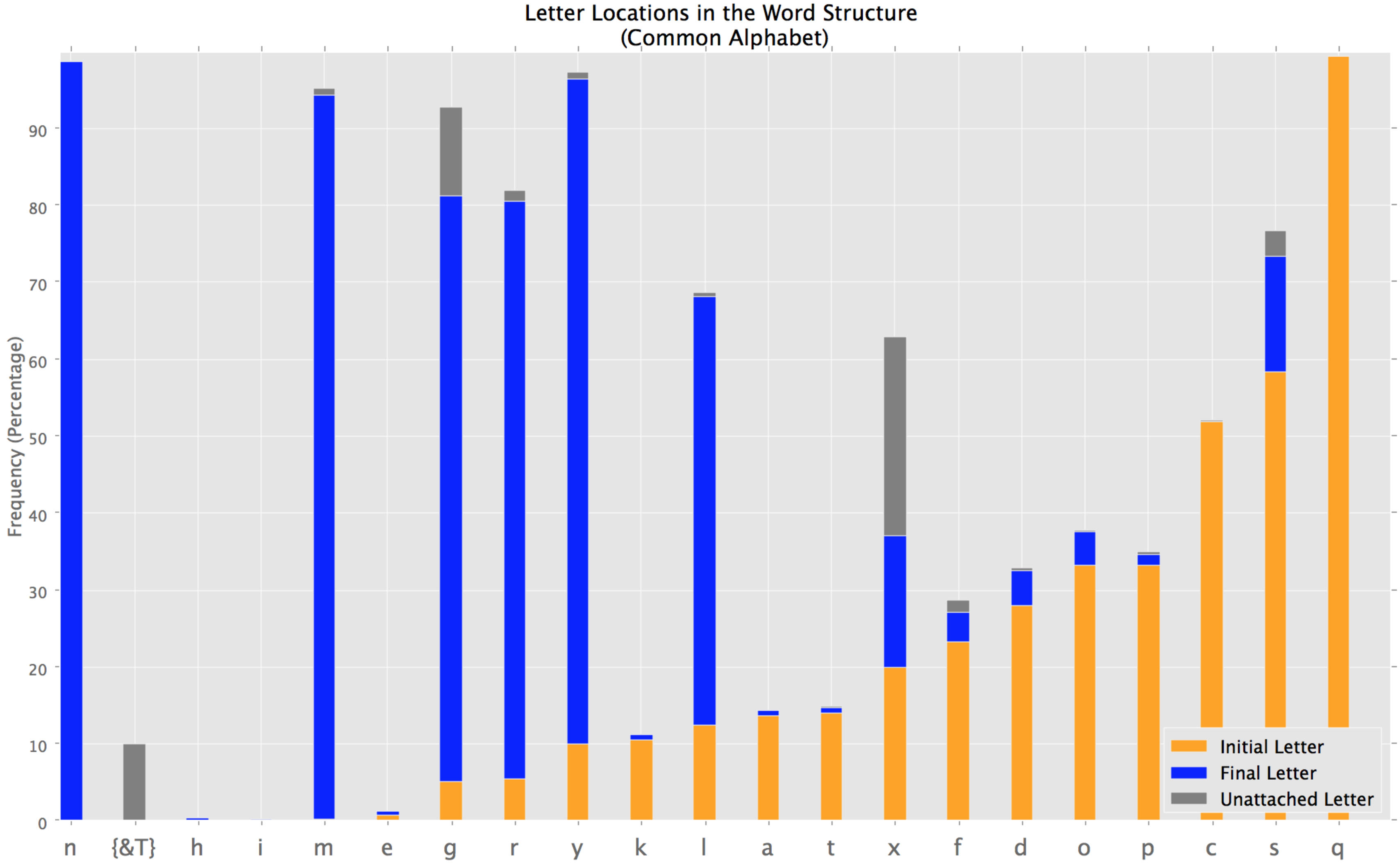
Word Entropy



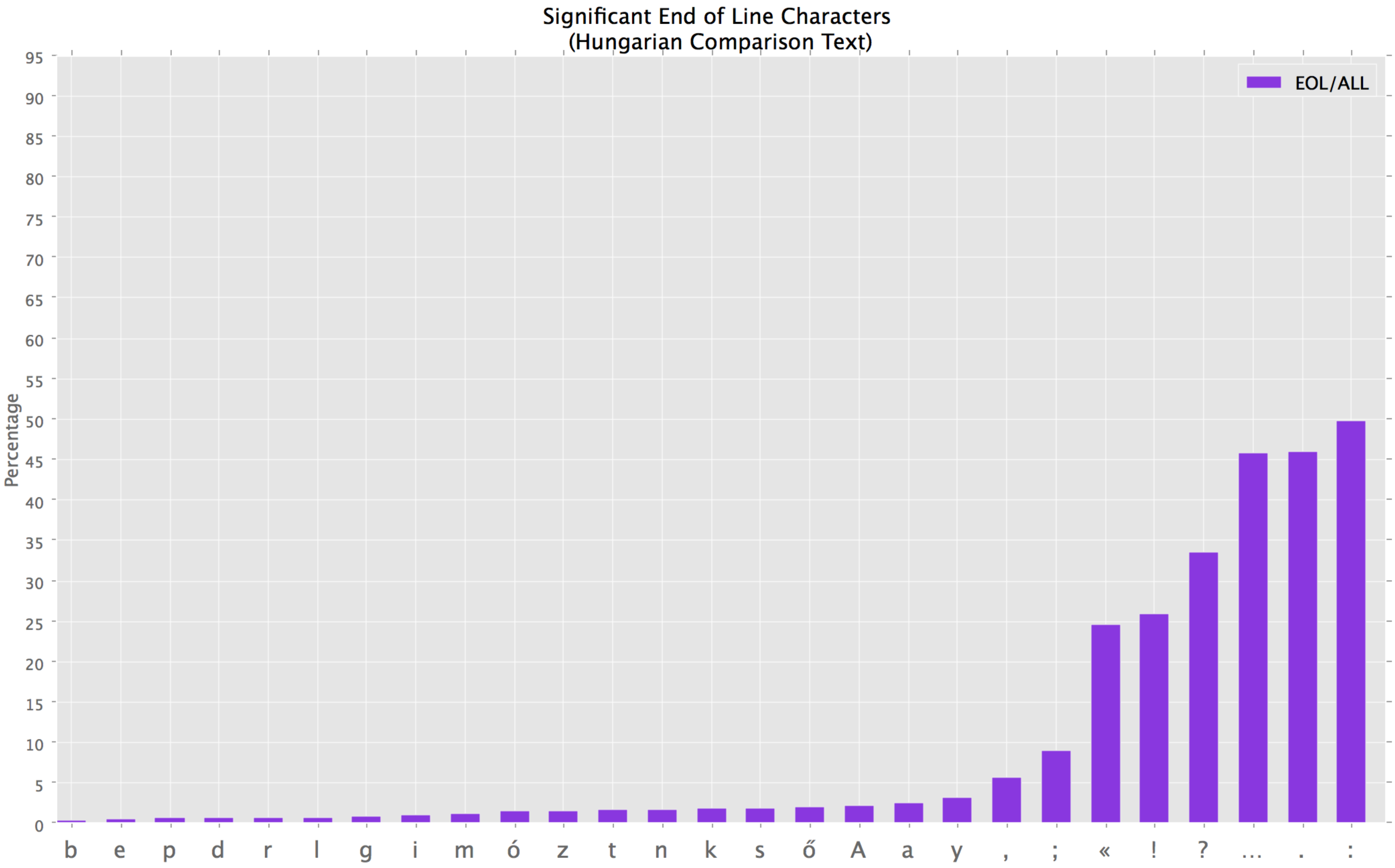
Collocations



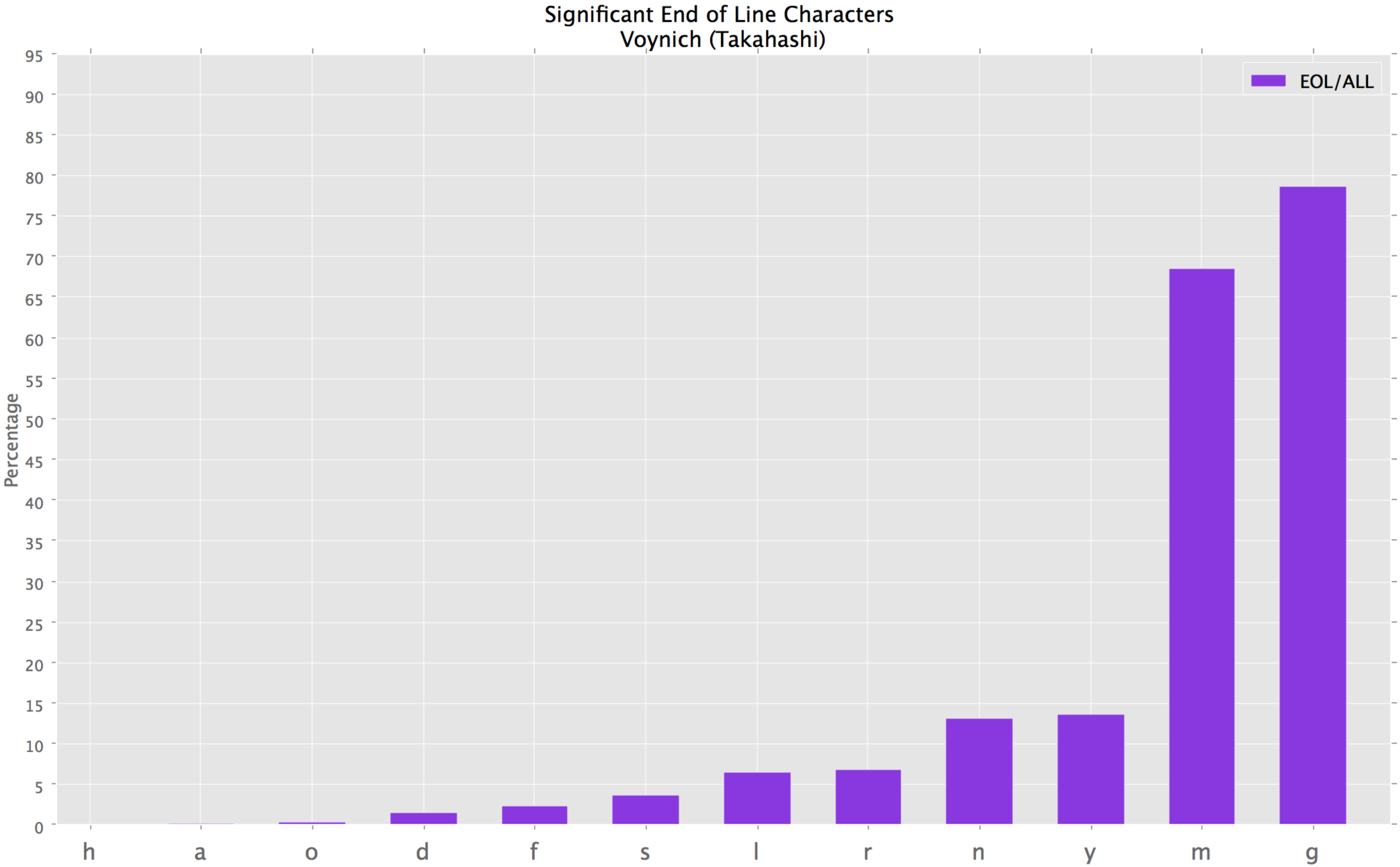
Word Structure



Punctuation



Punctuation



Support Vector Machine (SVM)

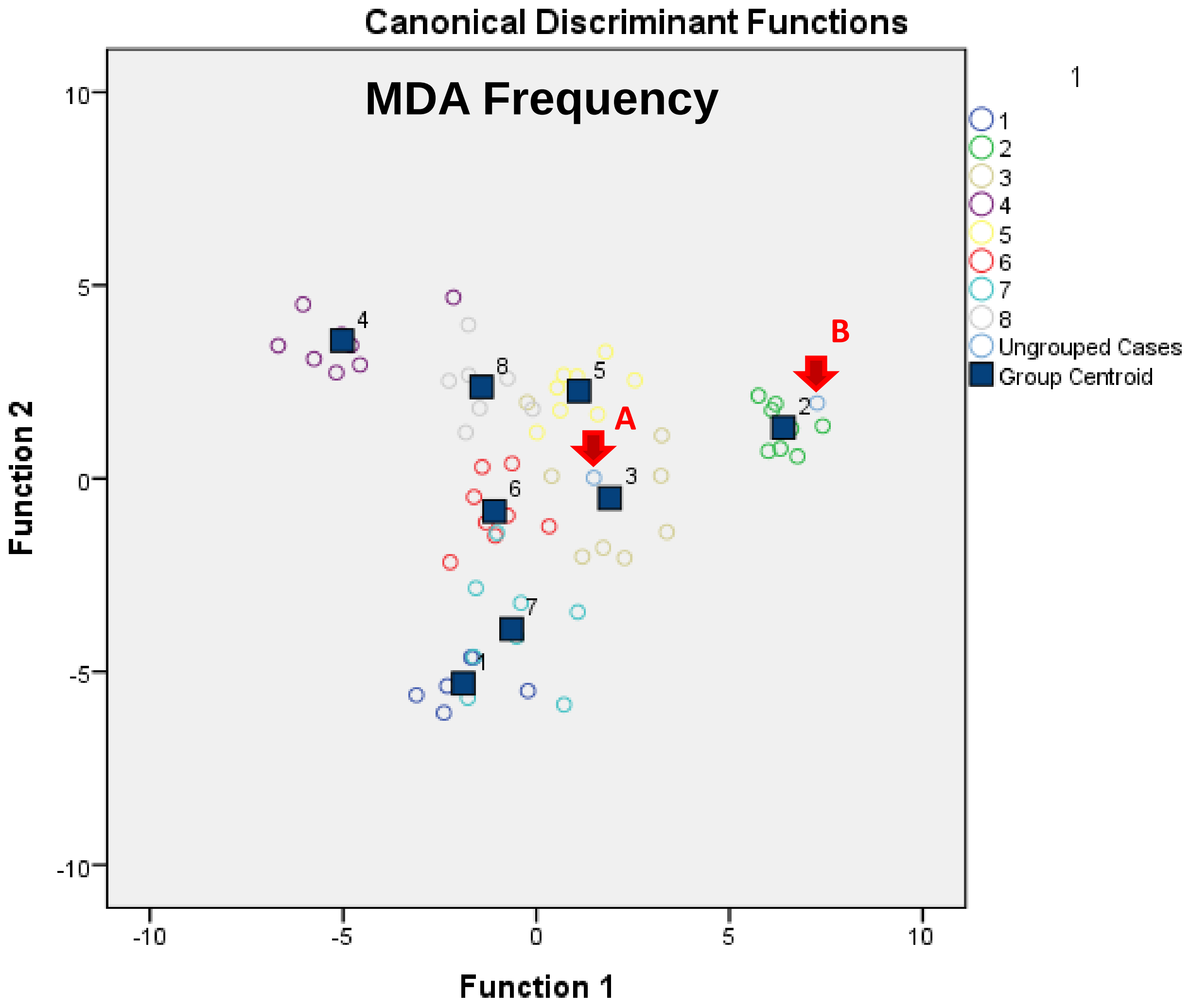
Language	Comparisons	Group
Voynich $_{\text{Takahashi}}$ Normalised	Frequency	**Hebrew**
Voynich $_{\text{Takahashi}}$ σ	WRI	**Russian**

Normal Languages Compared			
Chinese	English $_{\text{Sherlock Holmes}}$	Hebrew	Hungarian
Italian	Latin	PinYin	Russian

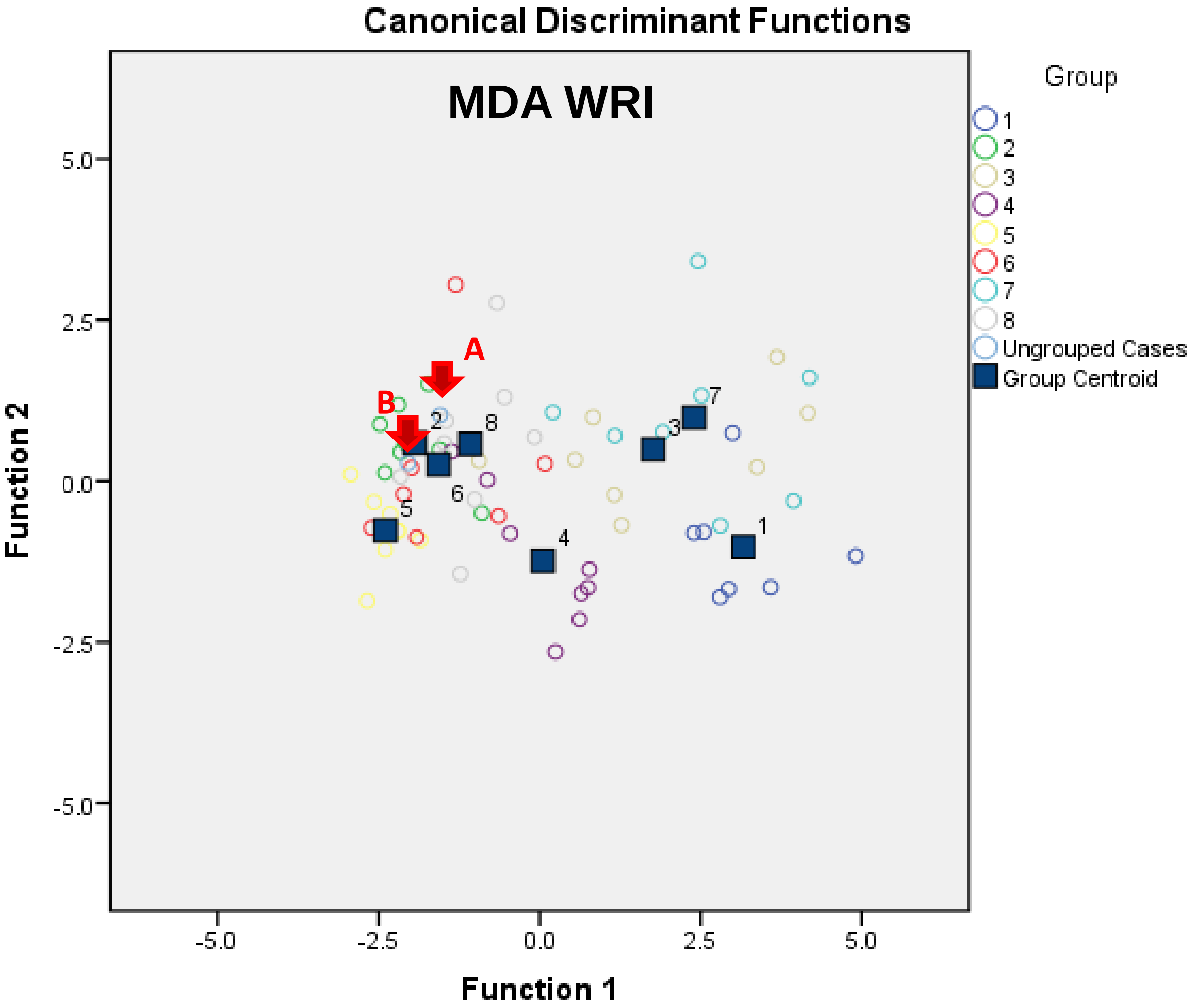
Language	Comparisons	Group
Voynich $_{\text{Takahashi}}$ Herbal A	Frequency	**Zodiac**
Voynich $_{\text{Takahashi}}$ Herbal A	WRI	**Pharmaceutical**

Voynich Languages Compared			
Biological	Cosmological	Herbal A	Herbal B
Pharmaceutical	Recipes	Unknown	Zodiacs

Multiple Discriminant Analysis (MDA)



Multiple Discriminant Analysis (MDA)



THE UNIVERSITY
of ADELAIDE

Risk Management, Budgeting, Timeframes and Approach

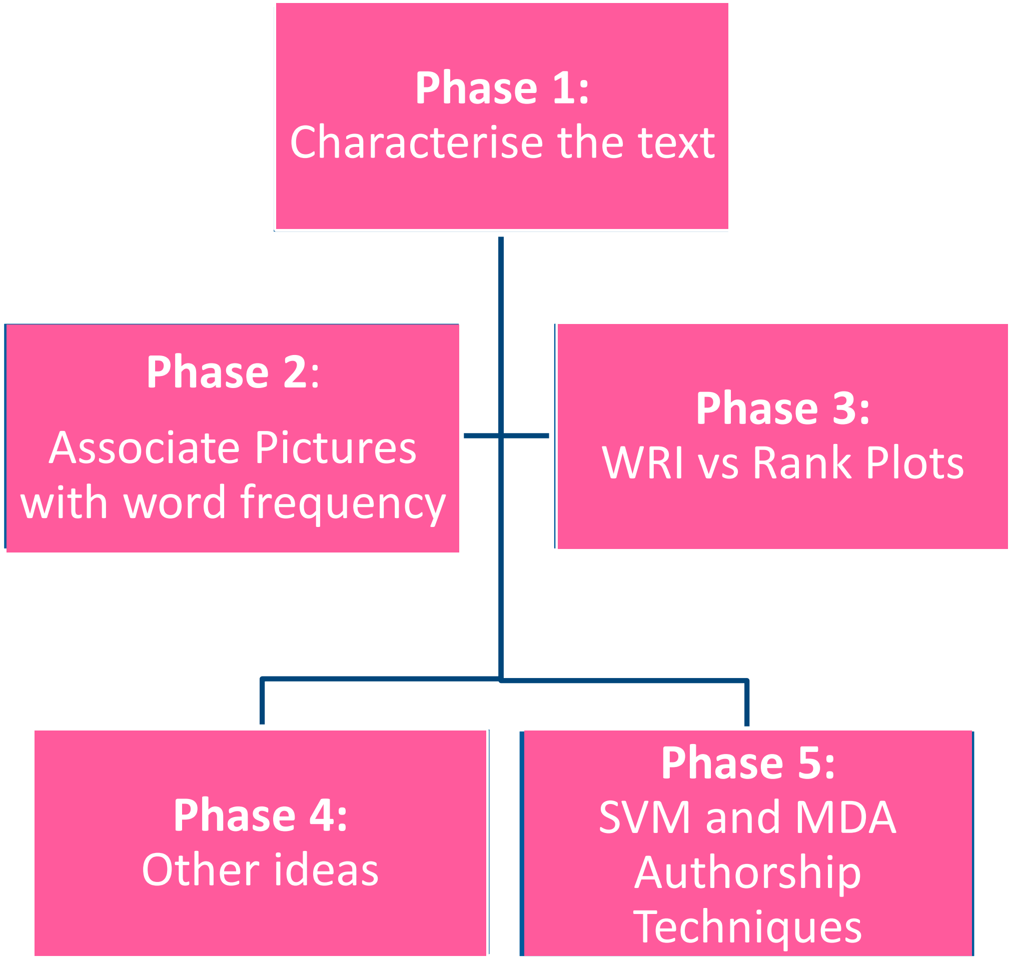
Project Management

Risk Management and Budget

No.	Risk	Likelihood	Consequence	Risk Level
1	Not understanding the project correctly and the processes required	Almost Certain	Moderate	Very High
2	Inaccurate allocation of time and resources to a particular area	Likely	Major	Very High
3	Health issues due to long periods of time sitting and working at a PC	Likely	Moderate	High
4	Files and working copies lost	Rare	Major	Medium
5	UofA Electrical Engineering server down for unknown reasons	Unlikely	Moderate	Medium
6	Not being able to solve the Voynich Manuscript code	Almost Certain	Negligible	Medium

- $396.46 (Spent on 3 books, printing and lamination)

Final Approach



Phase 1:
Characterise the text

Phase 2:

Associate Pictures with word frequency

Phase 3:
WRI vs Rank Plots

Phase 4:
Other ideas

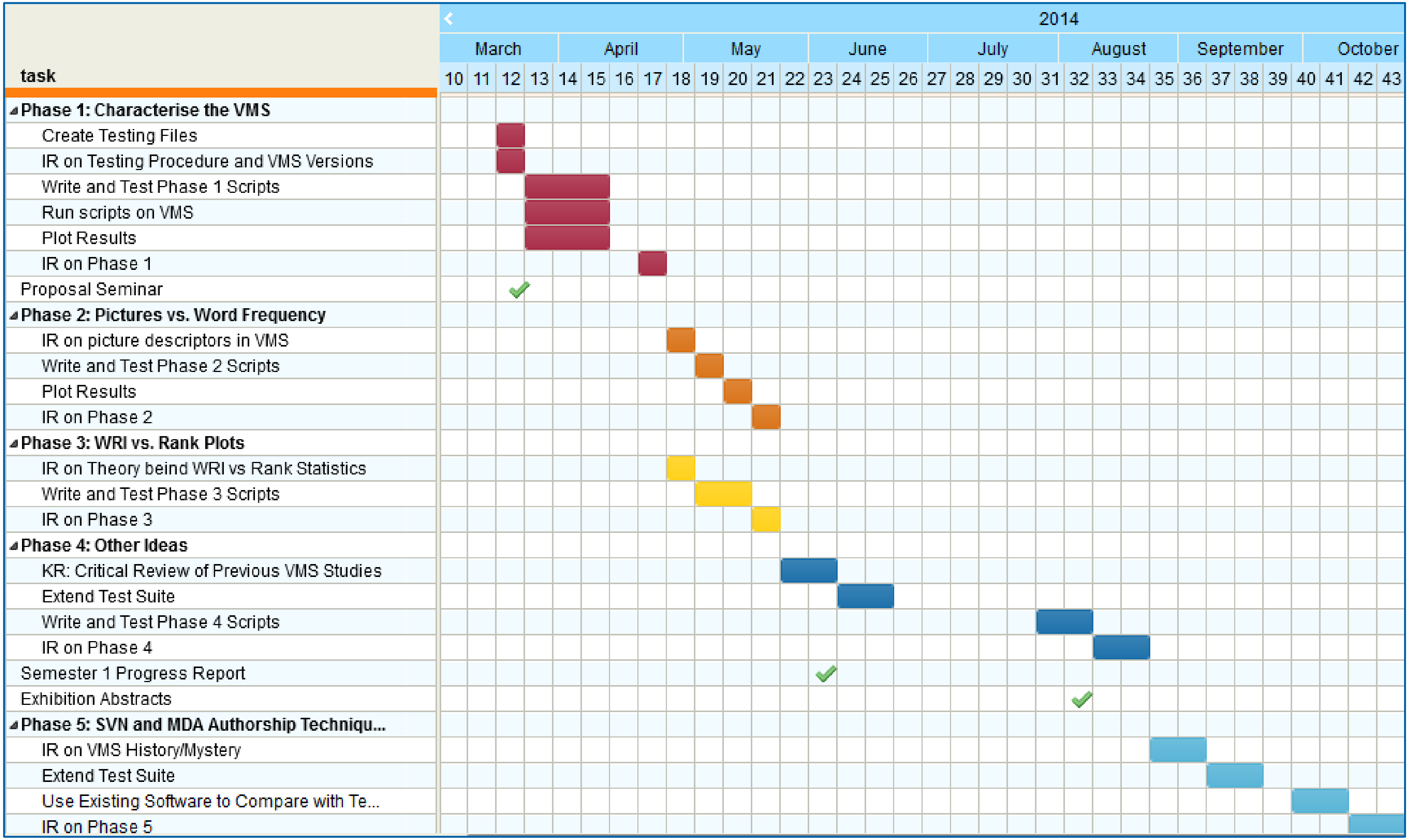
Phase 5:
SVM and MDA Authorship Techniques

Team Roles

- Peter
  - Python Code
  - Phase 2
  - Phase 4
  - Compilation of testing material
  - Research as necessary

- Bryce
  - MATLAB Code
  - Phase 3
  - Phase 5
  - Analysis of known 15th Century Text
  - Research as necessary

Project Progress



task

- Phase 1: Characterise the VMS
  - Create Testing Files
  - IR on Testing Procedure and VMS Versions
  - Write and Test Phase 1 Scripts
  - Run scripts on VMS
  - Plot Results
  - IR on Phase 1
- Proposal Seminar
- Phase 2: Pictures vs. Word Frequency
  - IR on picture descriptors in VMS
  - Write and Test Phase 2 Scripts
  - Plot Results
  - IR on Phase 2
- Phase 3: WRI vs. Rank Plots
  - IR on Theory beind WRI vs Rank Statistics
  - Write and Test Phase 3 Scripts
  - IR on Phase 3
- Phase 4: Other Ideas
  - KR: Critical Review of Previous VMS Studies
  - Extend Test Suite
  - Write and Test Phase 4 Scripts
  - IR on Phase 4
- Semester 1 Progress Report
- Exhibition Abstracts
- Phase 5: SVN and MDA Authorship Techniqu...
  - IR on VMS History/Mystery
  - Extend Test Suite
  - Use Existing Software to Compare with Te...
  - IR on Phase 5

Interpretation of Results

Conclusion

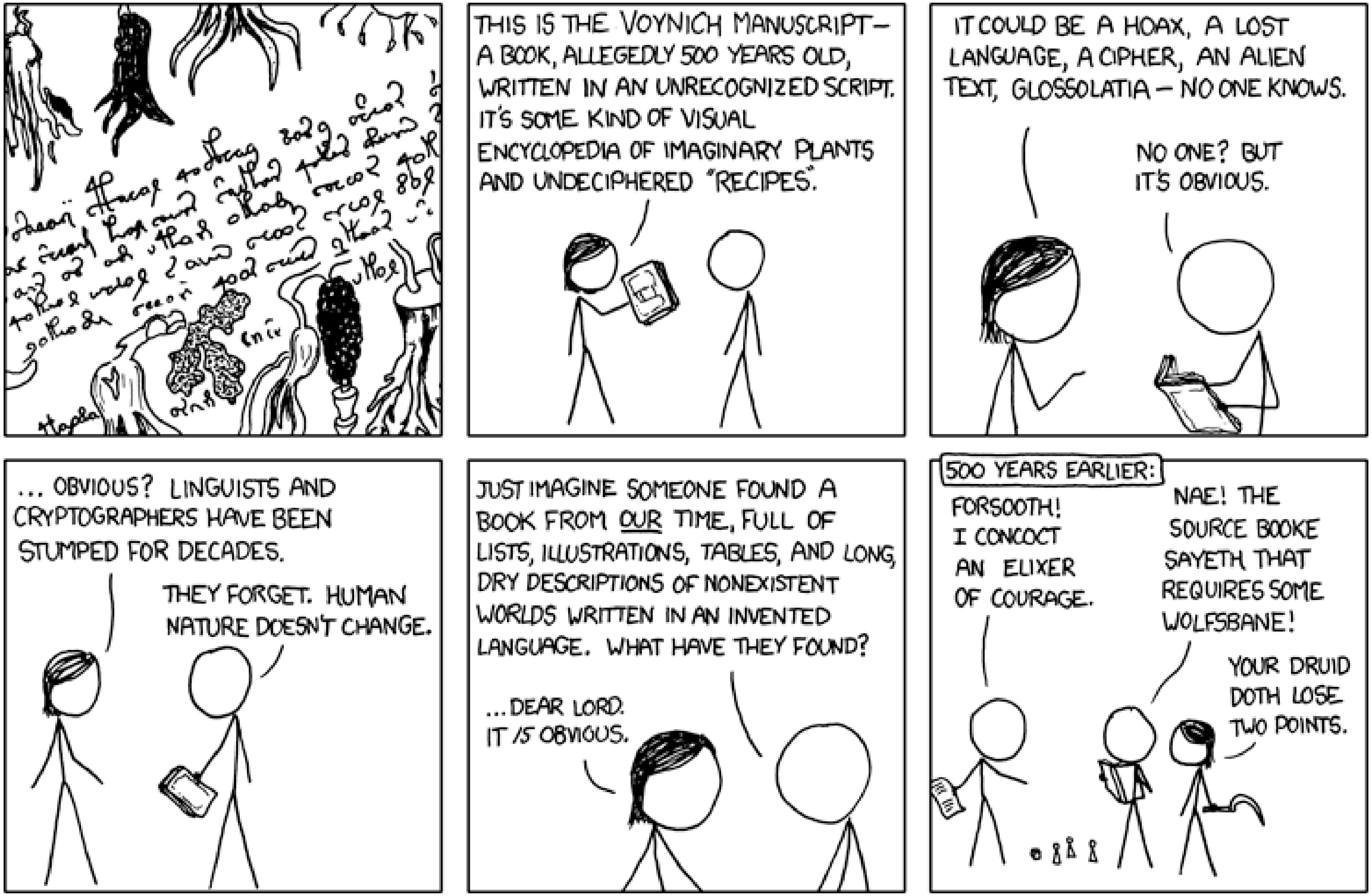
Conclusions

- The writing and language in the Voynich appears to have evolved over time, making analysis difficult.
- There is a relationship between language and section, but this may not have anything to do with illustrations
- Based on characteristics such as word length distribution and WRI, appears similar to languages such as Hebrew and Latin
- May contain punctuation, based on line characteristics.
- Weak word order, indicating lack of phrases and proper nouns, or perhaps indicating the characteristics of a code

Future Pathways

- Expand research into word/illustration relationship
- Test the effect of modified alphabets
- Expand research into authorship if possible
- Hidden Markov Model classification of text
- Develop a rule-based grammar for the the Manuscript if possible
- Test characteristics against transcriptions of known 15th century codes

Questions?


THIS IS THE VOYNICH MANUSCRIPT— A BOOK, ALLEGEDLY 500 YEARS OLD, WRITTEN IN AN UNRECOGNIZED SCRIPT. IT'S SOME KIND OF VISUAL ENCYCLOPEDIA OF IMAGINARY PLANTS AND UNDECIPHERED "RECIPES."
IT COULD BE A HOAX, A LOST LANGUAGE, A CIPHER, AN ALIEN TEXT, GLOSSOLATIA — NO ONE KNOWS.
NO ONE? BUT IT'S OBVIOUS.
... OBVIOUS? LINGUISTS AND CRYPTOGRAPHERS HAVE BEEN STUMPED FOR DECADES.
THEY FORGET. HUMAN NATURE DOESN'T CHANGE.
JUST IMAGINE SOMEONE FOUND A BOOK FROM OUR TIME, FULL OF LISTS, ILLUSTRATIONS, TABLES, AND LONG, DRY DESCRIPTIONS OF NONEXISTENT WORLDS WRITTEN IN AN INVENTED LANGUAGE. WHAT HAVE THEY FOUND?
...DEAR LORD. IT IS OBVIOUS.
500 YEARS EARLIER:
FORSOOTH! I CONCOCT AN ELIXER OF COURAGE.
NAE! THE SOURCE BOOKE SAYETH THAT REQUIRES SOME WOLFSBANE!
YOUR DRUID DOTH LOSE TWO POINTS.

Reproduced under the Creative Commons Attribution-Non Commercial 2.5 licence from **xkcd**